NATIONAL OPEN UNIVERSITY OF NIGERIA

APPLIED ECONOMETRICS
COURSE CODE: ECO 713

FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

COURSE CONTENT DEVELOPER
Joshua Sunday Riti (PhD)
Department of Economics, Faculty of Social Sciences
University of Jos, Jos-Nigeria

Course Content Editor
Professor Femi Saibu
Department of Economics, Faculty of Social Sciences
University of Lagos

National Open University of Nigeria,
Headquarters,
University Village,
Plot 91, Cadastral Zone,
Nnamdi Azikiwe Expressway,
Jabi, Abuja.

Lagos Office
14/16 Ahmadu Bello Way,
Victoria Island, Lagos.

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

 Printed: 2018 ISBN: 978-058-023-X.

CONTENT

Introduction
Course Content
Course Aims
Course Objectives
Working through This Course
Course Materials
Study Units
Textbooks and References
Assignment File
Presentation Schedule
Assessment
Tutor-Marked Assignment (TMAs)
Final Examination and Grading
Course Marking Scheme
Course Overview
How to Get the Most from This Course
Tutors and Tutorials
Summary

Introduction

Welcome to ECO 713: APPLIED ECONOMETRICS.

ECO 713: Applied Econometric is a three-credit and one-semester postgraduate course for postgraduate Economics students. The course is made up of twelve units spread across twelve lectures weeks. This course guide gives you an insight to Applied Econometrics in a broader way and how to study to make use and apply econometric issues in quantifying economic relationships. It tells you about the course materials and how you can work your way through these materials. It suggests some general guidelines for the amount of time required of you on each unit in order to achieve the course aims and objectives successfully. Answers to your tutor marked assignments (TMAs) are therein already.

Course Content

This course is basically on Applied Econometrics because as you are aspiring to become a quantitative economist, you must be able to apply the knowledge of statistics and mathematics to quantify and solve economic relationship problems. The topics covered include Definition and scope of econometrics, stages of econometric research. Regression analysis (Simple and multiple) and the statistical tests of significance, Econometric problems (heteroscedasticity, autocorrelation, multicollinearity): their causes, detection, consequences and correction. Basic ideas of the identification problem, dummy variables, and distributed lags. Simultaneous equation estimation methods (2SLS, 3SLS, etc); Matrix treatment of multiple regression; Advanced treatment of the simultaneous equation estimation techniques. Instrumental variables, Time series econmetr1cs.

Course Aims

The aim of this course is to give you in-depth understanding of the development as regards:

- Definition and Scope of Econometrics
- Correlation Analysis
- Simple Regression Model and Statistical Test of Significance
- Multiple Regression Model and Statistical Test of Significance

- Econometric Problems (Heteroscedasticity, Autocorrelation and Multicollinearity)
- Basic Ideas of the Identification Problem, Dummy variables and Distributed lag Models
- Simultaneous Equation Estimation Methods (2SLS, 3SLS, etc)
- Matrix treatment of Multiple Regression and Advanced treatment of Simultaneous Equation Estimation Techniques
- Advanced Treatment of the Simultaneous-equation Estimation Techniques
- Time Series Econometrics

Course Objectives

To achieve the aims of this course, there are overall objectives which the course is out to achieve though, there are set out objectives for each unit. The unit objectives are included at the beginning of a unit; you should read them before you start working through the unit. You may want to refer to them during your study of the unit to check on your progress. You should always look at the unit objectives after completing a unit. This is to assist the students in accomplishing the tasks entailed in this course. In this way, you can be sure you have done what was required of you by the unit. The objectives serves as study guides, such that student could know if he/she is able to grab the knowledge of each unit through the sets of objectives in each one. At the end of the course period, the students are expected to be able to:

- Define the term Econometrics
- Explain the scope/division of Econometrics
- State the objectives/goals of Econometrics
- Describe the stages of Econometrics
- Define the term Correlation
- Examine the nature of Correlation between Variables
- Distinguish between measures of Correlation: The Population correlation Coefficient, ρ and its Sample Estimate, r.
- Estimate numerical Value of the Correlation Coefficient.
- Explain the meaning of simple regression model

- Describe the assumptions of the linear stochastic regression model.
- Discuss the Least Squares Criterion and the Normal Equations of OLS.
- Evaluate the statistical test of significance of the Least Squares estimates.
- Illustrate models with two explanatory variables.
- Derive the normal equation of two explanatory variables.
- Estimate the coefficient of multiple determinations and the adjusted coefficient of multiple determinations.
- Calculate the mean and variance of parameter estimates ($\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$)
- Test the statistical significance of the parameter estimates
- Examine econometric problems of heteroscedasticity, autocorrelation and multicollinearity: their causes, detection, consequences and correction.
- Identify the basic ideas of identification, dummy variables and distributed lag models.
- Explain simultaneous equation estimation methods (2SLS, 3SLS etc.).
- Discuss matrix treatment of multiple regression and advance treatment of simultaneous equation estimation techniques.
- Define what identification problem is all about.
- State the implications of identification problems
- State the formal rules or conditions for identification.
- State the nature of dummy variables.
- Compute ANOVA models.
- Estimate ANCOVA models.
- Analyse regression with a mixture of quantitative and qualitative regressors.

- Examine the use of dummy variables in seasonal analysis
- State the nature of simultaneous-equation model.
- Identify simultaneous-equation bias in a model and the inconsistency of the OLS estimators.
- Describe approaches to simultaneous-equation estimators.
- Examine recursive models and the OLS
- Determine estimation of exactly identified and over-identified equations.
- Analyze matrix formulation of the regression model
- Estimate least squares estimate in matrix notation
- Analyze further matrix result for multiple regression
- Determine the method of Instrumental Variables (IV)
- Examine the method of Generalised Least Squares (GLS)
- Solve the method of Three Stage least Squares (3SLS)
- Analyze matrix formulation of the regression model
- Estimate least squares estimate in matrix notation
- Analyze further matrix result for multiple regression
- Differentiate between univariate and multivariate time series models.
- Understand Vector Autoregressive (VAR) models and discuss their advantages.
- Understand the concept of causality and its importance in economic applications.
- Estimate VAR models and test for Granger and Sims causality through the use of econometric software
- Understand the concept of stationarity.
- Understand the importance of stationarity and the concept of spurious regressions.
- Understand the concept of unit roots in time series.
- Estimate the DF, ADF and PP tests using appropriate software
- Understand the concept of cointegration in time series.

- Appreciate the importance of cointegration and long-run solutions in econometric applications.
- Understand the error-correction mechanism and its advantages.
- Test for cointegration using the Engle–Granger approach.
- Test for cointegration using the Johansen approach.
- Obtain results of cointegration tests error-correction models and using appropriate econometric software.

Working Through The Course

To successfully complete this course, you are required to read the study units, referenced books and other materials on the course.

Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about 15weeks to complete and some components of the course are outlined under the course material subsection.

Course Material

The major component of the course, What you have to do and how you should allocate your time to each unit in order to complete the course successfully on time are listed follows:

1. Course guide
2. Study unit
3. Textbook
4. Assignment file
5. Presentation schedule

Study Unit

There are 12 units in this course which should be studied carefully and diligently.

MODULE 1: DEFINITON AND SCOPE OF ECONOMETRICS, REGRESSON ANALYSIS AND THE STATISTICAL TEST OF SIGNIFICANCE

Unit 1: Definition and Scope of Econometrics

Unit 2: Unit 3: Simple Regression Model

Unit 3: Multiple Regression Model

Unit 4: Statistical Test of Significance for Simple and Multiple Regressions

MODULE 2: ECONOMETRIC PROBLEMS, BASIC IDEAS OF THE IDENTIFICATION PROBLEM AND SIMULTENEOUS EQUATION ESTIMATION METHODS

Unit 1: Econometric Problems (Heteroscedasticity, Autocorrelation and Multicollinearity)

Unit 2: Basic Ideas of the Identification Problem, Dummy variables and Distributed lag Models

Unit 3: Simultaneous Equation Estimation Methods (2SLS, 3SLS, etc)

Unit 4: Matrix treatment of Multiple Regression and Advanced treatment of Simultaneous Equation Estimation Techniques.

MODULE THREE: MATRIX TREATMENT OF REGRESSION ANALYSIS, TIME SERIES ECONOMETRICS

Unit 1: Matrix Treatment of Multiple Regressions

Unit 2: Vector Auto Regressive (VAR) Models

Unit 3: Non-Stationarity and Unit Roots

Unit 4: Cointegration and Error Correction Model

Each study unit will take at least three hours, and it include the introduction, objective, main content, self-assessment exercise, conclusion, summary and reference. Other areas

border on the Tutor-Marked Assessment (TMA) questions. Some of the self-assessment exercise will necessitate discussion, brainstorming and argument with some of your colleges. You are advised to do so in order to understand and get acquainted with historical economic event as well as notable periods.

There are also textbooks under the reference and other (on-line and off-line) resources for further reading. They are meant to give you additional information if only you can lay your hands on any of them. You are required to study the materials; practice the self-assessment exercise and tutor-marked assignment (TMA) questions for greater and in-depth understanding of the course. By doing so, the stated learning objectives of the course would have been achieved.

Textbook and References

For further reading and more detailed information about the course, the following materials are recommended:

Akerele, A.A. (2002). Operations Research. Dimis Publications, Jos.

Albert, J. H., and S. Chib (1993): \Bayesian analysis of binary and polychotomous response data," *J. Amer. Statist. Assoc.*, 88(422), 669{679.

Allen, R.G.D. (1956). Mathematical Economics, Macmillan, London.

Almon, S. (1965). The Distributed Lag between Capital Appropriations and Net Expenditure, Econometrica, 33, 178 – 196.

Anderson T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statist.* **20**, 46-63.

Anderson, T. W. (2000). Cointegration of economic time series. In *Probability and Statistical Models with Applications*. Chapman and Hall, New York.

Asteriou, D. & Hall, S. (2007). Applied Econometrics: A Modern Approach (Revised Edition), New York: Palgrave Macmillam.

Astrios, D. & Hall, S.G. (2006). Applied econometrics: A modern Approach (Revised Edition), Palgrave Macmillan, New York.

Basmann, R.L. (1957). A generalised Classical Method of Linear Estimation of Coefficients in a structural Equation, Econometrica, vol. 25, pp. 77-83.

Bickel, P. J., and J. A. Yahav (1969): \Some Contributions to the Asymptotic Theory of Bayes Solutions," *Z. Wahrsch. Verw. Geb*, 11, 257{276.

Brennan, 1965). Preface to Econometrics, South-Western Publishing Company, Cincinnati, Ohio

Brooks, C. (2008). Introductory Econometrics for Finance (Second Edition). Campbridge University Press, United Kingdom.

Cagan, P. (1956). The monetary dynamics of hyper-inflation. In *Studies in the Quantity Theory of Money* (Edited by Milton Friedman). University of Chicago Press, Chicago, Ill.

Cagan, P. (1956). The Monetary Dynamics of Hypernflation. In Friedman, Milton (ed.). Studies in Quantity Theory of Money, Chiicago: University of Chicago Press.

Del Negro, M., and F. Schorfheide (2004): Priors from General Equilibrium Models for VARs," *International Economic Review*, 45, 643{673.

Department of Economics, Princeton University, Princeton NJ 08544-1021, U.S.A.

Draper, N.R. & Smith, H. (1998). Applied Regression Analysis, Third Edition, John Wiley online; doi: 10.1002/9781118625590.Email: gchow@princeton.edu (Received October 2000; accepted January 2001)

Engle, R. and Granger, C. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica* **55**, 251-176.

Ericsson, N. R., Hendry, D. F. and Mizon, G. E. (1998). Exogeneity, cointegration, and economic policy analysis. *J. Business Econom. Statist. 16*, 370-387.

Evidence from the U.S. Stock Market. Department of Economics, Princeton University, Princeton, New Jersey.

Goldberger, A.S. (1964). Econometric Theory. Wiley, New York, P. 1.

Goldberger, A.S. (1964). Econometric Theory. Wiley, New York.

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

Gujarati, D.N. (2006). Essentials of Econometrics (Third Edition). McGraw-Hill, New York. P. 1.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1-12.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Harken, L. and Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* **50**, 1269-1286.

HILDRETH, C. (1963): "Bayesian Statisticians and Remote Clients," Econometrica, 31(3), 422–438.

Hirano, K., and J. R. Porter (2003): Asymptotic efficiency in parametric structural models with parameter-dependent support," *Econometrica*, 71(5), 1307{1338.

Ibragimov, I. A., and R. Z. Hasminskii (1981): *Statistical estimation*, vol. 16 of *Applications of Mathematics*. Springer-Verlag, New York, Asymptotic theory, Translated from the Russian by Samuel Kotz.

Johannes, M., and N. Polson (2003): \MCMC Methods for Continuous-Time Financial Econometrics," in *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *J. Econom. Dynamics Contr.* **12**, 231-254.

Johansen, S. (1989). Likelihood based inference in cointegration theory and applications. *Centro Interuniversitario di Econometrica*, Bagni di Lucca, Italy.

Johnston, J. (1962)Econometric Methods, 2nd Edition, p. 210.

Kane, E. (1970). Economic Statistics and Econometrics

Klein, L.R. (1962). An Introduction to Econometrics, Prentice-Hall International, London, pp. 64–6 6, 86-87, 104-105.

Klein, L.R. (1974). A Textbook of Econometrics, 2nd ed., Prentice Hall, Eaglewood Cliffs, N.J., p. 150.

Kmenta, J. (1986). Elements of Econometrics. New York: Macmillan.

Koopmans, T. C. (1950). *Statistical Inferencein Dynamic Economic Models*. Cowles Commision Monograph 10, John Wiley and Sons, New York.

Koopmans, T. C. and William C. H. (1953). *Studiesin Econometric Method*. Cowles Commission Monograph 14, John Wiley and Sons, New York.

Koutsoyannis, A. (1977). Theory of Econometrics (Second Edition). PALGRAVE. New York.

Koyck, L.M. (1954). Distributed Lag and Investment Analysis.. North-Holland, Amsterdam.

Kramer, J.S. (1991). The Logit Models for Economists, Edward Arnold Publishers, London.

KWAN, Y. K. (1998): "Asymptotic Bayesian analysis based on a limited information estimator," Journal of Econometrics, 88, 99–121.

Kwan, Y. K. and Chow, G. C. (1996). Estimating economic effects of political movements in China. *J. Comparative Econom.* **23**, 192-208.

Lancaster, T. (2004): *An introduction to modern Bayesian econometrics*. Blackwell Publishing, Malden, MA.

Laplace, P.-S. (1818): *Theorie analytique des probabilites*. Editions Jacques Gabay (1995), Paris.

Lehmann, E., and G. Casella (1998): *Theory of Point Estimation*. Springer.

Lin, J. (1998). *The Present Value Model of Stock Prices under Rational and Adaptive Expectations:*

Liu, J. S. (2001): *Monte Carlo strategies in scienti¯c computing*, Springer Series in Statistics. Springer-Verlag, New York.

Liu, T.-C. (1960). Underidentification, structural estimation and forecasting. *Econometrica* **28**, 855-865.

Liu, T.-C. (1963). An exploratory quarterly economic model of effective demand in the postwar U.S. economy. *Econometrica* **31**, 301-348.

Liu, T.-C. (1969). A monthly recursive econometric model of United States: a test of feasibility. *Rev. Econom. Statist.* **51**, 1-13.

Liu, T.-C. and Hwa, E. C. (1974). A monthly econometric model of the U.S. economy. *Inter. Econom. Rev.* **15**, 328-365.

Liu, T.-C., Chow, G. C., Koo, A. Y., Fei, J. C., Tsiang, S. C. and Hsing, M. H. (1974). Report on a Joint Study and Discussion of the Future Fiscal and Economic Policies of Taiwan. The Central Bank, Taipei, Taiwan.

Lucas, R. (1976). Econometric policy evaluation: a critique. In *The Philip Curve and Labor Markets* (Edited by K. Brunner and A. H. Meltzer). Carnegie -Rochester Series on Public Policy, Vol. 1. North-Holland, Amsterdam.

M., Lin, C. F., Mao, C. S., Ho, C. S., Liou, R. W. and Yang, Y. F. (1998). A time series approach to econometric models of Taiwan's economy. *Statist. Sinica* **8**, 991-1044.

Marschak, J. (1954). Economic measurements for policy and prediction. Chapter 1 In *Studies in Econometric Method* (Edited byW. C. Hood and T. C. Koopmans), 1-26. JohnWiley and Sons, New York.

McCulloch, R., and P. E. Rossi (1994): \An exact likelihood analysis of the multinomial probit model," *J. Econometrics*, 64(1-2), 207{240.

Mirer, T.W. (1995). Economic Statistics and Econometrics (Third Edition), Prentice-Hall Inc, London.

Muth, J. (1961). Rational expectations and the theory of price movements. *Econometrica* **29**, 315-335.

Poirier, D. J. (1995): *Intermediate statistics and econometrics*. MIT Press, Cambridge, MA, A comparative approach.

Ragnar, F. (1934). Statistical Confluence Analysis by Means of Complete Regresson Systems, Institute of Economics, Olso University, publ. no. 5.

Robert, C. P., and G. Casella (2004): *Monte Carlo statistical methods*, Springer Texts in Statistics. Springer-Verlag, New York, second edn.

Samuelson, P.A., Koopmans, T.C. & Stone, J.R.N. (1954). Report of the Evaluative Committee on Econometrica, Econometrica, Vol. 22, No. 2, PP. 141-146.

Sargan, J.D. (1958). The Estimation of Economic Relationships using Instrumental Variables, Econometrica, vol. 26, pp. 393-405.

SAVAGE, L. J. (1977): "The Shifting Foundations of Statistics," in Logic, Laws and Life,ed. by R. Colodny, pp. 3–18. University of Pittsburgh Press.

Schultz, H. (1938). *The Theory and Measurement of Demand*. University of Chicago Press, Chicago, Ill.

Sims, C. (1980). Macroeconomics and reality. *Econometrica* **48**, 1-48.

SIMS, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples," Journal of Econometrics, 95(2), 443–462, http://www.princeton.edu/˜sims/.

Sims, C. A., and H. Uhlig (1991): \Understanding unit rooters: a helicopter tour," *Econometrica*, 59(6), 1591{1599.

Smets, F., and R. Wouters (2003): \An estimated dynamic stochastic general equilibrium model of the euro area," *J. European Economic Association*, 1, 527{549.

Theil, H. (1953). Repeated least Squares Applied to Complete Equation Systems, The haque: The Central Planning Bureau, The Netherlands.

Theil, H. (1961). *Economic Forecasts and Policy*. North-Holland, Amsterdam.

Tinbergen, J. (1939). *Statistical Testing of Business-Cycle Theories*. League of Nations Economic Intelligence Service, Geneva.

Tinbergen, J. (1952). *On the Theory of Economic Policy*. North-Holland, Amsterdam.

Tinbergen, J. (1956). *Economic Policy: Principles and Design*. North-Holland, Amsterdam.

Tsiang, S. C., Chow, G. C., Hsing, M. H., Fei, J. and Koo, A. (1978). *Economic Planning and Efficient Utilization of Resources*. Economic Planning Council, Taipei, Taiwan.

Uhlig, H. (2005): \What are the e®ects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52, 381{419.

Wonnnacott, R & Wonnacott, T. (1970). Econometrics pp.322-6

Woodford, M. (1999). Optimal monetary policy inertia. mimeo., Princeton University, Princeton, NJ.

Working, E. J. (1927). What do statistical demand curves show? *Quarterly J. Economics* **41**, 212-235.

Zellner (1962). Three-stage Least Squares: Simultaneous Estimation of Simultaneous Equations, Econometrica, vol. 30.

Zellner, A. (1962). "An Efficient Method of Estimating Seemingly Uncorrelated regressions and Tests for Aggregation Bias". Journal of the American Statistical Association, vol. 57, pp. 348-368.

Zellner, A. (1996): *An introduction to Bayesian inference in econometrics*, Wiley Classics Library. John Wiley & Sons Inc., New York, Reprint of the 1971 original, A Wiley-Interscience Publication.

Assignment File

Assignment files and marking scheme will be made available to you. This file presents you with details of the work you must submit to your tutor for marking. The marks you obtain from these assignments shall form part of your final mark for this course. Additional information on assignments will be found in the assignment file and later in this Course Guide in the section on assessment.

There are four assignments in this course. The four course assignments will cover:

Assignment 1 - All TMAs' question in Units 1 – 4 (Module 1)

Assignment 2 - All TMAs' question in Units 5 – 8 (Module 2)

Assignment 3 - All TMAs' question in Units 9 – 12 (Module 3)

Presentation Schedule

The presentation schedule included in your course materials gives you the important dates for this year for the completion of tutor-marking assignments and attending tutorials. Remember, you are required to submit all your assignments by due date. You should guide against falling behind in your work.

Assessment

There are two types of the assessment of the course. First are the tutor-marked assignments; second, there is a written examination.

In attempting the assignments, you are expected to apply information, knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal Assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignments File. The work you submit to your tutor for assessment will count for 30 % of your total course mark.

At the end of the course, you will need to sit for a final written examination of three hours' duration. This examination will also count for 70% of your total course mark.

Tutor-Marked Assignments (TMAs)

There are four tutor-marked assignments in this course. You will submit all the assignments. You are encouraged to work all the questions thoroughly. The TMAs constitute 30% of the total score.

Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and materials contained in your set books, reading and study units. However, it is desirable that you demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad viewpoint of the subject and also to give you a deeper understanding of the subject.

When you have completed each assignment, send it, together with a TMA form, to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given in the Presentation File. If for any reason, you cannot complete your work on time, contact your tutor before the assignment is due to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances.

Final Examination and Grading

The final examination will be of three hours' duration and have a value of 70% of the total course grade. The examination will consist of questions which reflect the types of self-assessment practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed

Revise the entire course material using the time between finishing the last unit in the module and that of sitting for the final examination to. You might find it useful to review your self-assessment exercises, tutor-marked assignments and comments on them before the examination. The final examination covers information from all parts of the course.

Course Marking Scheme

The Table presented below indicates the total marks (100%) allocation.

Assignment	**Marks**
Assignments (Best three assignments out of four that is marked)	30%
Final Examination	70%
Total	**100%**

Course Overview

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Applied Econometrics (ECO 713).

Units	**Title of Work**	**Week's Activities**	**Assessment (end of unit)**
	Course Guide		
MODULE 1:DEFINITON AND SCOPE OF ECONOMETRICS, REGRESSON ANALYSIS AND THE STATISTICAL TEST OF SIGNIFICANCE			
1	Definition and Scope of Econometrics	Week 1	Assignment 1
2	Correlation Analysis	Week 2	Assignment 2
3	Simple Regression Model and Statistical Test of Significance	Week 3	Assignment 3

4	Multiple Regression Model and Statistical Test of Significance	Week 4	Assignment 4
Module 2: ECONOMETRIC PROBLEMS, BASIC IDEAS OF THE IDENTIFICATION PROBLEM AND SIMULTENEOUS EQUATION ESTIMATION METHODS			
1	Econometric Problems (Heteroscedasticity, Autocorrelation and Multicollinearity)	Week 5	Assignment 1
2	Basic Ideas of the Identification Problem, Dummy variables and Distributed lag Models	Week 6	Assignment 2
3	Simultaneous Equation Estimation Methods (2SLS, 3SLS, etc)	Week 7	Assignment 3
4	Matrix treatment of Multiple Regression and Advanced treatment of Simultaneous Equation Estimation Techniques.	Week 8	Assignment 4
MODULE THREE:MATRIX TREATMENT OF REGRESSION ANALYSIS, TIME SERIES ECONOMETRICS			
1	Matrix Treatment of Multiple Regressions	Week 9	Assignment 1
2	Non-Stationarity and Unit Roots	Week 10	Assignment 2
3	Vector Auto Regressive (VAR Models	Week 11	Assignment 3
4	Cointegration and Error Correction Model	Week 12	Assignment 4

	Examination	Week 13, 14 & 15	

How To Get The Most From This Course

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace and at a time and place that suit you best.

Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to embark on discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit.

You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course and getting the best grade.

The main body of the unit guides you through the required reading from other sources. This will usually be either from your set books or from a readings section. Some units require you to undertake practical overview of historical events. You will be directed when you need to embark on discussion and guided through the tasks you must do.

The purpose of the practical overview of some certain historical economic issues are in twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience and skills to evaluate economic arguments, and understand the roles of history in guiding current economic policies and debates outside your studies. In any event, most of the critical thinking skills you will develop during studying are applicable in normal working practice, so it is important that you encounter them during your studies.

Self-assessments are interspersed throughout the units, and answers are given at the ends of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each

self-assessment exercises as you come to it in the study unit. Also, ensure to master some major historical dates and events during the course of studying the material.

The following is a practical strategy for working through the course. If you run into any trouble, consult your tutor. Remember that your tutor's job is to help you. When you need help, don't hesitate to call and ask your tutor to provide it.

1. Read this Course Guide thoroughly.
2. Organize a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.
3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.
5. Assemble the study materials. Information about what you need for a unit is given in the `Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.
6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.
7. Up-to-date course information will be continuously delivered to you at the study centre.
8. Work before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.
9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.
10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.
11. When you have submitted an assignment to your tutor for marking do not wait for it return `before starting on the next units. Keep to your schedule. When the

assignment is returned, pay particular attention to your tutor's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.

12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this Course Guide).

Tutors and Tutorials

There are some hours of tutorials (2-hours sessions) provided in support of this course. You will be notified of the dates, times and location of these tutorials. Together with the name and phone number of your tutor, as soon as you are allocated a tutorial group.

Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties you might encounter, and provide assistance to you during the course. You must mail your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

Do not hesitate to contact your tutor by telephone, e-mail, or discussion board if you need help. The following might be circumstances in which you would find help necessary. Contact your tutor if.

• You do not understand any part of the study units or the assigned readings
• You have difficulty with the self-assessment exercises
• You have a question or problem with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.

You should try your best to attend the tutorials. This is the only chance to have face to face contact with your tutor and to ask questions which are answered instantly. You can raise any problem encountered in the course of your study. To gain the maximum benefit from course tutorials, prepare a question list before attending them. You will learn a lot from participating in discussions actively.

Summary

The course, Applied Econometrics (ECO 713), exposes you to the analysis of quantitative economics by applying the knowledge of statistics and mathematics to quantify and solve economic relationships. The topics covered include definition and scope of econometrics, stages of econometric research. Regression analysis (Simple and multiple) and the statistical tests of significance, Econometric problems

(heteroscedasticity, autocorrelation, multicollinearity): their causes, detection, consequences and correction. Basic ideas of the identification problem, dummy variables, and distributed lags. Simultaneous equation estimation methods (2SLS, 3SLS, etc); Matrix treatment of multiple regressions; advanced treatment of the simultaneous equation estimation techniques. Instrumental variables, Time series econometrics, VAR models, Non-stationarity and unit roots, cointegration and error correction models (ECM); estimation and some tests of statistical hypotheses.

On successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective discussion on Environment and Sustainable Development: overview of theoretical perspectives of environment and development, introduction to the theories of managing common pool resources and their implications for sustainable development and the environmental challenges and types of resources.
However, to gain a lot from the course please try to apply anything you learn in the course to term papers writing in other economics courses. We wish you success with the course and hope that you will find it fascinating and handy.

MODULE 1: DEFINITON AND SCOPE OF ECONOMETRICS, REGRESSON ANALYSIS AND THE STATISTICAL TEST OF SIGNIFICANCE

Unit 1: Definition and Scope of Econometrics

Unit 2: Simple Regression Model

Unit 3: Multiple Regression Model

Unit 4: Statistical Test of Significance of Parameter Estimates

UNIT 1: DEFINITION AND SCOPE OF ECONOMETRICS

1.0 Introduction
2.0 Objectives
3.0 Main Contents

- 3.1. Definition of Econometrics
- 3.2. Scope/Division of Econometrics
- 3.3. Goals of Econometrics
- 3.4. Stages of Econometrics
- 3.5. Concept of Model: Economic Model and Econometric Model

4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

When the Nobel Memorial Prize in Economics Science was first awarded, in 1969, it was given to Ragnar Frisch and Jan Tinbergen of The Netherlands for their pioneering work in econometrics. At the time, few people had heard of the subject and even fewer knew much about it. Today econometrics is widely recognised as the primary tool of empirical economic analysis. Put simply, econometrics involves the development and the use of special statistical methods with mathematics within the framework that is consistent with the ways of economic inquiry. It is an extension of the field of statistics, which deals with techniques for collecting and analyzing data that arise in many different contexts.

Econometrics analysis starts from a statement about a behavioural relation. This statement, which may come from some sophisticated economic theory or from some plain reasoning, is then developed into an equation that specifies how the value of one variable is determined by the values of other variables.

2.0 OBJECTIVES

At the end of this unit, students are expected to:

- Define the term Econometrics
- Explain the scope/division of Econometrics
- State the objectives/goals of Econometrics
- Describe the stages of Econometrics

3.0 MAIN CONTENT

3.1. Definition of Econometrics

In a simple parlance, econometrics means economic measurement. According to Goldberg (1964), 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. Samuelson, Koopmans and Stone (1954) as captured by Gujarati (2006) defined econometrics as the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results. According to Brooks (2008), the literal meaning of econometrics is 'measurement in economics'. The first four letters of the word suggest correctly that the origins of econometrics are rooted in economics. In the words of Koutsoyannis (1977), econometrics deals with the measurement of economic relationships. It is a combination of economic theory, mathematical economics and statistics, but t is completely distinct from each one of these three branches of science. Koutsoyannis (1977) further stated that the following quotation from the opening editorial of Econometrica written by R. Frish in 1933 may give a clear idea of the scope and method of econometrics, thus:

But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what is called general economic theory, although a considerable portion of this theory has a definite quantitative character. Nor should econometrics be taken as synonymous with the application of mathematics to economics. Experience has shown that each of these three view points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. It's this unification that constitutes econometrics.

Thus econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationship such as elasticities, propensities, marginal values etc and verifying economic theories. It is a special type of economic analysis and research in which the general economic theory, formulated in mathematical terms, is combined with empirical measurement of economic phenomenon. Starting from the relationships of economic theory, we express them in mathematical terms (i.e. we build a model) so that they can be measured. We then use specific method, called econometric methods, in order to obtain numerical estimates of the coefficients of the economic relationships. Econometric methods are statistical methods specifically adapted to the peculiarities of economic phenomena. The most important characteristic of economic relationships is that they contain a random element, which, however, is ignored by economic theory and mathematical economics which postulate exact relationships between the various economic magnitudes. Econometrics has developed methods of dealing with the random element of economic relationships. For example, economic theory has it that the demand for a commodity is a function of its price, prices of other commodities, consumer's income and tastes. This kind of a function is exact because it implies that demand is wholly determined by the above four factors. No other factors except those explicitly mentioned, influences the demand Koutsoyannis (1977). In mathematical

economics, the above abstract economic relationship can be expressed in mathematical terms as follows:

$$Q^d = a_0 + a_1P + a_2Po + a_3Y + a_4t \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1.1)$$

Where Q^d = quantity demanded of a particular commodity

P = price of that particular commodity

Po = Prices of other commodity

Y = Consumer's income

t = tastes of the consumer

a_0, a_1, a_2, a_3, a_4 = the coefficients of the quantity demanded equation

The above demand equation is exact, because it implies that the only determinants of the quantity demanded are the four factors which appear in the right hand side of equation (1.1.1). Quantity will change only if some of these factors change. No other factor may have any effect on demand. Yet it is common knowledge that in economic life many more factors may affect demand. For example, the invention of a new product, a war, professional changes, institutional changes, changes in law, changes in income distribution, massive population movements (migration), etc., change quantity demanded of a commodity. In addition, human behaviour is erratic. We are influenced by rumours, dreams, prejudices, traditions and other psychological and sociological factors which make us behave differently even though the conditions in the market (prices) and our incomes remain the same. In econometrics the influence of these 'other' factors is taken into account by the introduction into the economic relationships of a random variable, with specific characteristics. In the previous example, the demand function studied with the tools of econometrics would be of the stochastic form:

$$Q^d = a_0 + a_1P + a_2Po + a_3Y + a_4t + u \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1.2)$$

Where u represents the random factors which affect the quantity demanded of the commodity.

It is important to emphasise that econometrics presupposes the existence of a body of economic theory. Economic theory should suffix first, because it sets the hypotheses about economic behaviour which should be tested with the application of econometric techniques.

SELF ASSESSMENT EXERCISE 1

1. In your own words, define the term econometrics
2. With example, differentiate between economic model and econometric model?

3.2 Scope of Econometrics

Econometrics may be classified into two main branches: Theoretical econometrics and applied econometrics.

- **Theoretical Econometrics:** This is concerned with the development of appropriate methods for the measurement of various economic relationships. Such methods could be: single equation technique, which is applied to one relationship at a time; and simultaneous equation techniques applied to all the relationships of a given model at once.
- Applied Econometrics: this other aspect of econometrics deals with the application of econometric methods to specific branches of economic theory. It examines various problems encountered and proffers solutions to such problems. Essentially, applied econometrics uses the tool of Theoretical econometrics in analyzing economic phenomenon as well as predicting economic behaviour.

SELF ASSESSMENT EXERCISE 2

Distinguish between theoretical and applied econometrics

3.3 Objectives of Econometrics

The three main objectives of econometrics are: (1) analysis, i.e. testing of economic theory; (2) policy-making, i.e. supplying numerical estimates of the coefficients of economic relationships, which may be then used for decision-making; (3) forecasting, i.e. using the numerical estimates of the coefficients in order to forecast the future values of the economic magnitudes Koutsoyannis (1977).

(1). Analysis: Testing Economic Theory

In the earlier stages of the development of economic theory, economists formulated the basic principles of the functioning of the economic systems using verbal exposition and applying a deductive procedure. The earlier economic theories started from a set of observations concerning the behaviour of individuals as consumers and producers. Some basic assumptions were set regarding the motivation of individual economic units.

Econometrics aims primarily at the verification of economic theories. Econometrics is a useful tool for structural analysis. From this, one can analyse intersectional relationships such as the business, household, financial or monetary sectors. 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-Making

Econometrics is aimed at bringing out alternatives for the process of decision-making. It is very vital for policy-making. The numerical estimates of the coefficients of the economic relationships are essential for the decision of firms as well as for the formulation government's economic policy.

For example, the decision of government about devaluing the currency will depend to a great extent on the numerical values of the price elasticities of exports and imports. If the sum of the price elaticities of exports and imports is less than one in absolute value, the devaluation will not help in eliminating the deficit in the balance of payment.

Such example show how important is the knowledge of the numerical values of the coefficients of the economic relationships. Econometrics can provide such sound economic policies.

(3). Forecasting the Future Values of Economic Magnitudes

Econometrics is useful in forecasting the future values of economic magnitudes. Such forecasts will enable policy makers to take necessary measures in order to influence the relevant economic variables. Indeed, predicting or forecasting has been cited as the prime contribution of econometrics.

For example, suppose that the government wants to decide its employment policy. It is necessary to know what is the current situation of employment as well as what level of employment will be, say, in three or six years' time, if no measure whatsoever is taken by the government. With econometric techniques, we may obtain such an estimate of the level of employment. Government can take appropriate measures to curb reoccurrence if the forecast value is higher or avoid inflation if the forecast value is low.

SELF ASSESSMENT EXERCISE 3

Vividly explain the objectives of econometrics.

3.4 Stages of Econometric Research

The subject matter of econometrics consists of the following stages:

1. Stage A. Model Specification

The first step in any econometric research is the specification of the model with which one will attempt the measurement of the phenomenon being analysed. This stage is also known as the formulation of the maintained hypothesis. This stage involves expressing

economic relationships between the given variables in mathematical form. Here, one needs to determine the dependent variable as well as the explanatory variable(s) which will be included in the model. Also expressed here is the apriori theoretical expectation regarding the sign and size of the parameters of the function as well as the nature of the mathematical form the model will take. Model specification, therefore, presupposes knowledge of economic theory and the familiarity with the particular phenomenon under investigation.

From the above sources of information the econometrician will be able to make a list of the variables (regressors) which might influence the dependent variable (regressand). Economic theory indicates the general factors which affect the dependent variable in any particular case. For example, suppose that the econometrician wants to study the demand for a particular commodity. The first source of her information is the static theory of demand which suggests that the determinants of the demand for any commodity are its price, the price of other commodities (mainly substitutes or complements), the level of the income of consumers, and their preferences. On the basis of this information, the demand function may be specified as follows:

$$Q^d_m = f(P_m, Po, Y, T)u \ldots\ldots (1.1.3)$$

Where Q^d_m = quantity demanded of commodity m

P_m = price of commodity m

Po = Prices of other commodity

Y = Consumers' income

t = preferences of the consumers

u = random variable

Equation (1.1.3) can be expressed in a more explicit form to contain the coefficients as follows:

$$Q^d_m = a_0 + a_1P_m + a_2Po + a_3Y + u \ldots\ldots (1.1.4)$$

The same source of knowledge – theory, other applied research and information about possible special features of the phenomenon being studied – will contain suggestions about the likely sign of the parameters and possibly of their size. For example, in equation (1.1.4), the parameter a_1 is expected to have a negative sign, given the 'law of demand' which postulates an inverse relationship between quantity demanded and its price. The parameter a_2 of the variable Po is expected to have a positive sign if commodity o is a substitute of commodity m, and a negative sign if the two commodities are complementary. The parameter a_3 related to the variable Y is expected to have a positive sign, since income and quantity demanded are positively related, except in the case of inferior goods.

2. Stage B: Estimation of the Model

Model estimation is the second stage in econometric research. It entails obtaining numerical estimates (values) of the coefficients of the specified model by means of appropriate econometrics techniques. This gives the model a precise form with appropriate signs of the parameters for easy analysis. In estimating the specified model, the following steps are important:

i. Data collection based on the variables included in the model. The data used in estimation of a model may be of various types. For example, time series data, cross-sectional data, panel data, engineering data, legislation and other institutional regulations, Data constructed by the econometrician: Dummy variable data etc.

ii. Examining the identification conditions of the model to ensure that the function one is estimation is the real function in question. Identification is the procedure by which we attempt to establish that the coefficients which we shall estimate by the application of some appropriate econometric technique are essentially the true coefficients of the function in which we are interested.

iii. Examining the aggregation problems of the function to avoid bias estimates. Aggregation problems arise from the fact that we use aggregative variables in

our function such aggregative variables may involve: aggregation over commodities, aggregation over time periods, spatial aggregation etc.

iv. Ensuring that the explanatory variables are not collinear, the situation which always results in misleading results. Most variables are correlated, in the sense the sense that they tend to change simultaneously during the various phases of economic activity. For example income, employment, consumption, investment, exports, imports, taxes, tend to grow n periods of booms and decline in periods of depression. Thus a certain degree of multicollinearity is inherent in the economic variables due to the growth and technological progress.

v. Appropriate methods should be adopted on the basis of the specified model.

The coefficients of economic relationships may be estimated by various methods which may be classified into two main groups:

i. Single-equation techniques. These are techniques which applied to one equation at a time. The most important are: the Classical Least Squares or Ordinary least Squares method, the Indirect Least Squares or Reduced form technique; Two Stage least Squares method, the Limited Information Maximum Likelihood method and various methods of mixed Estimation.

ii. Simultaneous-equation techniques. These are techniques which are applied to all equations of a system at once, and give estimates of the coefficients of all the functions simultaneously. The most important are the Three-stage Least Squares method and the Full Information Maximum Likelihood technique.

The choice of any technique in any particular case is a function of many factors, such as (a) The nature of the relationship and its identification condition. (b) The properties of the estimates of the coefficients obtained from each technique. (c) However, which of these desirable characteristics is the most important, depends on the purpose of the econometric research. (d) In some cases the simplicity of the method is used as a

criterion of choice: a method may be preferred to another because the first involves simpler computations and has less data requirements than the other. (e) The time and cost requirements of the various methods are often important criteria for the choice of the technique for the estimation of parameters of a model.

3. Stage C: Evaluation of Estimates

Evaluation entails assessing the results of the calculation in order to test their reliability. The results from the evaluation enable us to judge whether the estimates of the parameters are theoretically meaningful and statistically satisfactory. For this purpose we use various criteria which may be classified into three groups. (1) Economic apriori criteria, which are determined by economic theory and refer to the sign and size of the parameters of economic relationships (2) Statistical criteria, determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameter of the model. The most commonly used statistical criteria are the correlation coefficient and standard deviation (or standard error) of the estimates. (3) Econometric criteria otherwise known as the second-order tests, determined by econometric theory aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as second-order tests; in other words they determine the reliability of the statistical criteria, and in particular of the standard errors of the parameter estimates. They help to establish whether the estimates have the desirable properties of unbiasedness, consistency, etc.

4. Stage D: Evaluation of the Forecasting Power of the Estimated Model

Before the estimated model can be put to use, it is necessary to test its forecasting power. This will enable one to be assured of the stability of the estimates in terms of their sensitivity to changes in the size of the model even outside the given sample data, whose ‘average’ variation it represents.

A particular way of establishing the forecasting power of a model is to use the estimates of the model for a period not included in the sample. The estimated value (forecast value) s compared with the actual (realised) magnitude of the relevant

dependent variable. Another way of establishing the stability of the estimates and the performance of the model outside the sample of data from which it has been estimated is to re-estimate the function with an expanded sample that is a sample including additional observations.

SELF ASSESSMENT EXERCISE

Describe the stages of econometric research

3.5 Concept of Model: Economic Model and Econometric Model

A model is a simplified representation of a real world process. That is a prototype of reality, and so describes the way in which variables are interrelated (Akerele, 2002). These models exhibit the power of deductive reasoning in drawing conclusions relevant to economic policy.

Economic model describes the way in which economic variables are interrelated. Such model is built from various relationships between the given variables.

Econometric model on the other hand, consists of a system of equations which relate observable variables and unobservable random variables using a set of assumptions about the statistical properties of the random variables. In this respect, econometric model is built on the basis of economic theory.

Econometric model differs from econometric model in the following ways:

i. For an econometric model, its parameter can be estimated using appropriate econometric techniques.

ii. In formulation econometric model, it is usually necessary to decide the variables to be included or not. Thus, the variables here are selective, depending on the available statistical data.

iii. Because of the specific nature of econometric model, it allows fitting-in line of best fit, and this is not possible with economic model.

iv. The formulation of an econometric model involves the introduction of a random disturbance term. This will enable random element that are not accounted for to be taken care of.

The "goodness" of an econometric model is judged on the basis of the following various properties:

i. Conformity with economic theory. A good model should agree with the postulate of economic theory. It should describe precisely the economic phenomena to which it relates.

ii. Accuracy- the estimate of the coefficients should be accurate. They should approximate as best as possible the true parameter of the structural model.

iii. The model should possess explanatory ability. That is, it should be able to explain the observations of the real world.

iv. The model should be able to correctly predict future values of the dependent variable.

v. The mathematical form of the model should be simple with fewer equations. Such model should represent economic relationships with maximum simplicity.

vi. The equations of the model should be easily identified, that is, it most have a unique mathematical form.

SELF ASSESSMENT EXERCISE

Differentiate between an economic model and an econometric model.

4.0 CONCLUSION

We can see that this course would equip students understand the need to measure economic phenomena based on the application of statistic, mathematics and economic theory. This would help in achieving the goals of econometrics which include: analysis i.e. testing of economic theory, policy making, i.e. supplying

numerical estimates of the economic relationships, which may be then used for decision-making, and forecasting, i.e. using numerical estimates of the coefficients in order to forecast the future values of economic magnitudes.

SEFL ASSESSMENT EXERCISE 4

Discuss the methodology of econometric research.

5.0 SUMMARY

This unit has discussed the meaning of econometrics which involves the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories. The unit further elaborated the scope of econometrics which involves theoretical and applied econometrics and highlighted the goals of econometrics (analysis, policy-making and forecasting). The unit concluded with the methodology/stages of econometrics which include model specification, model estimation, model evaluation and evaluation of the forecasting validity of the model.

6.0 TUTORED-MARKED ASSIGNMENT

1. How do you perceive the roles of econometrics in analysis, decision/policy making and forecasting of economic phenomena?

2. Enumerate and explain the stages of econometric research you know.

7.0 REFERENCES/FURTHER READINDS

Akerele, A.A. (2002). Operations Research. Dimis Publications, Jos.

Allen, R.G.D. (1956). Mathematical Economics, Macmillan, London.

Brooks, C. (2008). Introductory Econometrics for Finance (Second Edition). Campbridge University Press, United Kingdom.

Klein, L.R. (1962). An Introduction to Econometrics, Prentice-Hall International, London, pp. 64–6 6, 86-87, 104-105.

Goldberger, A.S. (1964). Econometric Theory. Wiley, New York, P. 1.

Gujarati, D.N. (2006). Essentials of Econometrics (Third Edition). McGraw-Hill, New York. P. 1.

Koutsoyannis, A. (1977). Theory of Econometrics (Second Edition). PALGRAVE. New York.

Samuelson, P.A., Koopmans, T.C. & Stone, J.R.N. (1954). Report of the Evaluative Committee on Econometrica, Econometrica, Vol. 22, No. 2, PP. 141-146.

UNIT 2: SIMPLE REGRESSION MODEL

1.0 Introduction
2.0 Objectives
3.0 Main Contents
- 3.1 Definition of Simple Linear Regression
- 3.2 Assumptions of the Linear Stochastic Regression Model
- 3.3 The Least Squares Criterion and the 'Normal' Equations of OLS.
- 3.4 Ordinary Least Squares Estimators
- 3.5 Practical Aspect of Simple Regression in Econometric Software

4.0 Conclusion

5.0 Summary
6.0 Tutor-Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

In developing a model of economic phenomenon (e.g., the law of demand and supply) econometricians make heavy use of a statistical technique known as regression analysis. The purpose of this unit is to introduce the basics of regression analysis I terms of the simple and multiple linear regression model, namely, the two-variable model. Subsequent unit will consider various modifications and extensions of the two-variable model to the multiple-variable model.

2.0 OBJECTIVES

At the end of this unit, students are expected to:

4 Explain the meaning of simple regression model

5 Describe the assumptions of the linear stochastic regression model.

6 Discuss the Least Squares Criterion and the Normal Equations of OLS.

7 Evaluate the statistical test of significance of the Least Squares estimates.

3.0 MAIN CONTENT

3.1 Meaning of Simple Linear Regression Model

Regression analysis in general is concerned with the study of the relationship between one variable called the explained, or dependent, variable and one or more other variables called independent, or explanatory, variables (Gujarati, 2006).

Thus we may be interested in studying the relationship between the quantity demanded of a commodity in terms of the price of that commodity, income of the consumer, and prices of other commodities which could be complement or substitutes. We may also be interested in studying how sales of a product are related to advertisement expenditure incurred in that product. In both examples there may be some underlying theory that specifies why we would expect one variable to be dependent or related to one or more

other variables. In the first example, the law of demand provides the rationale for the dependence of the quantity demanded of a commodity on its price and several other factors. For notational uniformity, from here on we will le Y stand for the dependent variable and X the independent or explanatory variable Gujarati and Sangeetha, 2007).

Simple linear regression model therefore is a model which shows the relationship between two variables. In this relationship, one variable is depending on the other variable. The model consists of the independent variable and the constant term, with their respective coefficients and we need to estimate the parameters of the model in other to know the magnitude of their relationship.

Example

Consider a familiar supply function of the form:

$$Y = f(X) \qquad (1.2.1)$$

Where Y = quantity supplied

X = Price of the product

The theory of supply postulates that there exists a positive relationship between quantity supplied of a commodity and its price, ceteris paribus. Our first task is the specification of the supply model (equation 1.2.1), that is, the determination of the dependent (regressand or explained) and the independent (regressor or explanatory) variables, the number of equations and their precise mathematical form, finally the apriori expectations regarding the sign and size of the coefficients. Economic theory provides the following information with respect to the supply function.

1) The dependent variable is the quantity supplied and the explanatory variable s the price.

2) Economic theory does not specify the type of equation; therefore, we start our analysis with a single equation model.

3) Economic theory is not clear about the mathematical form of the model whether linear or nonlinear again we start by assuming a linear supply function:

$$Y_i = b_0 + b_1Xi \qquad (1.2.2)$$

This form implies that there is one-way causation between the variables Y and X: price is the cause of changes in quantity supplied, but not the other way around. The parameters of the supply function are b_0 and b_1, and our aim is to obtain estimates of their numerical values.

The parameters of this model are to be estimated using the Ordinary Least Squares (OLS). We shall employ this technique for a start due to the following reasons:

i. The computational procedure using this method is easy and straight forward.

ii. The mechanics of the OLS method is simple and understand.

iii. This method always produces satisfactory results.

iv. The parameter estimates using the OLS method best, linear and unbiased. This makes the estimates to be more accurate compared to the estimates obtained using other methods.

v. The OLS method is essential component of most econometric techniques.

Note that the model of equation (1.3.6) implies an exact relationship between Y and X. That is, all the variation in Y is due to changes in X only and no other factor(s) responsible for the change. When this is represented on a graph or scatter diagram, the pairs of observation (Y and X) would all lie on a straight line a shown in Figure 1.4.

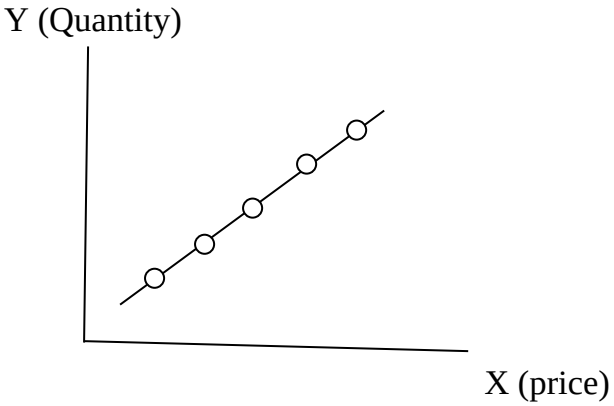


Figure 1.4: An exact relationship

Ideally, if we gather observations on the quantity actually supplied in the market at various prices and plot them on a diagram, we will notice that they do not really lie on

a straight line. Here, there are deviations of observations from the line. These deviations are attributable to the following factors:

1. Omission of variable(s) from the function on ground that some of these variables may not be known to be relevant.
2. Random behaviour of human beings. Human reactions at times are unpredictable and may cause deviation from the normal behavioural pattern depicted by the line.
3. Imperfect specification of the mathematical form of the model. A linear model, for instance, may mistakenly be formulated as a nonlinear model. It is also possible that some equations might have been left out in the model.
4. Error of aggregation. Usually, in model specification, we use aggregate data in which we add magnitudes relating to individuals whose behaviour differ.
5. Error of measurement. This error arises in the course of data collection, especially in the methods used in the collection of data.

The inclusion of a random variable usually denoted by U, into the econometric function helps in overcoming the above stated sources of errors.

This is so called because its introduction into the system disturbs the exact relationship which is assumed to exist between the Y and the X. Thus, the variation in Y could be explained in terms of explanatory variable, X and the random disturbance term, U. by introducing the random variable in the function, the model is rendered stochastic of the form:

$$Y_i = (b_0 + b_1X_i) + U_i \text{……..(1.2.3)}$$

Where Y_i = Variation in Y_i, $(b_0 + b_1X_i)$ = Systematic variation, U_i = random variation

Simply put: Variation in Y = Explained variation plus Unexplained variation

Thus, equation (1.2.3) is the true relationship that connects the variable Y and X. And this is our regression equation which we need to estimate its parameters using OLS method. To achieve this, we need observations on X, Y and U. However, U is not observed directly like any other variable.

SELF ASSESSMENT EXERCISE

Define the simple linear regression model and differentiate between economic model and econometric model.

3.2 Assumptions of the Linear Stochastic Regression Model

The linear regression model is based on certain assumptions, some of which refer to the distribution of the random variable U, some to the relationship between U and the explanatory variables, and finally some refer to the relationship between the explanatory variables themselves. We will group the assumptions into two categories: (a) stochastic assumptions, (b) other assumptions.

(a) Stochastic Assumptions of Ordinary least Squares

There are assumptions about the distribution of U. these assumptions are crucial for the estimates of the parameters.

i. U_i is a random variable. This means that the value which U_i takes in any one period depends on chance. Such values may be positive, negative or zero. For this assumption to hold, the omitted variables in the model should be numerous and should change in different directions.

ii. The mean value of "U_i" in any particular period is zero. That is, $E(U_i)$ denoted by $\bar{U}_i$ is zero. By this assumption, we may expressed equation (1.2.3) as:

$$Y_i = (b_0 + b_1X_i)\text{……..(1.2.4)}$$

iii. The variance of is U_i constant in each period. That is,

$Var(U_i) = E(U_i)^2 = \sigma(U_i)^2 = \sigma^2_{Ui,}$ which is zero. This implies that for all values of X, the U_i's will show the same dispersion about their mean. Violation of this assumption makes the U_i's heteroscedastic.

iv. U_i has a normal distribution. That is, a bell shaped symmetrical distribution about their zero mean. That is $U_i \sim N(0,\sigma^2_u)$.

v. The covariance of U_i and $U_j = 0$. $i \neq j$. this assumes the absence of autocorrelation among the U_i's. In this respect, the value of U_i in one period is not related to its value in another period.

(b) Other Assumptions

In terms of the relationship between U_i and the explanatory variables, the following assumptions also hold:

i. U and X do not covary. This means that there is no correlation between the disturbance term and the explanatory variable. Therefore, Cov (X_iU_i) = o

ii. The explanatory variables are measured without error. This is because the U absorbs any error of omission n the model.

In relation to the explanatory variable(s) alone, the following assumptions are made:

i. The explanatory variables are not linearly correlated. That is, there is absence of multicollinearity among the explanatory variables. This means Cov $X_iX_j = 0$, $i \neq j$ (This assumptions applies to multiple linear regression model).

ii. The explanatory variables are correctly aggregated. It is assumed that the correct procedure for such aggregate explanatory variables is used.

iii. The coefficients of the relationship to be estimated are assumed to have a unique mathematical form. That is, the variables are easily identified.

iv. The relationship to be estimated is correctly specified.

SELF ASSESSMENT EXERCISE

Describes the various assumptions of the linear stochastic regression model

3.3 The Least Squares Criterion and the Normal Equations of OLS.

Having specified the model and stated explicitly its assumptions in the previous unit, the next step is the estimation of the model, that is, the computation of the numerical values of its parameters.

- **Model Estimation**

The following procedure are used in finding numerical values of the parameters $\hat{b}_0$ and $\hat{b}_1$.

1. From the true relationship: $Y_i = (b_0 + b_1X_i) + U_i$ (1.2.3) and the estimated relationship: $\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_i + e_i$, the residual,

$$e_i = Y_i - \hat{Y}_i \text{ .....(1.2.4)}$$

and $e_i = Y_i - \hat{b}_0 - \hat{b}_1X_i$

2. Squaring the residuals and taking their sum gives:

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 or \sum e_i^2 = \sum (Y_i - \hat{b}_0 - b_1X_i)^2 \text{ .....(1.2.5)}$$

3. The expression in (1.2.5) is to be minimised with respect to $\hat{b}_0$ and $\hat{b}_1$ respectively.

Thus, $\frac{\partial \sum e_i^2}{\partial \sum \hat{b}_0} = 2\sum (Y_i - b_0 - b_1 X_i).(-1) = 0 = -2\sum (Y_i - b_0 - b_1 X_i) = 0$

$$\frac{\partial \sum e_i^2}{\partial \sum \hat{b}_1} = 2\sum (Y_i - b_0 - b_1 X_i).(-X_i) = 0 = -2X_i \sum (Y_i - b_0 - b_1 X_i) = 0$$

4. Dividing each term by -2, the OLS estimates of $\hat{b}_0$ and $\hat{b}_1$ could be written in the form:

$$\sum Y_i - \hat{b}_0 n - \hat{b}_1 \sum X_i = 0 \text{ …….....(1.2.6)}$$

$$\sum Y_i X_i - \hat{b}_0 \sum X_i - \hat{b}_1 \sum X_i^2 = 0 \text{ ……(1.2.7)}$$

The two equations (1.2.6) and (1.2.7) are the normal equations of the regression model.

5. Using the crammer's rule, the values of the parameters $\hat{b}_0$ and $\hat{b}_1$ are respectively:

$$\hat{b}_0 = \frac{\sum Y_i \sum X_i^2 - \sum X_i \sum Y_i X_i}{N\sum X_i^2 - \sum (X_i)^2}$$

……..(1.2.8)

$$\hat{b}_1 = \frac{N\sum Y_i X_i - \sum Y_i \sum X_i}{N\sum X_i^2 - \sum (X_i)^2}$$

……..(1.2.9)

Using lower case letters (i.e. deviation of the variables from their means), it can be shown that:

$$\hat{b}_0 = \bar{Y}_i - b_1 \bar{X}_i \text{ ……(1.2.10)}$$

$$\hat{b}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} \text{ ……..(1.2.11)}$$

Example

Given the following data on quantity supplied and price of a particular commodity, find the estimated supply function (Table 1.5).

N	1	2	3	4	5	6	7	8
Y_i(Quantity	64	68	44	48	50	65	45	56
X_i(Price)	8	10	6	9	6	10	7	8

Solution:

The expression for $\hat{b}_0$ and $\hat{b}_1$ in (1.2.8) and (1.2.9) as well as (1.2.10) and (1.2.11) lead us to produce the above table as seen in Table 1.6:

Y_i	X_i	X_i^2	X_iY_i	y_i	x_i	x_i^2	x_iy_i	$\hat{Y}$	e_i	e_i^2
64	8	64	512	9	0	0	0	55	9	81
68	10	100	680	13	2	4	26	64	4	16
44	6	36	264	-11	-2	4	22	46	-2	4
48	9	81	432	-7	1	1	-7	49.5	-11.5	132.25
50	6	36	300	-5	-2	4	10	46	4	16
65	10	100	650	10	2	4	20	64	1	1
45	7	49	315	-10	-1	1	10	50.5	-5.5	32.25
56	8	64	448	1	0	0	0	55	1	1
$\sum Y_i$=440	$\sum X_i$=64	$\sum X_i^2$=530	$\sum X_iY_i$ =3601	$\sum y_i$ = 0	$\sum x_i$ = 0	$\sum x_i^2$ =18	$\sum x_iy$ =81		$\sum e_i$ = 0	$\sum e_i^2$ = 281.5

Where e_i = Residual, $\hat{Y}_i = \hat{b}_0 + b_1 X_i$

From Table 1.6,

N= 8; $\bar{Y}_i = \frac{\sum Y_i}{n} = \frac{440}{8} = 55$; $\bar{X}_i = \frac{\sum X_i}{n} = \frac{64}{8} = 8$

Therefore, using the upper case letters:

$$\widehat{b}_0 = \frac{\sum Y_i \sum X_i^2 - \sum X_i \sum Y_i X_i}{N \sum X_i^2 - \sum (X_i)^2} = \frac{440(530) - (64)^2}{8(530) - (64)^2} = 19$$

$$\widehat{b}_1 = \frac{N \sum Y_i X_i - \sum Y_i \sum X_i}{N \sum X_i^2 - \sum (X_i)^2} = \frac{8(3601) - (64)(440)}{8(530) - (64)^2} = 4.5$$

Similarly, using the lower case letters:

$$\widehat{b}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{81}{18} = 4.5$$

$$\widehat{b}_0 = \bar{Y}_i - b_1 \bar{X}_i = 55 - 4.5(8) = 19$$

Note: (i) Only one of the two methods is to be used, and each gives the same result

(ii) Unless specified, one is free to use any of the methods.

From the values of $\widehat{b}_0$ and $\widehat{b}_1$, the estimated regression line or equation is gotten by substituting these values into $\widehat{Y}_i = \widehat{b}_0 + b_1 X_i$ and this gives:

$$\widehat{Y}_i = 19 + 4.5X_i \text{ ………………………………………………….(1.2.12)}$$

Thus, given the values of Xi (I = 1, 2 … N), the estimated values of Y can be obtained using the regression equation.

From the estimated regression line, one can estimate ***price elasticity***. Recall the estimated model, $\widehat{Y}_i = \widehat{b}_0 + b_1 X_i$. This is also the equation of the line with intercept, $\widehat{b}_0$ and slope, $\widehat{b}_1$.

Note that $\widehat{b}_1 = \partial Y_i / \partial X_i$. Therefore, price elasticity (e_p) = $\frac{\widehat{b}_1 \bar{X}_i}{\bar{Y}_i}$

Taking the mean of Xi and Yi, we have average elasticity:

$$e_p = \frac{\widehat{b}_1 \bar{X}_i}{\bar{Y}_i} = \frac{4.5(8)}{55} = 0.65 \text{ (Inelastic)}$$

SELF ASSESSMENT EXERCISE

Given a simple linear supply stochastic function, $Y_i = b_0 + b_1X_i + U_i$, derive the least squares estimates.

3.5 Ordinary Least Squares Estimators

The ordinary least squares (OLS) principle is among the best method used in obtaining estimates of parameter, $\hat{b_i}$.

The use of OLS method in estimating economic relationship is based on the fact that the estimates of the parameters have some optimal properties.

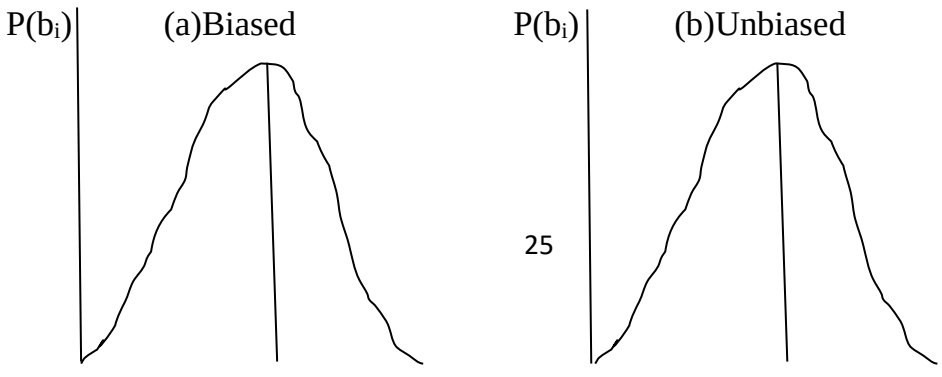
Generally, in choosing a particular method, one should aim at such method that gives an estimate, which (if at all it exists) will be within only a small range around the true parameter.

For any estimation method, the goodness of the estimator is judge on the basis of the following desirable properties:

i. Unbiasedness

An estimator is unbiased if its bias is zero, i.e. $E(\hat{b_i}) - b_i = 0$. In this case, the unbiased estimator changes to the true value of the parameter as the number of sample increases.

An unbiased estimator always gives, on the average, the true value of the parameter. The case of biased and unbiased estimator of the true parameter is illustrated diagrammatically:



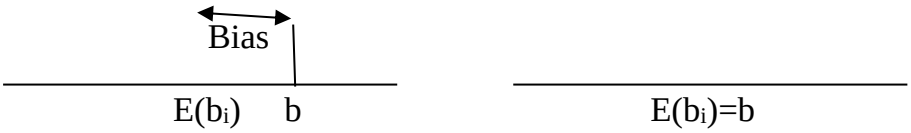


Figure 1.4: Biased and Unbiased estimator

ii. Minimum Variance

An estimator is best if it has the smallest variance compared with any other estimate obtained using other methods. By minimum variance, we mean that the values of the parameter bi clusters very closely around the true parameter b.

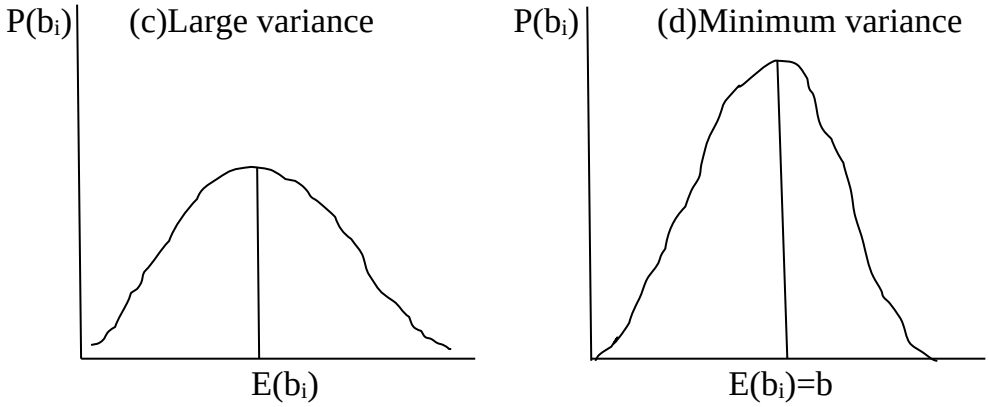


Figure 1.5: Large and Minimum Variance

In diagrams of Figure 1.5, panel (d) is best compared with (c) because (d) has minimum variance as seen in the narrowness of the distribution (Goldberger, 1964).

iii. Efficient Estimator

An estimator is efficient when it combines the property of unbiasedness and minimum variance property.

Symbolically, b_i is efficient if the following two conditions are fulfilled:

(a) $E(\hat{b}) = b$ (b) $E[(\hat{b}) - E(b)]^2 < E[b^* - E(b^*)]^2$

Where b* is another unbiased estimate of the true b.

This means that in the class of unbiased estimators, such estimator has a minimum variance

iv. Linear Estimator

An estimator is linear if it is a combination of the given sample data. Thus, with the sample observation, Y_1, Y_2,…Yn, a linear estimator take the form:

$k_1Y_1 + k_2Y_2+\ldots+k_nY_n$, where k is some constants.

v. BLUE (Best, Linear, Unbiased Estimator)

This property is abbreviated to BLUE meaning that the estimator is best (having minimum variance), linear, and unbiased as compared with all other linear, unbiased estimators. Thus, all the properties (i-iv) are included n the BLUE property.

vi. Minimum Mean Square Error (MSE) Estimator

This property combines unbiasedness and minimum variance properties. An estimator, therefore, is a minimum MSE estimator if it has the smallest mean square error, defined as the expected value of the squared difference of the estimator around the true population parameter, b. that is, $MSE(\hat{b}) = \sum[\hat{b} - b]^2$.

vii. Sufficiency

This property implies that the estimator uses all the available information a sample contains about the true parameter.

For this property to hold, the estimator should accommodate all the observations of he sample, and should not give room for any additional information in connection with the true population parameter.

The OLS method satisfies the above stated properties. For this reason, the method seems to be the best and most widely used of all the estimation methods. In a nutshell, the OLS has the BLUE (best, linear, unbiased properties) among the class of linear and unbiased estimators. The linearity property as previously discussed implies that the parameter estimates are linear functions of the observed Y_i. That is, the estimates $\hat{b}_0$ and $\hat{b}_1$ includes the variable Y and X in the first power.

Thus, $\hat{b}_1$ = f(Y). This property enables one to compute the values of the parameter estimates with ease.

The unbiased property of OLS estimates implies that the expected value of the estimated parameter is equal to the true value of the parameter. That is, $E(\hat{b}_i) = b_i$.

The importance of this property lies in the fact that for large samples, the parameter estimates obtained will on the average give a true value of the b's.

The minimum variance property becomes desirable when combined with unbiasedness. The importance of this property is obvious when we want to apply the standard error test of significance for b_0 and b_1, and to construct confidence intervals for these estimates. Because of minimum variance they have, their respective confidence intervals will be narrower than for other estimates obtained using any other econometric procedures.

The smaller confidence interval obtained is interpreted to mean in effect that we are extracting more information from our sample than we would be, if we were to use any other methods which yielded the same unbiased estimates.

SELF ASSESSMENT EXERCISE

Discuss the properties of OLS

3.6 Practical Aspect of Simple Regression in Econometric Software

Simple Regression in EViews

Step 1: Open EViews.

Step 2: Click on File/New/Workfile in order to create a new file.

Step 3: Choose the frequency of the data in the case of time series data or Undated or Irregular in the case of cross-sectional data, and specify the start and end of your data set. EViews will open a new window which automatically contains a constant (c) and a residual (resid) series.

Step 4: On the command line type:

genr x = 0 (press enter)

genr y = 0 (press enter)

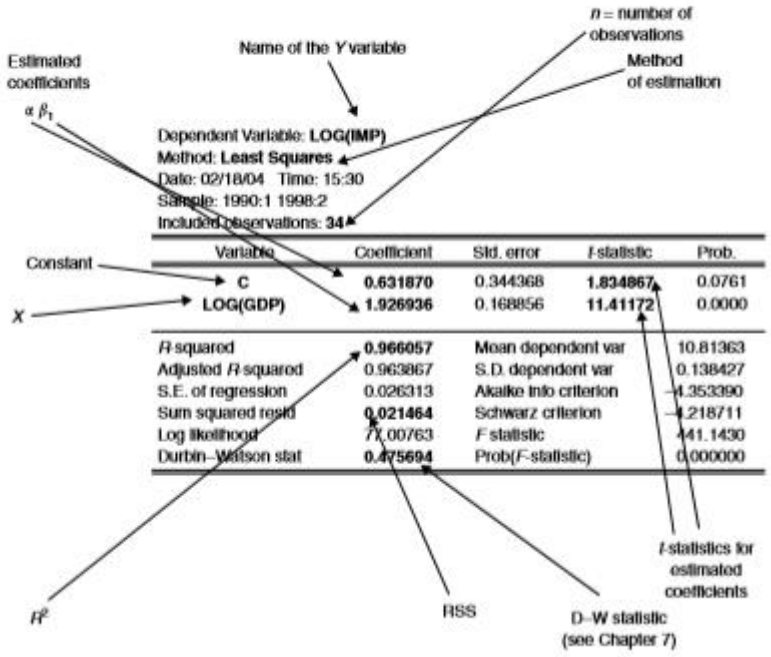
which creates two new series named x and y that contain zeros for every observation. Open x and y as a group by selecting them and double-clicking with your mouse.

Step 5: Then either type the data into EViews or copy/paste the data from Excel. To be able to type (edit) the data of your series or to paste anything into the EViews cells, the edit +/− button must be pressed. After editing the series press the edit +/−button again to lock or secure the data.

Step 6: Once the data have been entered into EViews, the regression line (to obtain alpha and beta) may be estimated either by typing:

ls y c x(press enter) on the command line, or by clicking on Quick/Estimate equation and then writing your equation (that is ycx) in the new window. Note that the option for OLS (LS Least Squares (NLS and ARMA)) is chosen automatically by EViews and the sample is automatically selected to be the maximum possible. Either way, the regression result is shown in a new window which provides estimates for alpha (the coefficient of the constant term), beta (the coefficient of X) and some additional statistics that will be discussed in later sections of this material.

Reading the EViews simple regression results output



Presentation of regression results

The results of a regression analysis can be presented in a range of different ways. However, the most common way is to write the estimated equation with standard errors of the coefficients in brackets below the estimated coefficients and to include further statistics below the equation. For the consumption function that will be presented in Computer Example 1, the results are summarized as shown below:

$$\hat{C} = 15.116 + 0.116\ Y_t^d$$

(6.565) (0.038)

$R^2 = 0.932$ n = 20 $\sigma = 6.879$

From this summary we can (a)read estimated effects of changes in the explanatory variables on the dependent variable; (b) predict values of the dependent variable for given values of the explanatory variable; (c) perform hypothesis testing for the estimated coefficients; and (d) construct confidence intervals for the estimated coefficients.

Computer example: the Keynesian consumption function

Table 1.7 provides data for consumption and disposable income for 20 randomly selected people.

S/N	1	2	3	4	5	6	7	8	9	10
Consumption (Y)	72.3	91.65	135.2	94.6	163.5	100	86.5	42.36	120	112.56
Disposable Income (X)	100	120	200	130	240	114	126	213	156	167
S/N	11	12	13	14	15	16	17	18	19	20
Consumption (Y)	132.3	149.8	115.3	132.2	149.5	100.25	79.6	90.2	116.5	126

Disposable Income (X)	189	214	188	197	206	142	112	134	169	179

Use Eviews to calculate α and β.

Solution: To obtain regression results in EViews, the following steps are required:

1. Open EViews.

2. Choose File/New/Workfile in order to create a new file.

3. Choose Undated or Irregular and specify the number of observations (in this case 20). A new window appears which automatically contains a constant (c) and a residual (resid) series.

4. In the command line type:

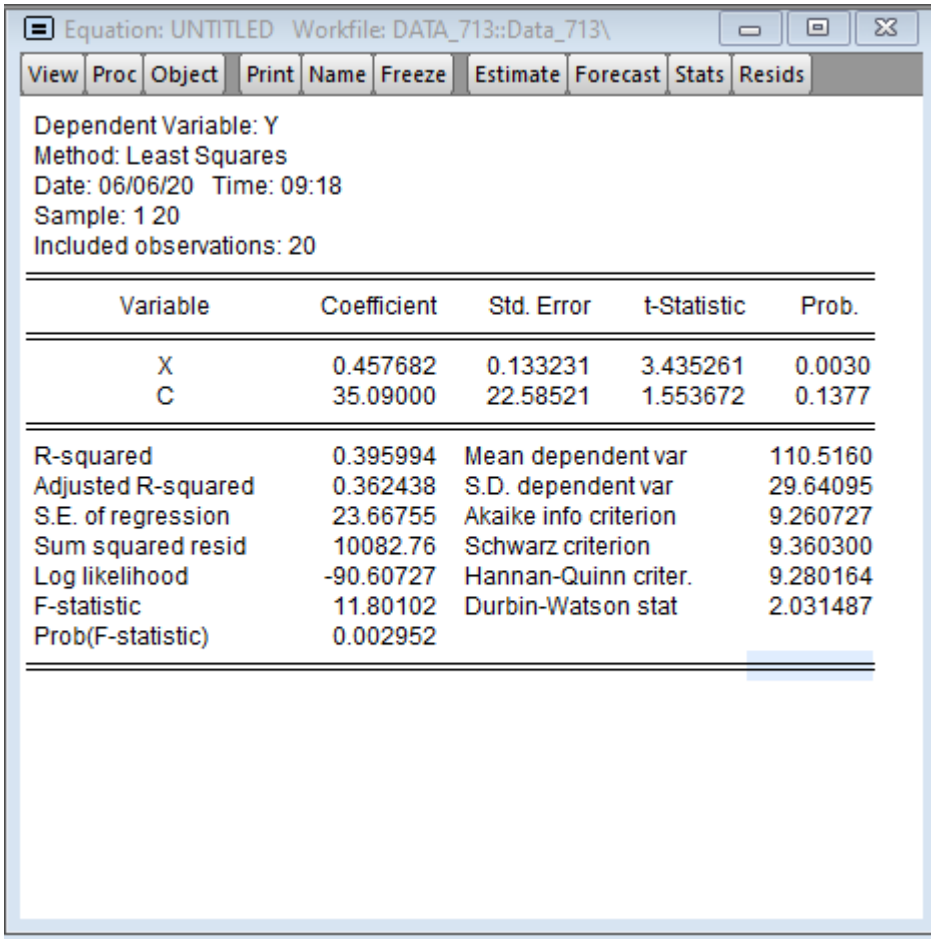
genr x = 0 (press enter) genr y = 0 (press enter) which creates two new series named x and y that contain zeros for every observation. Open x and y as a group by selecting them and double-clicking with the mouse.

5. Either type the data into EViews or copy/paste the data from Excel. To edit the series press the edit +/− button. After you have finished editing the series press the edit +/−button again to lock or secure the data.

6 After entering the data into EViews, the regression line (to obtain alpha and beta) can be estimated either by writing:

ls y c x(press enter)

on the EViews command line, or by clicking on Quick/Estimate equation and then writing the equation (that is y c x) in the new window. Note that the option for OLS (LS – Least Squares (NLS and ARMA)) is chosen automatically by EViews and the sample is automatically selected to be from 1 to 20. Either way, the output in Table 1.7 is shown in a new window which provides estimates for alpha (the coefficient of the constant term) and beta (the coefficient of X).



SELF ASSESSMENT EXERCISE

The following data in Table 18 refer to the quantity sold of good Y (measured in kg), and the price of that good X (measured in naira per kg), for 10 different market locations:

Y	198	181	170	179	163	145	167	203	251	147
X	23	24.5	24	27.2	27	24.4	24.7	22.1	21	25

(a) Assuming a linear relationship between the two variables, use Eviews software, obtain the OLS estimators of α and β. (b) On a scatter diagram of the data, and draw your OLS sample regression line. (c) Estimate the elasticity of demand for this good at the point of the sample means (that is when Y = $\bar{Y}$ and X = $\bar{X}$).

4.0 CONCLUSION

This unit throws light on the meaning of simple linear regression model as a model which shows the relationship between two variables with one of the variable called the dependent viable while the other one called the independent or explanatory variables. The unit concludes that the parameter estimates (b_0 and b_1) of the regression model can be estimated with the use of an econometric technique called the ordinary least squares (OLS) because the method is best, linear and unbiased. As a result makes the parameter estimates to be accurate compared to the estimates obtained from other methods.

5.0 SUMMARY

This unit discusses the meaning of simple linear regression model and assumptions of the linear stochastic regression model such as stochastic assumptions regarding the error term and other assumptions regarding the random term and explanatory variable(s) as well as regarding the explanatory variables themselves. The unit further discusses the least squares criteria and the least squares normal equation where the parameters b_0 and b_1 can be estimated using the OLS which is adjudged to be best, linear and unbiased. Statistical test of significance such as the coefficient of determination (R^2), standard error test, t-test, confidence intervals were also discussed. The unit rounds up with the explanation of the properties of the OLS such as unbiasedness, minimum variance, efficient estimator, linear estimator, BLUE, minimum mean square error and sufficiency.

6.0 TUTOR-MARKED ASSIGNMENT

1) The following table includes the price and quantity demanded of the product of a monopolist over a six-year period.

Table 1.7: Price and quantity demanded of a given Monopolist's product

Year	2014	2015	2016	2017	2018	2019

Quantity (Y)	80	30	40	70	80	10
Price (N)	20	40	30	10	30	50

(a) Using both manual and Eviews computer based, estimate the demand function for the monopolist's product. Use economic theory to comment on the values of the estimated parameters ($\hat{b}_0$ and $\hat{b}_1$).

(b) Estimate the average elasticity of demand.

2) Distinguish carefully between the following concepts:

a) The true relationship between X and Y.

b) The true regression line.

c) The estimated relationship.

d) The estimated regression line.

e) What assumptions do we normally make about the random term, U? Why are these assumptions necessary?

3) The following table includes the gross domestic product (X) and demand for food respectively measured in thousand naira and tons in Nigeria over the ten-year period, 2010-19.

Table 1.8: GDP and demand for food in Nigeria, 2010-2019

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Y	6	7	8	10	8	9	10	9	11	10
X	50	52	55	59	57	58	62	65	68	70

(a) Use both manual and Eviews computer base, estimate the food function

$Y = b_0 + b_1X + u$

What is the economic meaning of your result?

(b) Compute the coefficient of determination and interpret your result.

(c) Compute the standard errors of the regression estimates and conduct tests of significance at the 5 percent level of significance.

(d) Find the 99 percent confidence interval for the population parameters.

(e) Obtain annual data for the inflation rate and the unemployment rate of your country.

(i) Use Eviews to estimate the following regression, which is known as the Phillips curve: $\pi_t = a_0 + a_1UNEMP_t + u_t$

where π_t is inflation and UNEMPt is unemployment. Present the results in the usual way.

(ii) Estimate the alternative model:

$\pi_t - \pi_{t-1} = a_0 + a_1UNEMP_{t-1} + u_t$ and calculate the Non-Accelerating Inflation Rate of Unemployment (NAIRU) (that is when $\pi_t - \pi_{t-1} = 0$).

(iii) Re-estimate the above equations splitting your sample into different decades. What factors account for differences in the results? Which period has the 'best-fitting' equation? State the criteria you have used.

7.0 REFERENCES/FURTHER READINGS

Dimitrios Asteriou and Stephen G. Hall (2011). Applied Econometrics (Second Edition). PALGRAVE MACMILLAN, UK.

Goldberger, A.S. (1964). Econometric Theory. Wiley, New York.

Gujarati, D.N. (2006). Essentials of Econometrics (Third Edition). McGraw-Hill, New York.

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

Koutsoyannis, A. (1977). Theory of Econometrics (Second Edition). PALGRAVE. New York.

UNIT 3: MULTIPLE REGREEIONS MODEL

1.0 Introduction
2.0 Objectives
3.0 Main Contents
- 3.1 Models with two explanatory variables.
- 3.2 The coefficient of multiple determinations and the adjusted coefficient of multiple determinations.
- 3.3 The mean and variance of parameter estimates ($\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$)
- 3.4 Test the statistical significance of the parameter estimates
- 3.5 Practical Applications of Multiple Regression in Econometric Soft wares

4.0 Conclusion

5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

The two variable model studied in the previous unit is often inadequate in practice. Most economic relations, and the processes they describe, involve more than one determinant of some particular dependent variable. In consumption-income example, for instance, it is assumed implicitly that only income, X affects consumption, Y. But economic theory is seldom so simple for, besides income, a number of other variables are also likely to affect consumption expenditure. An obvious example is the wealth of the consumer. Another example is the demand for a commodity which is likely to depend not only on its own price but also on the prices of other competing or complementary goods, income of the consumer, social status, etc. therefore, we need to extend our two-variable regression model to cover models involving more than two variables. Adding more variables leads us to the discussion of multiple regression models, that is, models in which the dependent variable, or the regressand, Y, depends on two or more explanatory variables, or regressors.

2.0 OBJECTIVES

At the end of this unit, students are expected to:

8.0 Illustrate models with two explanatory variables.

9.0 Derive the normal equation of two explanatory variables.

10.0 Estimate the coefficient of multiple determinations and the adjusted coefficient of multiple determinations.

11.0 Calculate the mean and variance of parameter estimates ($\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$)

12.0 Test the statistical significance of the parameter estimates

3.0 MAIN CONTENT

3.1 Two Explanatory Variables Models

As a first step in learning about multiple regressions, we consider an economic process in which the variable Y is determined by two variables, X_1 and X_2 (Mirer, 1995). We shall illustrate the three-variable model with an example from the theory of demand. Economic theory postulates that the quantity demanded for a given commodity (Y) depends on its price (X_1) and the consumer's income (X_2):

$$Y = f(X_1, X_2) \quad \text{(1.3.1)}$$

Given that the theory does not specify the mathematical form of the demand function, we start our investigation by assuming a linear relationship between Y, X_1 and X_2.

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} \quad (i = 1,2,\ldots,n) \quad \text{(1.3.2)}$$

Equation (1.3.2) shows an exact relation in the sense that variations in quantity demanded are wholly explained by variations in price and income. If this claim is true, then any observation on Y, X_1 and X_2 would determined a point which would lie on a plane. However, if we gather observations on these variables during a certain period of time and plot them on a diagram, we will notice that not all the observations lie on the plane: some will lie on it, but others will lie above or below t. this scatter is due to various factors omitted from the function and to other types of error which have been examined in the previous unit. The influence of such factors may be taken into account by introducing a random variable u, n the function, which thus becomes stochastic:

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i}) + (u_i) \quad \text{(1.3.3)}$$

Systematic component   *random component*

Economic theory suggests that the coefficient $\hat{b}_1$ should have negative sign, given the 'law of demand', while $\hat{b}_2$ is expected to be positive, since for normal commodities the quantity demanded changes in the same direction as income (Koutsayiannis, 1977).

In equation (1.3.3), b_0 is the intercept term. As usual, it gives the mean or average effect on Y of all the variables excluded from the model, although its equal to zero. The coefficients $\hat{b}_1$ and $\hat{b}_2$ are called the partial regression coefficients (Gujarati & Sangeetha, 2007).

In order to complete the specification of our model, we continue to operate within the framework of the classical linear regression model (CLRM) introduced in the previous unit.

3.1.1 Assumptions with Respect to the Three-Variable Regression Model

Some certain assumptions are needed to complete the formulation of the model. Some of the assumptions are about the random term u. these assumptions are the same as in a two-variable regression model discussed in the previous unit. That is:

1. Randomness of u

The variable u is a real random variable.

2. Zero mean value of the random variable u, or

$$E(u_i \mid X_{1i}, X_{2i}) = 0$$

3. No serial correlation of the u's or

$$Cov\ (u_i, u_j) = 0\ i \neq j$$

4. Homoscedasticity or the variance of each u_i is the same for all the X_i values

$$Var(u_i) = E(u_i^2) = \sigma^2$$

5. Normality of u

$$ui \sim (0, \sigma_u^2)$$

6. Independence of u_i and X_i

   Every disturbance term u_i is independent of the explanatory variables

   $$E(u_i X_{1i}) = E(u_i X_{2i}) = 0$$

   This condition is automatically met if we assume that the values of the X's are a set of fixed numbers in all assumed samples.

7. No errors of measurement in the X's.

   The explanatory variables are measured without error.

8. No exact linear relationship between X_1 and X_2.

The explanatory variables are not perfectly correlated. We assume that the multiple regression is linear in the parameters that the values of the regressors are fixed in repeated sampling, and that there is sufficient variability in the values of the regressors.

9. Correct aggregation of the macro-variables.

   The appropriate 'aggregation bridge' has been constructed between the aggregate macro-variables used n the function and their individual components (micro-variables).

10. Identifiability of the model/function.

    The relationship under study is fully identified and has a unique mathematical formation.

11. Correct specification of the model.

    All the explanatory variables appear explicitly in the function and the mathematical form s correctly defined (linear or nonlinear form and the number of equations in the model). Therefore, the model has no specification error.

3.1.2 (a) Interpretation of Multiple Regression Model

Given the assumptions of the least squares regression model, it follows that on taking the conditional expectation of Y on both sides of equation (1.2.3), we obtained:

$$E(Y_i \mid X_{1i}, X_{2i}) = b_0 + b_1X_{1i} + b_2X_{2i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.3.4)$$

In words, (1.2.4) gives the conditional mean or expected value of Y conditional upon the given or fixed values of X_1 and X_2. Therefore, as in the two-variable case, multiple regression analysis is regression analysis conditional upon the fixed values of the regressors, and what we obtain is the average or mean response of Y for the given values of the regressors.

(b) **The Meaning of Partial Regression Coefficients**

The regression coefficients X_1 and X_2 are known as partial regression or partial slope coefficients. The meaning of partial regression coefficient is as follows: b_1 measures the change in the mean value of Y, E(Y), per unit change in X_1, holding the value of X_2 constant. Put differently, it gives the "direct" or the "net" effect of a unit change in X_1 on the mean value of Y, net of any effect that X_2 may have on mean Y. Likewise, b_2 measures the change in the mean value of Y per unit change in X_2, controlling for the value of X_1 constant. That is, it gives the "direct" or "net" effect of a unit change in X_2 on the mean value of Y, net of any effect that X_1 may have on mean Y.

Having specified our model (1.3.4), we next use sample observations on Y, X_1 and X_2 and obtain estimates of the true parameters b_0, b_1 and b_2:

$$\widehat{Y}_i = \widehat{b}_0 + \widehat{b}_1 X_{1i} + \widehat{b}_2 X_{2i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.3.5)$$

Where $\widehat{b}_0$, $\widehat{b}_1$ and $\widehat{b}_2$ are estimates of the true parameters b_0, b_1 and b_2 of the demand relationship.

As n the case of two-variable regression mode, the estimates will be obtained by minimizing the sum of squared residuals.

$$\sum_{i=1}^{n} e^2 = \sum_{i}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i}^{n} (Y_i - b_0 - b_1 X_{1i} - b_1 X_{2i})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.3.6)$$

A necessary condition for this expression to assume a minimum value is that its partial derivatives with respect to $\hat{b}_0$, $\hat{b}_1$ and $\hat{b}_2$ be equal to zero:

1. Partial derivative with respect to $\hat{b}_0$:

$$\left.\begin{aligned} &\frac{\partial \sum e^2}{\partial \hat{b}_0} = 2\sum (Y_i - b_0 - b_1 X_{1i} - b_1 X_{2i})(-1) = 0 \\ &-2\sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_1 X_{2i}) = 0 \\ &\sum Y_i - \sum \hat{b}_0 - \sum \hat{b}_1 X_{1i} - \sum \hat{b}_2 X_{2i} = 0 \\ &\sum Y_i = n\hat{b}_0 + \sum \hat{b}_1 X_{1i} + \hat{b}_2 \sum X_{2i} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

2. Partial derivative with respect to $\hat{b}_1$

$$\left.\begin{aligned} &\frac{\partial \sum e^2}{\partial \hat{b}_1} = 2\sum (Y_i - b_0 - b_1 X_{1i} - b_1 X_{2i})(-X_{1i}) = 0 \\ &-2\sum X_{1i}(Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_1 X_{2i}) = 0 \\ &\sum X_{1i} Y_i - \sum \hat{b}_0 X_{1i} - \sum \hat{b}_1 {X_{1i}}^2 - \sum \hat{b}_2 X_{1i} X_{2i} = 0 \\ &\sum X_{1i} Y_i = \hat{b}_0 \sum X_{1i} + \hat{b}_1 \sum {X_{1i}}^2 + \hat{b}_{2i} \sum X_{1i} X_{2i} \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

3. Partial derivative with respect to $\hat{b}_2$

$$\left.\begin{aligned} &\frac{\partial \sum e^2}{\partial \hat{b}_2} = 2\sum (Y_i - b_0 - b_1 X_{1i} - b_1 X_{2i})(-X_{2i}) = 0 \\ &-2\sum X_{2i}(Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_1 X_{2i}) = 0 \\ &\sum X_{2i} Y_i - \sum \hat{b}_0 X_{2i} - \sum \hat{b}_1 X_{1i} X_{2i} - \sum \hat{b}_2 {X_{2i}}^2 = 0 \\ &\sum X_{2i} Y_i = \hat{b}_0 \sum X_{2i} + \hat{b}_1 \sum X_{1i} X_{2i} + \hat{b}_{2i} \sum {X_{2i}}^2 \end{aligned}\right\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

From equation (1), divide through by n to estimate the mean of the equation to solve for $\hat{b}_0$:

$$\frac{\sum Y_i}{n} = \frac{n\hat{b}_0}{n} + b_1\frac{\sum X_{1i}}{n} + \hat{b}_2\frac{\sum X_{2i}}{n}$$

$$\bar{Y}_i = \hat{b}_0 + \hat{b}_1\bar{X}_{1i} + \hat{b}_2\bar{X}_{2i} \quad \text{.....(1.3.7)}$$

$$\hat{b}_0 = \bar{Y}_i - \hat{b}_1\bar{X}_{1i} - \hat{b}_2\bar{X}_{2i}$$

The other identities whose values are still unknown ($\hat{b}_1$ and $\hat{b}_2$) can be determined. The first rule to determine these unknown is to convert the upper case variables into lower case or deviations from the sample mean.

Secondly, the crammer's rule of matrix is applied to the deviation values algebraically to determine the solution equation for the slope of $\hat{b}_1$ and $\hat{b}_2$.

$$\left.\begin{aligned}\sum x_{1i}y_i &= \hat{b}_1\sum x_{1i}^{\,2} + \hat{b}_{2i}\sum x_{1i}x_{2i}\\ \sum x_{2i}y_i &= \hat{b}_1\sum x_{1i}x_{2i} + \hat{b}_{2i}\sum x_{2i}^{\,2}\end{aligned}\right\} \text{.....(1.3.8)}$$

Where $y_i = Y_i - \bar{Y}$, $x_{1i} = X_{1i} - \bar{X}_1$ and $x_{2i} = X_{2i} - \bar{X}_2$

Applying crammer's rule:

$$\begin{vmatrix}\sum x_{1i}^{\,2} & \sum x_{1i}x_{2i}\\ \sum x_{1i}x_{2i} & \sum x_{2i}^{\,2}\end{vmatrix}\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\end{bmatrix} = \begin{bmatrix}\sum x_{1i}y_i\\ \sum x_{2i}y_i\end{bmatrix} \text{.....(1.3.9)}$$

$$\Delta = \begin{vmatrix}\sum x_{1i}^{\,2} & \sum x_{1i}x_{2i}\\ \sum x_{1i}x_{2i} & \sum x_{2i}^{\,2}\end{vmatrix} \quad \text{.....(1.3.10)}$$

$$\Delta = \sum x_{1i}^{\,2}\sum x_{2i}^{\,2} - (\sum x_{1i}x_{2i})^2$$

Solve for $\hat{b}_1$

$$\Delta\hat{b}_1 = \begin{vmatrix}\sum x_{1i}y_i & \sum x_{1i}x_{2i}\\ \sum x_{2i}y_i & \sum x_{2i}^{\,2}\end{vmatrix}$$

$$\Delta\hat{b}_1 = \sum x_{1i}y_i\sum x_{2i}^{\,2} - (\sum x_{2i}y_i)(\sum x_{1i}x_{2i}) \quad \text{.....(1.3.11)}$$

$$\hat{b}_1 = \frac{\Delta\hat{b}_1}{\Delta} = \frac{\sum x_{1i}y_i\sum x_{2i}^{\,2} - (\sum x_{2i}y_i)(\sum x_{1i}x_{2i})}{\sum x_{1i}^{\,2}\sum x_{2i}^{\,2} - (\sum x_{1i}x_{2i})^2}$$

Solve for $\hat{b}_2$

$$\Delta\hat{b}_2 = \begin{vmatrix} \sum x_{1i}^2 & \sum x_{1i}y_i \\ \sum x_{1i}x_{2i} & \sum x_{2i}y_i \end{vmatrix}$$

$$\Delta\hat{b}_2 = \sum x_{2i}y_i \sum x_{1i}^2 - (\sum x_{1i}y_i)(\sum x_{1i}x_{2i}) \quad \text{.....(1.3.12)}$$

$$\hat{b}_2 = \frac{\Delta\hat{b}_2}{\Delta} = \frac{\sum x_{2i}y_i \sum x_{1i}^2 - (\sum x_{1i}y_i)(\sum x_{1i}x_{2i})}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}$$

Example

The following table shows the amount spent on education in some families (X_1), parents' literacy rates (X_2) and the corresponding students' performance (Y) in the school over a period of time.

Table 1.9: Amount spent on education, parent literacy rate and students' performance in school

Period	1	2	3	4	5	6	7	8	9	10
Y	5	6	8	10	7	8	10	11	15	20
X_1	18	20	25	35	45	60	72	80	85	90
X_2	16	15	12	8	7	6	5	5	4	2

Estimate the regression equation of Y on X_1 and X_2. Interpret the estimated parameters in line with economic theory.

Solution:

Estimated Model: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2$

Table 1.10: Solution to the example

Y	X_1	X_2	Y	x_1	x_2	x_1^2	x_2^2	$y\,x_1$	yx_2	$x_1\,x_2$	y^2
5	18	16	-5	-35	8	1225	64	175	-40	-280	25
6	20	15	-4	-33	7	1089	49	132	-28	-231	16
8	25	12	-2	-28	4	784	16	56	-8	-112	4
10	35	8	0	-18	0	324	0	0	0	0	0
7	45	7	-3	-8	-1	64	1	24	3	8	9
8	60	6	-2	7	-2	49	4	-14	4	-14	4
10	72	5	0	19	-3	361	9	0	0	-57	0
11	80	5	1	27	-3	729	9	27	-3	-81	1
15	85	4	5	32	-4	1024	16	160	-20	-128	25
20	90	2	10	37	-6	1369	36	370	-60	-222	100
100	**530**	**80**	**0**	**0**	**0**	**7018**	**204**	**930**	**-152**	**-1117**	**184**

N = 10 $\bar{Y} = \sum Y/N = 100/10 = 10$

$\bar{X}_1 = \sum X_1/N = 530/10 = 53$

$\bar{X}_2 = \sum X_2/N = 80/10 = 8$

$$\hat{b}_1 = \frac{(\sum y\,x_1)(\sum x_2^2) - (\sum y\,x_2)(\sum x_1\,x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1\,x_2)^2}$$

$$\hat{b}_1 = \frac{930(204) - (-152)(-1117)}{7018(204) - (-1117)^2}$$

$\hat{b}_1 = 0.10835784 \approx 0.1084$

$$\hat{b}_2 = \frac{(\sum y\, x_2)(\sum x_1^2) - (\sum y\, x_1)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_2 = \frac{7018(-152) - (-1117)(930)}{7018\,(204) - (-1117)^2}$$

$$\hat{b}_2 = 0.15178576 \approx 0.1518$$

$$\hat{b}_0 = \bar{Y} - b_1X_1 - b_2X_2$$

$$\hat{b}_0 = 10 - 0.1084\,(53) - 8\,(0.1518)$$

$$\hat{b}_0 = 3.0427$$

$$Y = 3.0427 + 0.1084X_1 + 0.1518X_2$$

Interpretation

b_0 = the intercept of the model or the value of the model at it mathematical origin is 3.0427. The value of Y when X's are zeros.

b_1 = the slope of X_1 is 0.1084. This implies that all things being equal, a 1% change in X_1 would lead to a 0.11% change in Y in the positive direction.

b_2 = -0.1518. This means that the magnitude of change in Y as a result of change in X_2. A 1% change in X_2 would bring about a 0.15 change in Y in the negative direction.

3.4 Practical Applications of Multiple Regression in Econometric Software

Multiple Regression Analysis in STATA

Step 1: Open Stata

Step 2: Click on the Data Editor button to open the Data Editor Window, which looks like a spreadsheet. Start entering the data manually or copy/paste the data from Excel or any other spreadsheet. After you have finished entering the data, double-click on the

variable label (the default name is var1, var2 and so on) and a new window opens up where you can specify the name of the variable and can (optionally) enter a description of it in the Label area. We will assume that for this example we entered data for the following variables given in Step 2 (variable y is the dependent variable and variables x1, and x2 are four explanatory variables).

Step 3: In the Command Window, type the command:

regress y x_2 x_3 (press enter) and you will obtain the regression results. Note that there is no requirement to provide a constant here as Stata includes it automatically in the results (in the output it is labelled as _cons). The β_1 coefficient is the one next to _cons in the Stata regression output and β_2 and β_3 are the coefficients derived in Stata, and you will see them next to the x_2 and x_3 variables in the results.

Example: using the data in Table 1.9, estimate the coefficients, β_1, β_2 and β_3.

SELF ASSESSMENT EXERCISE

i. Use the output of Table 1.10 and the estimated regression equation to estimate the standard errors of the parameter estimates and test for their respective statistical significance at 5 percent level of significance.

ii. Use the same information in (i) to estimate the coefficient of multiple determination and interpret your result.

4.0 CONCLUSION

This unit discussed extensively a three-variable linear regression model which is in many ways an extension of the two-variable regression model. The unit concludes that there are some new concepts involved, such as partial regression coefficients, multiple correlation coefficient or coefficient of multiple determination, adjusted and unadjusted (for degrees of freedom) R^2. Although, R^2 and adjusted R^2 are overall measures of how the chosen model fits a given set of data, their importance should not be overplayed.

What are critical are the underlying theoretical expectations about the model in terms of a priori signs of the coefficients of the variables entering the model.

5.0 SUMMARY

This unit introduced the simplest possible multiple linear regression model, namely, the three-variable regression model. It is understood that the term linear refers to the linearity in the parameters. The unit discussed models with two explanatory variables and derived the normal equation of the three-variable regression model. The unit further reiterated the assumptions of the linear regression model previously discussed with emphasis assumptions regarding the explanatory variables themselves. Some new concepts have been highlighted in the three-variable model which is an extension of the two-variable model such as partial regression coefficients, multiple correlation coefficient otherwise known as coefficient of multiple determination, adjusted and unadjusted (for degrees of freedom) R^2.

6.0 Tutor-Marked Assignment

The following Table 1.11 shows the values of expenditure on clothing (Y), total expenditure (X_1) and the price of clothing (X_2).

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Y	3.5	4.3	5	6	7	9	8	10	12	14
X_1	15	20	30	42	50	54	65	72	85	90
X_2	16	13	10	7	7	5	4	3	3.5	2

(1) (a) Find the least squares regression equation of Y on X_1 and X_2.

(b) Compute the coefficient of multiple determination and the standard errors of the estimated parameters and conduct tests of significance.

(2) The following results were obtained from a sample of 12 companies on their output (Y), labour input (X_1) and capital input (X_2), measured in arbitrary units.

$\sum Y = 753$ $\sum Y^2 = 48139$ $\sum YX^1 = 40830$

$\sum X_1 = 643$ $\sum X_1^2 = 34843$ $\sum YX_2 = 6796$

$\sum X_2 = 106$ $\sum X_2^2 = 976$ $\sum X_1X_2 = 5779$

(a) Find the least squares equation of Y on X_1 and X_2. What is the economic meaning of your coefficients?

(b) Given the following sample values of output (Y) in table 1.12, compute the standard errors of the estimates and test their statistical significance.

Companies	A	B	C	D	E	F	G	H	I	J	K	L
Output	64	71	53	67	55	58	77	57	56	51	76	68

(c) Find the coefficient of multiple determination and the unexplained variation in output

7.0 REFERENCES/FURTHER READINGS

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

UNIT 4: STATISTICAL TESTS OF SIGNIFICANCE

1.0 Introduction

2.0 Objective

3.0 Main Content

3.1 Statistical Test of the OLS Estimates of a Simple Regression

3.2 Statistical Test of the OLS Estimates of a Simple Regression

4.0 Conclusion

5.0 Summary

6.0 Tutored-Marked Assignment

7.0 References

1.0 INTRODUCTION

Under the assumptions of the CLRM, we know that the estimators $\hat{\alpha}$ and $\hat{\beta}$ obtained by OLS follow a normal distribution with means a and β and variances $\sigma_{\hat{\alpha}}^2$ and $\sigma_{\hat{\beta}}^2$, respectively. This test involves the means and variances of the estimates of the parameters.

2.0 OBJECTIVES

At the end of this unit students should be able to:

- Test the statistical significance of the OLS estimates of simple regression
- Test the statistical significance of the OLS estimates of multiple regression
- Construct confidence interval of parameter estimates

3.0 MAIN CONTENT

3.1 Statistical Tests of Significance of the OLS Estimates of a Simple Regression

3.2 Statistical Tests of Significance of the OLS Estimates of a Simple Regression

3.1 Statistical Tests of Significance of the OLS Estimates of a Simple Regression

After the estimation of the model in the previous sub-unit, we need to test the explanatory power or the goodness of fit of the model, as well as the statistical reliability/significance at a given level in respect of b_i (i = 0,1,2,…n).

This is achieved using the following tests:

i. Coefficient of determination, r^2.

ii. Standard error test

iii. Z and t-statistic (test)

3.1.1. R^2 and the Simple Regression Line

The coefficient of determination, r^2, is used in determining the goodness of fit of the regression line obtained using the OLS method. That is, it is used in testing the explanatory power of the linear regression of Y on X.

Thus, in order to determine the degree to which the explanatory variables is able to explain the variation in the dependent variable, Y, the r^2provides a useful guide.

If we measure the dispersion of observations around the regression line, some may be closer to the line while others may be far away from it. Our argument is that the closer these observed values are to the line, the better the goodness of fit. On the basis of this, we may turn out to state that changes in the dependent variable, Y, is explained by changes in the explanatory variable, X. to know precisely the extent to which the total variation in Y is explained by the independent variable, X, we compute the value of r^2 as the ratio of explained variation to total variation. That is,

r^2 = Explained variation/ Total variation = $\frac{\sum \hat{y}^2}{\sum y^2}$ …….(1.4.1)

However, the coefficient of determination, r^2 is also expressed as:

$$r^2 = \frac{\sum (xy)^2}{\sum x^2 \sum y^2} \text{ or } r^2 = \frac{\hat{b}_1 \sum xy}{\sum y^2} \text{ ……..….(1.4.2)}$$

The first expression of equation (1.4.2) shows that r^2 determines the proportion of the variation in Y which is explained by variation in X.

If for instance, $r^2 = 0.75$, it means that 75% of the variation in Y is due to the variation in X, while 25% of the variation is explained by the disturbance term, U. Thus, this regression line gives a good fit to the observed data.

If, however, $r^2 = 0.45$, it means that only 45% of the variation in Y is as a result of variation in X while 55% of the variation is due to the disturbance term. This is a poor indication and the regression line does not give a good fit to the observed data. If r^2 is 0.5 and above, it shows a good fit while a value of r^2 less than 0.5 shows poor fit.

Note that the r^2 is much of relevance when the estimated model is used for forecasting. Note also that the value of $r^2 = 0$ signifies that the independent variable cannot explain any changes in the dependent variable, hence variation in the independent variable has no effect on the dependent variable.

Using the data on Table 1.6 with the r^2 formulae, where $\sum y^2 = 646$

$$r^2 = \frac{\sum (xy)^2}{\sum x^2 \sum y^2} \quad \text{or} \quad r^2 = \frac{\hat{b}_1 \sum xy}{\sum y^2}$$

$$r^2 = \frac{(81)^2}{18(646)} \qquad \text{or} \qquad r^2 = \frac{4.5(81)}{646}$$

$$r^2 = 0.56 \qquad r^2 = 0.56$$

This implies that 56% of the variation in Y is due to the variation in X, while 44% of the variation is explained by the disturbance term, U. Thus, this regression line gives a good fit to the observed data.

3.1.2 The Coefficient of Multiple Determination ($R^2_{Y.X_1X_2}$)

When the explanatory variables are more than one then we have a situation called coefficient of multiple determination. This can be obtained by taking the square of correlation coefficient. That is why it is also called squared multiple correlation coefficient. The coefficient of multiple determination is designated R^2, with subscripts the variables whose relationship is under study. For example, in the three-variable model, the squared multiple correlation coefficient is $R^2_{Y.X_1X_2}$. As in the two-variable model, R^2 shows the proportion of the total variation of Y explained by the regression plane, that is, by changes in X_1 and X_2.

$$R^2_{Y.X_1X_2} = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = 1 - \frac{\sum e^2}{\sum y^2} = \frac{\sum y^2 - \sum e^2}{\sum y^2} \quad \text{.....(1.4.3)}$$

Given the relationship between b_i and R^2, equation (1.48) can be written as:

$$R^2_{Y.X_1X_2} = \frac{\hat{b}_1 \sum x_1 y + \hat{b}_2 \sum x_2 y}{\sum y^2} \quad \text{.....(1.4.4)}$$

The value of R^2 lies between 0 and 1. The higher the value of R^2, the greater the proportion of variation of Y explained by the plane, that is the better the 'goodness of fit' of the regression plane to the sample observations. The closer the coefficient of multiple determination (R^2) to zero the worse the fit of the regression plane.

The above formula for R^2 does not take into consideration the loss of degrees of freedom from the introduction of additional explanatory variables in the function. An adjusted expression of R^2 is discussed in the next section.

3.1.3. Statistical Significance of b_i using the Standard Error Test

To use this test, it is important to know the mean and variance of the parameter estimates $\hat{b}_0$ and $\hat{b}_1$. It has been established that the mean and variance of $\hat{b}_0$ and $\hat{b}_1$ are respectively:

$$E(\hat{b}_0) = b_0 \ \therefore \ \text{var}\,\hat{b}_0 = [\hat{b}_0 - b_0] = \frac{\hat{\sigma}^2 u \sum x_i^2}{n\sum x_i^2} \quad \text{.....(1.4.5)}$$

$$E(\hat{b}_1) = \hat{b}_1 \ \therefore \ \text{var}\ \hat{b}_1 = [\hat{b}_1 - b_1] = \frac{\hat{\sigma}^2 u}{\sum x_i^2} \quad \text{.....(1.4.6)}$$

The standard error test enables us to determine the degree of confidence in the validity of the estimate. That is, from the test, we are able to know whether the estimates $\hat{b}_0$ and $\hat{b}_1$ are significantly different from zero. The test is mainly useful when the purpose of the research is the explanatory (analysis) of economic phenomena and estimation of reliable values.

We formally start by stating null hypothesis: (H_0): $\hat{b}_0 = 0$

Against the alternative hypothesis: H_1: $\hat{b}_0 \neq 0$

The standard error for the parameter estimates $\hat{b}_0$ and $\hat{b}_1$ are respectively computed as shown:

i. $$S(\hat{b}_0) = \sqrt{\text{var}(\hat{b}_0)} = \sqrt{\frac{\hat{\sigma}_U^2 \sum x_i^2}{n\sum x_i^2}} \quad \text{.....(1.4.7)}$$

But $$\hat{\sigma}_U^2 = \frac{\sum e_i^2}{n-2}$$

$$\therefore \ S(\hat{b}_0) = \sqrt{\frac{\sum e_i^2 \sum x_i^2}{(n-2)n\sum x_i^2}} \quad \text{.....(1.4.8)}$$

ii. $S(\hat{b}_1) = \sqrt{\text{var}(\hat{b}_1)} = \sqrt{\frac{\hat{\sigma}_U^{\,2}}{\sum x_i^{\,2}}} = \sqrt{\frac{\sum e_i^{\,2}}{(n-2)\sum x_i^{\,2}}}$ ……...(1.4.9)

When the numerical values for the $S(\hat{b}_0)$ and $S(\hat{b}_1)$ are each compared with the numerical values of $\hat{b}_0$ and $\hat{b}_1$, the following decision rules apply:

Decision Rules

i. If $(b_i) < ½(b_i)$, we reject the null hypothesis and conclude that $\hat{b}_1$ is statistically significant.

ii. If on the other hand, $S(b_i) > ½(b_i)$, we accept the null hypothesis that the true population parameter $b_i = 0$. This concludes that the estimate is not statistically significant. Therefore, change in X has no effect on the values of Y.

The acceptance of the null hypothesis has economic implication. Thus, the acceptance of the null hypothesis, say $\hat{b}_1 = 0$ implies that the explanatory variable to which this estimate relates does not influence the dependent variable, Y, and should not be included in the function. This situation renders the relationship between Y and X, hence the regression equation is parallel to axis of the explanatory variable, X. In this case, $\hat{Y}_i = \hat{b}_0 + 0X_i$ or $\hat{Y}_i = \hat{b}_0$. The zero slope shows that no relationship exists between Y and X.

Similarly if the null hypothesis of $\hat{b}_0 = 0$ is accepted, on the basis that $S(b_0) > 1/2\hat{b}_0$, it implies that the intercept of this regression line is zero. Therefore, the line passes through the origin. In this case, the relationship between Y and X will be $\hat{Y}_i = 0 + b_1 X_i$ or $\hat{Y}_i = b_1 X_i$. The two situations are diagrammatically presented as show in Figure 1.5

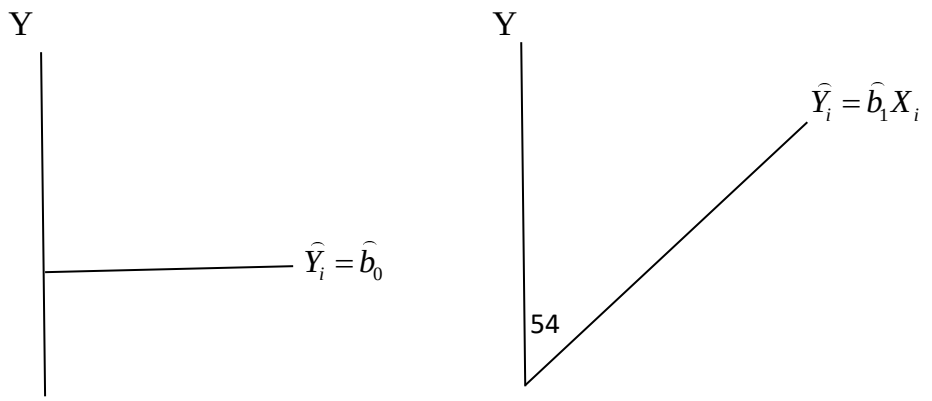


Figure 1.5: (a) $\hat{b}_1 = 0$ (b) $\hat{b}_0 = 0$

Example

Refer to the example on the regression analysis, the r^2 and the standard errors, $S(\hat{b}_0)$ and $S(\hat{b}_1)$ are as computed:

From the computed values of the table,

$\sum x_i y_i = 81$; $\sum x_i^2 = 18$; $\sum y_i^2 = 646$; $\sum e_i^2 = 281.5$; n = 8; K = 2 where K = degree of freedom

$$r^2 = \frac{\sum (xy)^2}{\sum x^2 \sum y^2} = \quad r^2 = \frac{(81)^2}{18(646)}$$

$$r^2 = 0.56$$

Hence X is a fairly good predictor of Y. About 56% of the variation in the dependent variable, Y is explained by variation in the explanatory variable, X, while 44% of the variation is due to the disturbance term, U. Thus, the regression line fairly gives a good fit to the observed data.

On the statistical significance of the parameter estimates, we compute:

$$S(\hat{b}_0) = \sqrt{\frac{\sum e_i^2 \sum x_i^2}{(n-2)n\sum x_i^2}} = \sqrt{\frac{281.5(530)}{(8-2)(8)(18)}} = 13.14$$

$$S(b_1) = \sqrt{\frac{\sum e_i^2}{(n-2)\sum x_i^2}} = \sqrt{\frac{281.5}{(8-2)}} = 1.61$$

Thus, the result of our regression may be presented formally as:

$$\hat{Y}_i = 19 + 4.5X_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.4.10)$$

(13.14) (1.61)

R^2 =0.56, n = 8

From the above estimated result,

$\hat{b}_0 = 19$, $\hat{b}_1 = 4.5$, $S(\hat{b}_0) = 13.14$, $S(\hat{b}_1) = 1.61$

Therefore, $1/2(\hat{b}_0) = 9.5$ and $1/2(\hat{b}_1) = 2.25$

Thus, $S(\hat{b}_0) > 1/2(\hat{b}_0)[13.14 > 9.5]$ and we accept the null hypothesis that $H_0 : \hat{b}_0 = 0$. This shows that $\hat{b}_0$ is not statistically significant.

Similarly, $S(\hat{b}_1) < 1/2(\hat{b}_1)[1.61 < 2.25)$ and we reject the null hypothesis. Thus, the estimate $\hat{b}_1$ is statistically significant.

Note: the acceptance and meaningful interpretation of econometric result entails a combination of high r^2 and low standard errors.

3.1.4 Z and t-statistic Test

i. Z-test of Statistical Significance of OLS Estimate

The z-test is employed when the sample size is sufficiently large (i.e. n > 30). It could be applied whether the population variance is known or not. The z-test is applied using the formula:

$$z* = \frac{\hat{b}_1}{S(\hat{b}_1)} \quad \text{.....(1.4.11)}$$

Where z* is the calculated z which is to be compared with the z table (theoretical value of z) at a given level of significance, say 5%.

Decision Rule

If -z < z* < +z at 0.025, we accept the null hypothesis that $H_0 : b_1 = 0$, and conclude that our estimate is not statistically significant.

If however, z* > z, then we accept that $H_1 : b_1 \neq 0$, and conclude that our estimate is statistically significant (Koutsoyannis, 1977).

Example:

As an illustration, consider an estimated function from a sample of 50 observations in the form: $\hat{Y}_i = 15. + 6.8X_i$

(2.8) (1.05)

To conduct Z test for the estimates $\hat{b}_0$ and $\hat{b}_1$ at 5% level of significance, we proceed as follows:

H_0: $\hat{b}_i = 0$ (Null hypothesis)

H_1: $\hat{b}_i \neq 0$ (Alternative hypothesis

$Z^* = \frac{\hat{b}_0}{S(\hat{b}_0)} = 15/2.8 = 5.36$ ($\hat{b}_0$)

$Z^* = \frac{\hat{b}_1}{S(\hat{b}_1)} = 6.8/1.05 = 6.48$ (for $\hat{b}_1$)

Z table at 5% level of significance = 1.96

For $\hat{b}_0$ Z* (5.36) > Z-table (1.96); and for $\hat{b}_1$, Z* (6.48) > Z-table (1.96)

Therefore, we reject the null hypothesis, and conclude that estimates $\hat{b}_0$ and $\hat{b}_1$ are statistically significant at 0.05 level.

ii. t-Test of Significance of $\hat{b}_i$

The student's t-test is used when the sample size is small (i.e. n < 30) provided that the population parameter follows a normal distribution. With this n view, and taking degree of freedom into consideration, we need to compare this with the theoretical t, at a given level of significance say 5%.

Our null and alternative hypotheses are respectively formulated thus:

H_0: $\hat{b}_i = 0$

H_1: $\hat{b}_i \neq 0$

Following a normal distribution, our *t* is computed as follows:

$$t^* = \frac{\hat{b}_i}{S(\hat{b}_i)} \quad \text{.....(1.4.12)}$$

As stated previously, the empirical t (t*) value is compared with the t-table (t^c) with n-k degree of freedom, given a 5% level of significance.

Decision Rule:

If $-t_{0.025} < t^* < t_{0.025}$ (with n-k degree of freedom), we accept the null hypothesis, and conclude that our estimate $\hat{b}_i$ is not statistically significant at 0.05 level of significance.

If however, $t^* > t_{0.025}$, we reject the null hypothesis, and accept the alternative hypothesis. This concludes that the estimate $\hat{b}_i$ is statistically significant.

Example:

Given a sample size of n = 18, the model was estimated to be:

$$\hat{Y}_i = 21 + 0.75X_i$$

(10.2) (1.4)

We wish to test the statistical reliability of $\hat{b}_0$ and $\hat{b}_1$ respectively.

From the estimated model,

$\hat{b}_0 = 21,\ S(\hat{b}_0) = 10.2;\ \hat{b}_1 = 0.75,\ S(\hat{b}_1) = 1.4$

Therefore, $t^* = \frac{(\hat{b}_0)}{S(\hat{b}_0)} = 21/10.2 = 2.06$ (for $\hat{b}_0$)

$t^* = \frac{(\hat{b}_1)}{S(\hat{b}_1)} = 0.75/1.4 = 0.54$ (for $\hat{b}_1$)

The critical value of t for (n-k) or 8 – 2 = 16 degree of freedom [$t_{0.025(16)}$] are:

$t_{0.025(16)} = -2.12$ and $+2.12$.

For the estimate $\hat{b}_0$, we see that $t^* < t_{0.025(n-k)}$ [2.06 < 2.12), we accept the null hypothesis and conclude that the estimate $\hat{b}_0$ is not statistical significant.

In the case of estimate $\hat{b}_1$, since $t^* < t^* < t_{0.025(n-k)}$ [0.54 < 2.12), we also accept the null hypothesis and conclude that $\hat{b}_1$ is not statistically significant at 5% significant level.

3.1.5 Confidence Interval for the Parameter Estimates

a. Confidence for Z-statistic

The construction of confidence intervals for the estimates $\hat{b}_0$ and $\hat{b}_1$ enables us to state how close to these estimates the true parameter lies. It shows the limiting values within which the true parameter is expected to lie within a certain degree of confidence.

In econometrics, we usually chose 95%. This implies that the confidence limit computed from a given sample would include from a population parameter in 95% of the cases. That s, we are 95% confident that our parameter estimates represent the true population parameter.

Using the Z-distribution, the 95% confidence interval for the parameters $\hat{b}_i$ is constructed as follows:

$$\hat{b}_i - 1.96(S\hat{b}_i) < b_i < \hat{b}_i + 1.96(S\hat{b}_i) \quad \text{.....(1.4.13)}$$

This means that the unknown population parameter $\hat{b}_1$ will lie within the limits 95 times out of 100.

Example:

From the previous regression model, equation (1.2.12):

$$\hat{Y}_i = 19 + 4.5X_i$$

(13.14) (1.61)

And choosing 95% for the confidence coefficient, our confidence interval for b_0 is:

$$\hat{b}_0 - 1.96(S\hat{b}_0) < b_0 < \hat{b}_0 + 1.96(S\hat{b}_0)$$

$$19 - 1.96\ (13.14) < b_0 < 19 + 1.96\ (13.14)$$

$$-6.75 < b_0 < 44.75$$

This implies that the true population parameter b_0 will lie between -6.75 and 44.75 with a probability of 0.95.

Similarly, the confidence interval for b_1 is constructed as shown:

$$\hat{b}_1 - 1.96(S\hat{b}_1) < b_1 < \hat{b}_1 + 1.96(S\hat{b}_1)$$

$$4.5 - 1.96\ (1.61) < b_1 < 4.5 + 1.96\ (1.61)$$

$$1.34 < b_0 < 7.66$$

We also conclude that the true population will lie between 1.34 and 7.66 with a probability of 0.95.

b. Confidence interval for the t-statistic

The confidence interval for t-statistic is also constructed the same way as for the Z-distribution. The difference is that the t-distribution uses n-k degree of freedom. Thus, the 95% confidence interval for the parameter estimates b_i is constructed as follows:

$$\hat{b}_i - t_{0.025}(S\hat{b}_i) < b_i < \hat{b}_i + t_{0.025}(S\hat{b}_i) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.4.14)$$

with n-k degree of freedom.

Example:

From a sample of 20, if the estimated model was:

$$\hat{Y}_i = 13.25 + 1.8X_i$$

(6.04) (0.35)

The 95% confidence interval for the parameter estimates is constructed as shown:

d.f = 20 – 2=18

For $\hat{b}_0$: $\hat{b}_0$ -2.101 $(S\hat{b}_0) <$ b_0 < $\hat{b}_0$ + 2.101 $(S\hat{b}_0)$

13.25 – 2.101(6.04) < b_0 < 13.25 + 2.101(6.04)

0.560 < b_0 < 25.94

This shows that the true population parameter b_0 (or the intercept) will lie between 0.56 and 25.94 with a probability of 0.95.

For $\hat{b}_1$: $\hat{b}_1$ - 2.101 $(S\hat{b}_1) <$ b_1 < $\hat{b}_1$ + 2.101 $(S\hat{b}_1)$

1.8 – 2.101(0.35) < b_1 < 2.101(0.35)

1.064 < b_1 < 2.535

This implies that the value of true parameter b_1 will be between 1.064 and 2.535, given a probability of 0.95.

SELF ASSESSMENT EXERCISE

Explain confidence interval in relation to the how close to the estimate the true parameter lies.

3.2 Statistical Test of Significance of the OLS of a Multiple Regression

3.2.1 The Adjusted Coefficient of Multiple Determination ($R^2_{Y.X_1X_2}$)

Worthy of note is that the inclusion of additional explanatory variables does not in any way reduce the R^2 and rather raise it. When new explanatory variables are included in a model, the numerator of R^2 equation increases, while the denominator remains the same because total variation ($\sum y^2$) is given in any particular sample.

In order to correct for such defect, the R^2 is adjusted by taking into consideration the degrees of freedom, which clearly decrease as new explanatory variables are introduced in the function. The expression for the adjusted R^2 is:

$$\bar{R}^2 = 1-(1-R2)\frac{n-1}{n-K}$$

$$\bar{R}^2 = 1-\left[\frac{\sum e^2/(n-K)}{\sum y^2/(n-1)}\right] \quad \text{.....(1.4.15)}$$

R^2 as usual is the unadjusted coefficient of multiple determination, n is the number of observations and K is the number of parameter estimated from the sample.

3.2.2 The Mean and Variance of the Parameter Estimates ($\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$)

The mean of the estimates of the parameters in the three-variable model is derived in the same way as the two-variable model. The estimates $\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$ are assume to be unbiased estimates of the true parameters of the relationship between Y, X_1 and X_2: their mean expected value is the true parameter itself.

$$E(\hat{b}_0) = b_0, \qquad E(\hat{b}_1) = b_1, \qquad E(\hat{b}_2) = b_2 \quad \text{.....(1.4.16)}$$

The variances of the parameter estimates are obtained by the following formulae:

$$\text{var}(\hat{b}_0) = \sigma_u^{\,2}\left[\frac{1}{n} + \frac{\bar{X}_1^{\,2}\sum x_2^{\,2} + \bar{X}_2^{\,2}\sum x_1^{\,2} - 2\bar{X}_1\bar{X}_2\sum x_1x_2}{\sum x_1^{\,2}\sum x_2^{\,2} - (\sum x_1x_2)^2}\right] \quad \text{.....(1.4.17)}$$

$$\text{var}(\hat{b}_1) = \sigma_u^{\,2}\frac{\sum x_2^{\,2}}{\sum x_1^{\,2}\sum x_2^{\,2} - (\sum x_1x_2)^2} \quad \text{.....(1.4.18)}$$

$$\text{var}(\hat{b}_2) = \sigma_u^{\,2}\frac{\sum x_1^{\,2}}{\sum x_1^{\,2}\sum x_2^{\,2} - (\sum x_1x_2)^2} \quad \text{.....(1.4.19)}$$

Where $\sigma_u^{\,2} = \sum e^2/(n\text{-}K)$, K being the total number of parameters which are estimated. In the there-variable model, K = 3.

3.2.3 Test the Reliability/significance of the parameter estimates

The conventional test of reliability of parameter estimates has been explained in the previous unit. This test includes the standard error test which is equivalent to the Student's t-test. Here, a summary of the procedures is provided to guide the students on how to conduct hypothesis testing.

Conventionally in econometrics usage, researchers test the null hypothesis H_0: $b_i = 0$ for each parameter, against the alternative hypothesis H_1: $b_i \neq 0$. The hypothesis under discussion is the two-tailed test at a chosen level of significance, usually at the 5 percent significant level.

1. The standard error test

The value of the standard error $(\sqrt{\text{var}(\hat{b}_i)})$ is compared with the numerical values of the estimates and a decision is taken o the basis of the comparison.

(a) The null hypothesis is accepted if $s_{(\hat{b}_i)} > 1/2(b_i)$; that is, the estimate b_i is not statistically significance at the 5 percent level of significance for a two-tailed test.

(b) The null hypothesis is rejected while alternative hypothesis is accepted if $s_{(\hat{b}_i)} < 1/2(b_i)$; that is, the estimate b_i is statistically significance at the 5 percent level of significance for a two-tailed test.

In conventional terms, the smaller the standard errors, the stronger is the evidence that the estimates are statistically reliable.

2. The Student's test of significance

The student's t ratio for each b_i is computed as follows:

$$t^* = \frac{\hat{b}_i}{s_{(b_i)}} \quad \text{.....(1.4.20)}$$

Equation (1.3.20) is the sample or observed t ratio which is compared with the theoretical value of t obtainable from the student's t-table with n-K degrees or freedom (Koutsoyiannis, 1977).

(a) If t* falls in the acceptance region; that is, if $-t_{0.025} < t^* < t_{0.025}$(with n-K degrees of freedom), we accept the null hypothesis that $\hat{b}$ is not statistically significant and hence the corresponding explanatory variable does not explain the variation in the dependent variable.

(b) If on the other hand, t* falls in the critical region, we reject the null hypothesis that $\hat{b}$ is statistically significant and hence the corresponding explanatory variable contributes to the explanation of the variation in the dependent variable.

It is important to note that the greater the value of t*, the stronger is the evidence that $\hat{b}$ is significant and vice versa.

SELF ASSESSMENT EXERCISE

i. Use the output of Table 1.10 and the estimated regression equation to estimate the standard errors of the parameter estimates and test for their respective statistical significance at 5 percent level of significance.

ii. Use the same information in (i) to estimate the coefficient of multiple determination and interpret your result.

4.0 CONCLUSION

This unit discussed extensively the statistical significance of parameter estimates of both simple and multiple regression. The unit concludes that although, R^2 and adjusted R^2 are overall measures of how the chosen model fits a given set of data, their importance should not be overplayed. What are critical are the underlying theoretical expectations about the model in terms of a priori signs of the coefficients of the variables entering the model.

5.0 SUMMARY

This unit introduced the statistical significance of the parameter estimates to ascertain their reliability. The unit discussed the R^2, statistical significance of α and β in terms of simple regression estimates and β_0, β_1 and β_2 in terms of multiple regrssion.

6.0 Tutor-Marked Assignment

The following Table 1.11 shows the values of expenditure on clothing (Y), total expenditure (X_1) and the price of clothing (X_2).

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Y	3.5	4.3	5	6	7	9	8	10	12	14
X_1	15	20	30	42	50	54	65	72	85	90
X_2	16	13	10	7	7	5	4	3	3.5	2

(1) (a) Find the least squares regression equation of Y on X_1 and X_2.

(b) Compute the coefficient of multiple determination and the standard errors of the estimated parameters and conduct tests of significance.

(c) Construct 95 percent confidence intervals for the population parameters

(2) The following results were obtained from a sample of 12 companies on their output (Y), labour input (X_1) and capital input (X_2), measured in arbitrary units.

$\sum Y = 753$ $\sum Y^2 = 48139$ $\sum YX^1 = 40830$

$\sum X_1 = 643$ $\sum X_1^2 = 34843$ $\sum YX_2 = 6796$

$\sum X_2 = 106$ $\sum X_2^2 = 976$ $\sum X_1X_2 = 5779$

(a) Find the least squares equation of Y on X_1 and X_2. What is the economic meaning of your coefficients?

(b) Given the following sample values of output (Y) in table 1.12, compute the standard errors of the estimates and test their statistical significance.

Companies	A	B	C	D	E	F	G	H	I	J	K	L
Output	64	71	53	67	55	58	77	57	56	51	76	68

(c) Find the coefficient of multiple determination and the unexplained variation in output

(d) Construct 99 percent confidence intervals for the population parameters.

7.0 REFERENCES/FURTHER READINGS

Mirer, T.W. (1995). Economic Statistics and Econometrics (Third Edition), Prentice-Hall Inc, London.

MODULE 2: ECONOMETRIC PROBLEMS, BASIC IDEAS OF THE IDENTIFICATION PROBLEM AND SIMULTENEOUS EQUATION ESTIMATION METHODS

Unit 1: Econometric Problems (Heteroscedasticity, Autocorrelation and Multicollinearity)

Unit 2: Basic Ideas of the Identification Problem, Dummy variables and Distributed lag Models

Unit 3: Simultaneous Equation Estimation Methods (2SLS, 3SLS, etc)

Unit 4: Matrix treatment of Multiple Regression and Advanced treatment of Simultaneous Equation Estimation Techniques

UNIT 1: ECONOMETRIC PROBLEMS

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 Heteroscedasticity: Causes, Detection, Consequences and Correction

3.2 Autocorrelation: Detection, Consequences and Correction
3.3 Multicollinearity: Detection, Consequences and Correction
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

- **INTRODUCTION**

In the previous units, we stated that the third stage in any econometric research is the evaluation of the reliability of the estimates of the parameters. After the estimation of the parameters with the OLS method or any other econometric technique, there is need to establish how trustworthy these estimates are. The evaluation is on the basis of three criteria. Firstly, is the a priori economic criterion, which is determined by the postulates of economic theory and relate to the sign and magnitude of the parameters. Secondly, statistical criteria otherwise known as the first-order tests, defined by statistical theory. Thirdly, econometric criteria, also known as second-order tests, defined by econometric theory.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- Examine econometric problems of heteroscedasticity, autocorrelation and multicollinearity: their causes, detection, consequences and correction.
- Identify the basic ideas of identification, dummy variables and distributed lag models.
- Explain simultaneous equation estimation methods (2SLS, 3SLS etc.).
- Discuss matrix treatment of multiple regression and advance treatment of simultaneous equation estimation techniques

3.0 MAIN CONTENT

3.1 Heteroscedasticity

3.1.1 Meaning of Heteroscedasticity

Heteroscedasticity is an econometric problem which arises as a result of the violation of the assumption of homoscedasticity. Recall that the variance of each disturbance term in a given model is the same (constant) [homoscedastic] for all values of the explanatory variables. That is, var(u) = $E[(u_i - E(u)\}^2 = \sigma_u^2$, which is constant.

If however, this assumption s violated, then:

$$var(u_i) = \sigma_u^2 \text{ (not constant)}$$

where the subscript I signifies that the individual variances of u may all be different (Koutsoyiannis, 1977).

According to Mirer (1995), Heterohcedasticity is the situation in which the standard deviations of the disturbances are not the same for all observations. This often arises in the analysis of cross-section data, although t may be present n time-series data also.

Heteroscedasticity manifests itself with the variance of the u's tending to change with changes in the values of the regressors.

3.1.2 Causes of Heteroscedasticity

Heteroscedasticity may be attributable to the following factors:

i. Error of specification due to the non-inclusion of all relevant regressors in the equation of the model.

ii. Accumulated error of measurement which tends to increase over time. This makes the variance of u_i to increase with increase in the value of X.

3.1.3 Consequences of heteroscedasticity in a model

i. the presence of heteroscedascity disturbance terms render the formula for variances of parameter estimates for conducting test of significance rather difficult. More also, confidence intervals cannot be constructed with ease.

ii. Since it is no longer constant, it is not possible to factor out the variance to ease further computation. Sequel to the above, the test of significance using the t or z will yield inaccurate result.

iii. The presence of heteroscedacity renders the OLS estimates (parameter estimates) inefficient. Thus, the property of minimum variance in the class of unbiased estimates no longer holds for these parameters.

iii. The explanatory power of the regressor(s) will be affected. Predictions based on the estimates of the paraneters will no longer be accurate.

3.1.4 Test for Detection of the Presence of Hetroscedasticity

To test for the presence of heteroscedasticity or otherwise, various devices are used. Such tests include:

i. Spearman's Rank correlation tes

ii. The Goldfeld and Quandt test

iii. The Glejser test

1. The Spearman's Correlation Test

This test can be applied to either small or large samples. N applying the test, the following procedures hold:

(a) Y is regressed on X, and the value of the individual e's obtained.

(b) These values (e's) are ordered (from smallest to highest) on the basis of its absolute values (i.e. ignoring the signs) together with the X values either in ascending or descending order of magnitude (in terms of their positions in the data) and we compute the rank correlation coefficient using:

$$r' = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \quad \text{.....(2.1.1)}$$

n = number of observations; d = rank difference

A high value of r' suggests the presence of heteroscedasticity while a low value suggest its absence.

For the case of more than one explanatory variable, we perform the rank correlation between e's and each of the explanatory variable separately.

2. The Goldfeld and Quandt Test

This test is mainly used for large sample sizes. It is based on the assumption that:

i. u_i is a normal distribution with zero mean and standard deviation, σ_u^2 i.e.

$$u_i \sim n(0, \sigma_u^2)$$

ii. No serial dependence of u_i or correlation between the u's and X's, i.e.

$$E(u_i\, u_j) = 0;\ E(u_i\, X_i) = 0$$

To conduct this test, we proceed as follows:

(a) Order the observations according to the magnitude of the explanatory variable X (in ascending order).

(b) Select arbitrarily a certain number, C (central observations) which we need to omit from the analysis. Note; it has been observed that for n = 30, the number to be omitted = 8; for n = 60, the number = 16. If $n \geq 30$, the number of central observations to be omitted is approximately equal to ¼ of n. The remaining (n - c) observations are divided into two sub-samples of equl size (n-c/2), one including the small values of X and the other including large values of X.

(c) Regress each sub-sample and obtain the sum of squared residuals from each, such that:

Σe_1^2 = Residuals from the sub-sample of low values of X's with $[(n - c)/2] - K$ degree of freedom, where K = number of parameters in the model.

Σe_2^2 = Residuals from the sub-sample of high values of X's with [(n – c)/2] – K degree of freedom.

Using F ratio, the value is:

$F^* = \Sigma e_2^2 / \Sigma e_1^2$……………………………………………………(2.1.2)

Decision Rule:

If F^* > F-table, we accept that there is heteroscedasticity (i.e. reject null hypothesis, H_0: u_i's are homoscedastic).

If however, F^* < F-table, then we conclude that u_i's are homoscedastic.

Thus, the larger the F ratio (F^*), the stronger the heteroscedasticity of the u_i's.

3. The Glejser Test

This test involves the performance of the following regressions:

(i) Y on X's and computes the residuals, e's.

(ii) We regress the absolute values of e's (│e's│) on the explanatory variable with which σui2 is thought on a priori grounds, to be associated.

The actual form of this regression is usually not known, so that one may experiment with various formulations, containing various powers of X. hence, the first tests are preferred to this.

Example:

Considering personal savings and personal income of a certain community over a 31 year period shown in Table 2.1:

Table 2.1

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Saving (NM)	10	12	8	14	15	7	11	16	18	20	13	17	6	5	19	24
Income(NM)	54	60	38	63	65	32	56	68	72	80	64	70	30	25	75	84
Period	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
Saving(NM)	25	26	27	30	32	35	28	29	21	22	23	33	34	38	40	
Income(NM)	87	88	92	95	100	104	96	97	78	85	89	93	94	110	120	

Question: Use Goldfeld and Quandt to test for the presence or otherwise of heteroscedasticity.

Solution:

Income = X, Saving = S

Using Goldfeld and Quandt test, we order the observations n ascending order of the X's and omitting the central observations, we are left with two sub-samples of data, one with the lower values of X and one with the higher values of X as shown below in Table 2.2:

Table 2.2

X	25	30	32	38	54	556	60	63	65	68	70	72	75	78	80
S	5	6	7	8	10	11	12	14	13	16	17	18	19	21	20
X	84	85	87	88	89	92	93	94	95	96	98	100	104	110	120
S	24	22	25	26	23	27	33	34	30	28	29	32	35	38	40

Dividing the observations into two sub-sets after omitting the nine central observations, we have as presented in Table 2.3 below:

Table 2.3

n_1	S_1	X_L	n_2	S_2	X_H
1	5	25	12	23	89
2	6	30	13	27	92
3	7	32	14	33	93
4	8	38	15	34	94
5	10	54	16	30	95
6	11	56	17	28	96
7	12	60	18	29	98
8	14	63	19	32	100
9	13	64	20	35	104
10	15	65	21	38	110
11	16	68	22	49	120

We regress the following:

$S_1 = f(X_L) \rightarrow S = a_0 + a_1X_1$……(i)

$S_2 = f(X_H) \rightarrow S = b_0 + b_1X_2$……...(ii)

Where $X_L = X_1$ and $X_H = X_2$

From equation (i),

$$\hat{a}_1 = \frac{\sum sx_1}{\sum x_1^2} \text{ and } \hat{a}_0 = \bar{S}_1 - \hat{a}_1\bar{X}_1$$

Similarly, from equation (ii),

$$\hat{b}_1 = \frac{\sum sx_2}{\sum x_2^{\,2}} \text{ and } \hat{b}_0 = \bar{S}_2 - \hat{b}_1 \bar{X}_2$$

Continue to get Σe_1^2; Σe_2^2

Then $F^* = \Sigma e_2^2 / \Sigma e_1^2$

F-table with $V_1 = V_2 = n\text{-}c\text{-}2K/2$ d.f = 31-9-2(2)/2 = 9

$F^* = 3.18$

Since F*(3.1) < F-table (9), we conclude that the u's are homoscedastic.

Example:

Test using rank correlation

(i) Using a given data set, regress the dependent variable on the independent variable, say X

(ii) obtain the e's

(iii) Rank the X values and the e's (ignoring the sign of e) in either ascending order or descending order and using the formula, $r' = 1 - \frac{6\sum d^2}{n(n^2-1)}$, we get the correlation between the e's and the X's, and conclude appropriately.

Illustration:

The following data is in respect of quantity demanded, D and price, P, of a commodity in a certain period shown in the Table 2.4 below:

Table 2.4

D	24	30	33	36	45	48	54	57	60	63	66	60
P	48	45	45	36	39	30	27	24	18	15	9	6

If D = a0 + a1P + U; $\hat{D} = 74.50 - 0.90P$. The P's and the e's are also tabulated in Table 2.5 below:

Table 2.5

D	24	30	33	36	45	48	54	57	60	63	66	60
$\hat{D}$	31.3	34	34	42.1	39.4	47.5	50.2	52.9	58.3	61	66.4	69.1
E	-7.3	-4	-1	-6.1	5.6	0.5	3.8	4.1	1.7	2	-0.4	-9.1
P	48	45	45	36	39	30	27	24	18	15	9	6

Taking the absolute values of e and rank the respective /e/ and P, we have the following in Table 2.6:

R_P	1	2.5	2.5	5	4	6	7	8	9	10	11	12
R_e	2	5	10	3	4	11	7	6	9	8	12	1
D	-1	2.25	-7.5	2	0	-5	0	2	0	2	-1	11
d^2	1	6.25	56.25	4	0	25	0	4	0	4	1	121

R_P = Rank of P; R_e = Rank of e

$\sum d^2 = 222.5$; n = 12

$$\therefore\ r' = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(222.5)}{12(12^2-1)} = 1 - 0.778 \approx 0.22$$

This value of r[!] Shows a weak correlation between the explanatory variable, P and the residuals, e. thus, there is the absence of heteroscedasticity in the model.

3.1.5 Solutions for Heteroscedasticity

For a given model, heteroscedastic disturbance term can be corrected by transforming the original model n a way to obtain a form in which the transformed disturbance terms have constant variances.

Supposed we concluded that the disturbance terms in the regression model: $Y_i = b_0 + b_1X_i + ui$ are homoscedastic as specified as:

$\sigma u_i = \sigma X_i$,……(2.1.3)

where σ stands simply as a constant of proportionality.

Suppose we have concluded that the disturbance terms in $Y_i = b_0 + b_1X_i + ui$ are heteroscedastic. In this case the technique for estimating the coefficients is straight forward to apply. The object is to re-specify the original model in such a way that the resulting disturbance terms are homoscedastic (i.e., free from heteroscedasticity)

The first step is to divide through the original model ($Y_i = b_0 + b_1X_i + ui$) by X_i, the measure to which $\sigma(u_i)$ is proportional to (2.3). Letting $\in_i = u_i/X_i$, this yield:

$$\left[\frac{Y_i}{X_i}\right] = b_0\left[\frac{1}{X_i}\right] + b_1 + \in_i \text{ ……..(2.1.4)}$$

Equation (2.1.4) is simple regression specification with regressand and regressor given in brackets. Note that the original intercept b_0 is the slope coefficient here, and the original slope b_1 appears as the intercept. Now, since X_i is fixed for each observation, the transformation $\in_i = (1/X_i)u_i$ is a simple case of a linear transformation. Substituting from (2.1.3), we see that:

$$\sigma(\in_i) = \left|\frac{1}{X_i}\right|\sigma(u_i) = \frac{1}{X_i}\sigma X_i = \sigma \text{ ……(2.1.5)}$$

For $X_i > 0$. Thus (2.4) specifies a regression model that is free from heteroscedasticity. Since the ϵ_i satisfy all the regular disturbance terms assumptions, OLS can be used to make estimates b_0 and b_1 that are unbiased, efficient and consistent (Mirer, 1995).

The transformed model can then be estimated using OLS method. The transformation technique to be adopted depends entirely on the nature of the relation between the variances of the disturbance term, σ_u^2 and the values of the explanatory variables(s).

Generally, we transform the original model by dividing the original relationship by the square root of the term, which is responsible for the homoscedasticity.

SELF ASSESSMENT EXERCISE

Define heteroscedasticity and its causes, consequences, detection and solutions.

3.2 Autocorrelation

3.2.1 Meaning of Autocorrelation

The term "autocorrelation" or "serial correlation" refers to a situation which arises when the value of the disturbance term u n any particular period is correlated with its own preceding value (Koutsoyiannis, 1977). In the words of Mirer (1995), autocorrelation is a situation which successive disturbances are related to each other rather than independent. Almost by definition, this is time-series problem, because the ordering of the observations plays a very special role. Accordingly, in this sub-unit, we use t as a subscript (instead of i) to index individual observations.

The circumstance surrounding time-series models make autocorrelation a plausible occurrence in many cases. Recall that one of the factors contributing to the disturbance term in a regression model is error of measurement for the dependent variable. Measurement errors may be serially correlated because data-gathering techniques may be modified gradually over time. A second factor usually contributing to te disturbance term is the exclusion of some unimportant explanatory variables. Each of these is likely to vary systematically with time, and their combination may be serially correlated (Mirer, 1995).

One of the assumptions with respect to the disturbance term discussed n the previous unit s that, the successive values of the random term u are temporarily independent. This assumption implies that the covariance of u_i and u_j is equal to zero:

$$\text{Cov}(u_i\, u_j) = E\{[u_i - E(u_i)][u_j - E(u_j)]\}$$

$$= E(u_i u_j) = 0 \quad (\text{for } i \neq j)$$

If however, this assumption is not satisfied, i.e. $\text{cov}(u_i\, u_j) \neq 0$ we say that there is autocorrelation or serial correlation of the random variable. T is the violation of the zero covariance of the disturbance term.

Auto correlation is a special case of correlation. It is a relationship between the successive values of the same variable. The autocorrelation of the u's is similar to the concept of correlation in general.

Consider the simple linear regresson model:

$$Y_t = b_0 + b_1X_t + u_t \text{ ……..(2.1.6)}$$

Under the assumption of normal regression model, the disturbance terms are independent. This means the probability of different values occurring for one period's disturbance term is not affected by the value that occurred for the previous period's disturbance term.

In order to understand the consequences of this situation and to take corrective measures, we need to develop a formal model of autocorrelation. There are varieties of ways in which the disturbance terms may be related, and each requires a separate analysis we examine only the most common one. The model of ***first-order autocorrelation*** starts with the regression model in (2.6). The disturbance term u_t is assumed to be related to the previous period's disturbance term according to:

$$u_t = \rho u_{t-1} + v_t \quad 0 < \rho < 1 \text{…….(2.1.7)}$$

where ρ = first order autocorrelation coefficient and v = random variable.

This gives the relationship between the u's in the form $u_t = f(u_{t-1})$

3.2.2 Test for Autocorrelation

To test for the presence or absence of autocorrelation in a model, we begin by:

(i) Plotting the values of the regression residuals e's on a scatter diagram.

[Note; the e's are estimates of the true values of u]

If these show certain pattern, it suggests autocorrelation of the u's.

(ii) Another way of obtaining a rough idea of autocorrelation is to plot the residuals e's against time. If the e's in successive periods show a regular pattern, one concludes that auto correlation exists in the given function.

A case of positive autocorrelation is seen in Figure 1.1(a) where several positive e's are followed by several negative e's.

If however, the e's assume positive and negative signs in successive time periods, the autocorrelation is negative. This is shown in Figure 1.1(b)

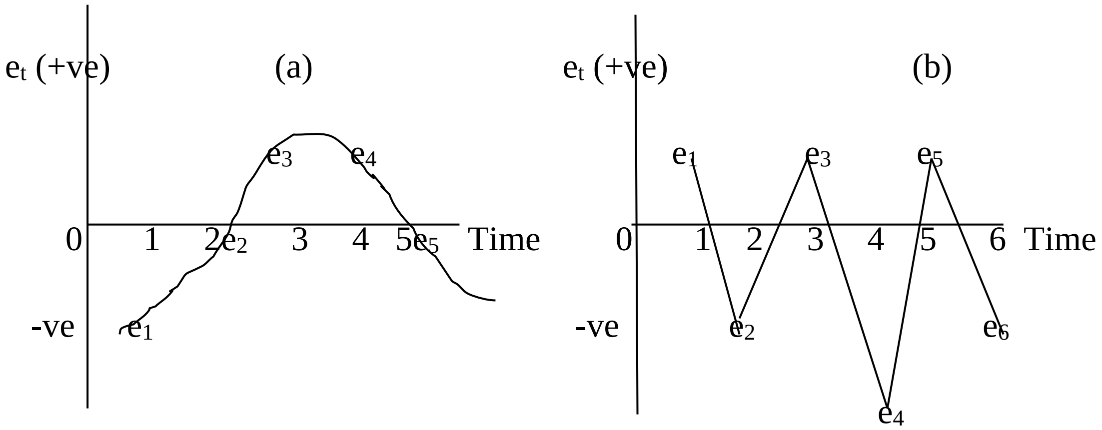


Figure 2.1: +ve autocorrelation -ve autorrelation

i. The degree to which the e's in one period s related to the e's in another period is measured using:

$$r_{e_t e_{t-1}} = \frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^2}\sqrt{\sum e_{t-1}^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1.8)$$

This s called first-order linear autocorrelation.

In the regression result on Table 2.7, the values of e_t and e_{t-1} are given as follows:

Table 2.7

N	e_t	e_{t-1}	e_t^2	e_{t-1}^2	$e_t e_{t-1}$
1	9	-	81	-	-

2	4	9	16	81	36
3	-2	4	4	16	-8
4	11.5	-2	132.25	4	-23
5	4	11.5	16	132.25	46
6	1	4	1	16	4
7	-5.5	1	30.25	1	-5.5
8	1	-5.5	1	30.25	-5.5
N = 8			281.5	280.5	44

$$r_{e_t e_{t-1}} = \frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^{\,2}}\sqrt{\sum e_{t-1}^{\,2}}} = \frac{44}{\sqrt{281.5}\sqrt{280.5}} = 0.$$

This shows that the autocorrelation of the disturbance term is

ii. Formal Test for Autocorrelation: The Durbin Watson Test

Durbin and Watson suggested a test which is applicable to only a small size, with n given in the range; 15 < d < 30.

In this test, the first auto regressive scheme of the form: $u_t = \rho u_{t-1} + v_t$ is adhered to.

The test may be outlined as follows:

H_0: $\rho = 0$, i.e. the u's are not auto correlated with the first-order scheme

H_1: $\rho \neq 0$, i.e. the u's are serially correlated.

The test is administered using the Durbin-Watson statistic:

$$d^* = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^{\,2}} \quad \text{.....(2.1.9)}$$

Where d* is the empirical value of Durbin-Watson statistic. This value (d*) is compared with the theoretical value of "d" with K' degrees of freedom where K' = number of explanatory variable(s), excluding constant term.

The following decision rules apply:

(i) Reject H_0 if $d^* < d_L$, or $d^* > 4d_L$

(ii) Accept H_0 if $d_U < d^* < (4 - d_U)$

Where d_L is the Durbin-Watson lower significance limit

d_U is the Durbin-Watson upper significance limit

Example:

Assume a linear demand function: $D_t = a_0 \pm a_1 P_t + u_t$, where P = price of the commodity and u = error form.

The data in respect of the estimated demand and actual values are given in Table 2.8:

Table 2.8: Calculations for the Test of Autocorrelation

D_t	$\hat{D}_t$	e_t	e_{t-1}	$e_t - e_{t-1}$	e_t^2	$(e_t - e_{t-1})^2$
24	31.3	-7.3	-	-	53.29	-
30	34	-4	-7.3	3.3	16	10.89
33	34	-1	-4	3	1	9
36	42.1	-6.1	-1	-5.1	37.21	26.01
45	39.4	5.6	-6.1	11.7	31.36	136.89
48	47.5	0.5	5.6	-5.1	0.25	26.01
54	50.2	3.8	0.5	3.3	14.44	10.89

57	52.9	4.1	3.8	0.3	16.81	0.09
60	58.3	1.7	4.1	-2.4	2.89	5.76
63	61	2	1.7	0.3	4	0.09
66	66.4	-0.4	2	-2.4	0.16	5.76
60	69.1	-9.1	-0.4	-8.7	82.81	75.69
62	70	-8	-9.1	1.1	64	1.21
65	71.6	-6.6	-8	1.4	43.56	1.96
64	73.5	-9.5	-6.6	-2.9	90.25	8.41
68	72.8	-4.8	-9.5	4.7	23.04	22.09

$\sum e_t^2 = 481.07$; $\sum (e_t - e_{t-1})^2 = 340.75$

$$\therefore\ d^* = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^{\,2}} = \frac{340.75}{481.07} = 0.708$$

From the Durbin-Watson table, $K' = 1$ and n = 16, $d_L = 1.10$ and $d_U = 1.37$

Since $0.70 < d_L$, we reject the null hypothesis and conclude that the error terms are auto correlated.

3.2.3 Sources of Autocorrelation

Autocorrelation of the error terms arises under a number of different circumstances. Some of these include:

(i) Mis-specification of the mathematical form of the model.

Specification error arises here when instead of adopting a linear relation between the Y and the X, a nonlinear relationship is specified. Thus, a mathematical form which differs from the true form of the relationship makes the values of the u's temporarily dependent.

(ii) Omission of explanatory variables

The exclusion of an auto correlated variable from a set of explanatory variables renders the values of u's dependent. This is because the influence of the omitted auto correlated variable will be captured by the error term, hence cause autocorrelation of the disturbance term.

(iii) Mis-specification of the random variable

This is also responsible for the autocorrelation of the disturbance term. If the true random variables are not correctly specified, the successive error terms will be correlated.

(iv) Autocorrelation may also occur as a result of using interpolated values in our estimation. This leads to the interrelationship between the successive value of u's, which exhibits autocorrelation patterns.

3.2.4 Consequences of Autocorrelation

Autocorrelation affects our estimation in the following ways:

(i) The estimates of the parameter will not have statistical bias. The bias is necessary because even in the presence of autocorrelated residuals such parameter estimates will be statistically unbiased. In this respect, their expected values will be equal to the true population parameters (for example, bias in $= E(\hat{b}_1) - b_1 = \frac{\sum (x_i)E(u_i)}{\sum x_i^2} = 0$

(ii) The OLS method cannot be successfully applied to the model whose disturbance terms are auto correlated. The application of this leads to the underestimation of their variances.

(iii) The presence of autocorrelation renders the variance of the random term, u, underestimated. The underestimation becomes pronounced in the case of positive autocorrelation.

(iv) If the values of the u's are serially correlated, the predictions based on the OLS estimates will be inefficient. That is, predictons with needlessly large sampling variances.

It should be noted however, that auto correlation does not affect the properties of unbiasedness nor is the property of consistency necessarily affected.

3.2.5 Solution for Autocorrelation

Problem of autocorrelation is solved using the following methods:

(i) Inclusion of variables in the set of the explanatory variables if the autocorrelation is due to omission of some variables.

(ii) Changing the initial form of a given equation, if the autocorrelation is attributable to mis-specification of the mathematical form of the relationship.

SELF ASSESSMENT EXERCISE

Define autocorrelation as an econometric problem, stating the sources, consequences, detection and solutions. Why is autocorrelation an econometric problem associated with time-series?

3.3 Multicollinearity

3.3.1 Meaning of Multicollinearity

One of the assumptions of least squares (classical linear) regression model is that there is no multicollinearity among the regressors included in the regression model. That I, the explanatory variables are not perfectly linearly correlated ($r_{xixj} \neq 1$).

The term multicollinearity is due to Ragnar Frisch (1934) which originally meant the existence of a "perfect" or exact, linear relationship among some or all explanatory variables of a regression model Gujarati and Sangeetha, (2007).

If the explanatory variables are perfectly linearly correlated, that is, if the correlation coefficient for these variables is equal to unity, the parameters become indeterminate:

it is impossible to obtain numerical values for each parameter separately and the method of least squares breaks down. At the other extreme, if the explanatory variables are not intercorrelated at all (i.e. if the correlation coefficient for these variables is equal to zero), the variables are called orthogonal (Note: orthogonal variables are the variables whose covariance is zero: $\Sigma x_i x_j/n = 0$) and there are no problems concerning the estimates of the coefficients, at least so far as multicollinearity is concerned.

To understand multicollinearity, consider the following model:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + u \quad \text{.....(2.1.10)}$$

Where the hypothetical sample values for X_1 and X_2 are given below:

X_1: 1 2 3 4 5 6

X_2: 2 4 6 8 10 12

From this, we can easily observed that $X_2 = 2X_1$. Therefore, while equation (2.1.10) seems to contain two explanatory variables X_1 and X_2 which are distinct, in fact the information provided by X_2 is not distinct from that of X_1. When this situation occurs, X_1 and X_2 are perfectly collinear (Dimitrios & Stephen, 2006). More formally, two variables X_1 and X_2 are linearly dependent if one variable can be expressed as a linear function of the other variable. When this occurs then the equation:

$$\delta_1 X_1 + \delta_2 X_2 = 0 \text{.....(2.1.11)}$$

can be satisfied for non-zero values of both δ_1 and δ_2. In our example, we have $X_2 = 2X_1$, therefore $(-2)X_1 + (1)X_2 = 0$, so $\delta_1 = -2$ and $\delta_2 = 1$. Obviously, if the only solution in (2.1.11) is $\delta_1 = \delta_2 = 0$ (usually called the trivial solution), the X_1 and X_2 are linearly independent. The absence of perfect multicollinearity requires that (2.1.11) does not hold exactly.

In the case of more than two explanatory variables (let us take five), the case for linear dependence is that one variable can be expressed as an exact linear function of one or more or even all of the other variables. So this time, the expression:

$$\delta_1X_1 + \delta_2X_2 + \delta_3X_3 + \delta_4X_4 + \delta_5X_5 = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1.12)$$

can be satisfied with at least two non-zero coefficients.

An application to better understand this situation can be given by the dummy variable trap. Take for example X_1 to be the intercept 9so as X_1= 1) and X_2, X_3, X_4 and X_5 to be seasonal dummies for quarterly time series data (i.e. X_2 takes the value of 1 for the first quarter, zero otherwise; X_3 takes the value of 1 for the second quarter, zero otherwise and so on). Therefore, n this case we have that, $X_2 + X_3 + X_4 + X_5 = 1$; and because $X_1 = 1$ then $X_1 = X_2 + X_3 + X_4 + X_5$. So the solution is $\delta_1 = 1$, $\delta_2 = -1$, $\delta_3 = -1$, $\delta_4 = -1$, and $\delta_5 = -1$, and this set the variables is linearly dependent.

In practice, neither the orthogonal X's nor perfect collinear X's is often met. In most cases there is some degree of interrelationships among the explanatory variables, due to the interdependence of economic magnitude over time. In this event the simple correlation coefficient will have a value between 0 and 1 and the multicollinearity problems may impair the accuracy and stability of the parameter estimates (Koutsoyiannis, 1977).

3.3.2 Causes of Multicollinearity

The following reasons may be attributed as the causes of multicollinearity in a model:

1. There is the tendency for economic variables to move together over time. Economic magnitudes are influenced by the same factors and in consequences once these determining factors are in operative the economic variables show the same broad pattern of behaviour over time. For example, in the period of boom or rapid economic growth, the basic economic magnitudes grow, although some tend to lag behind others. Thus, incomes, consumption, savings, investment, prices employment, tend to rise in periods of economic expansion and decrease

in periods of economic down turn or recession. These growth and trend factors can seriously caused multicollinearity.

2. The use of lagged values of some explanatory variables as independent factors in the relationship. Naturally, the successive values of certain variables are interrelated since the current value of the variable is partly determined by its own value in the previous period. Thus, multicollinearity is almost certain to exist in distributed lagged model (i.e., $Y_t = f(X_t, X_{t-1}, X_{t-2},\ldots,X_{t-n})$. it should be noted however that multicollinearity is usually connected with time-series data.

3.3.3 Consequences of Multicollinearity

1. If there is a perfect multicollinearity among the X's, their regression coefficients are indeterminate i.e. undefined.

2. The standard errors of these estimates become infinitely large

3. While it is possible to obtain least squares estimates of the regression coefficients, the interpretation of the coefficients will be quite difficult.

It is fairly easy to show that under conditions of perfect multicollinearity, the OLS estimators are not unique. Consider, for example, the model:

$$Y_t = a_1 + a_2 X_2 + a_3 X_3 + u_t \quad \ldots\ldots(2.1.13)$$

Where $X_3 = \delta_1 + \delta_2 X_2$; and δ_1 and δ_2 are known constant. Substituting this into (2.3.13) gives:

$$Y_t = a_1 + a_2 X_2 + a_3(\delta_1 + \delta_2 X_2) + u$$

$$= (a_1 + a_3\delta_1) + (a_2 + a_3\delta_2)X_2 + u$$
$$= \vartheta_1 + \vartheta_\partial X_2 + \varepsilon \quad \ldots\ldots(2.1.14)$$

Where of course $\vartheta_1 = (a_1 + a_3\delta_1)$ and $\vartheta_1 = (a_2 + a_3\delta_2)$.

So what we can estimate from our sample data is that coefficients ϑ_1 and ϑ_2. However, no matter how good the estimates of ϑ_1 and ϑ_2 will be, we will never be able to obtain unique estimate of a_1, a_2 and a_3.

3.3.4 Test for detecting Multicollinearity

There are various methods for detecting the presence of multicollinearity, among these are:

(1) A method base on Frisch confluence analysis

The procedure is to regress the dependent variable on each of the independent variable separately. Thus, we obtain all the elementary regression and we examine their results on the basis of a priori and statistical criteria.

We choose the elementary regression which appears to give the most plausible results, on the basis of these criteria (i.e., on a priori and statistical criteria). Then we gradually insert additional variables and we examine their effects on the individual coefficients, on their standard errors and on the overall R^2. A new variable is classified as useful, superfluous or detrimental as follows:

(a) If it improves R^2 without rendering the individual coefficient unacceptable on a priori consideration, the variable is considered useful and retain as explanatory variable.

(b) If it affects considerably the size or values of the coefficients, it is considered detrimental.

(c) If the individual coefficients are affected n such a way as to be commonly acceptable on theoretical a priori consideration, then we may say that this is a warning that multicollinearity is a serous problem.

(2) The Farrar-Glauber test

This is a statistical test for multicollinearity developed by Farrar and Glauber. It is a set of three tests:

(i) Chi-square test for the detection of the existence and severity of multicollinearity in a function including several explanatory variables.

The basic hypothesis here is:

H_0: the X's are orthogonal

H_1: the X's are not orthogonal

A formula for computing chi-square test is given as:

$$*X^2 = -n[-1-\frac{1}{6}(2k+5).\log_e \begin{bmatrix} 1 & rx_1x_2 \\ rx_1x_2 & 1 \end{bmatrix} \quad \text{.....(2.1.15)}$$

(where $*X^2$ = observed (computed from the sample) value of X^2, n = sample size and k = number of explanatory variables) has a X^2 distribution with degree of freedom of v = 1/2k(k-1); the term in parenthesis is value of the standardize determinant obtained from the partial correlation coefficients.

From the sample data, we obtain the empirical value of $*X^2$ which we compare with the theoretical value of X^2 at a chosen level of significance which may be obtained from a X^2 table.

If the observed $*X^2$ > than the theoretical value of X^2 with v degree of freedom, we reject the assumption of orthogonality, that is we accept that there is multicollinearity in the function. The higher the observed $*X^2$, the more severe the multicollinearity.

If the observed $*X^2 < X^{2,}$ we accept the assumption of orthogonality, that is, e accept that there is no significant multicollinearity in the function.

(ii) The second test is the F-test for locating which variables are multicollinear.

To test the F-test for the location of variables that are collinear, Glauber and Farrar compute the multiple correlation coefficients among the explanatory variables ($R^2_{x_1.x_2x_3\ldots x_k}$, $R^2_{x_2.x_1x_3\ldots x_k....x_k}$, and in general $R^2_{x_i.x_1x_2\ldots x_k}$) and they test the statistical significance of these multiple correlation coefficients with:

For each multiple correlation coefficient, we compute the observed F*,

$$F^* = \frac{(R^2_{x_i.x_1x_2\ldots x_k})/(k-1)}{(1-R^2_{x_i.x_1x_2\ldots x_k})/(n-k)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.1.16)$$

Where n = sample size, k = number of explanatory variables

The hypothesis being tested is:

H_0: $R^2_{x_i.x_1x_2\ldots x_k} = 0$

And the alternative hypothesis is:

H_1: $R^2_{x_i.x_1x_2\ldots x_k} \neq 0$

The observed value F* is compared with theoretical value F (from the F table) with v1=(k-1) and v2=(n-k) degrees of freedom (at a chosen level of significance).

If F* > F-table, we accept that the variable X_i is multicollinear, that is we accept the null hypothesis.

If F* < F-table, we accept that the variable X_i is not multicollinear.

(iii) The third test is a t-test for finding out the pattern of multicollinearity.

To find which variable are responsible for multicollinearity, we compute the partial correlation coefficients among the explanatory variables and test their statistical significance with the t-ststistic.

The basic hypothesis being tested is:

$$H_0 : r_{x_ix_j.x_1x_2\ldots x_k=0}$$

Against the alternative hypothesis

$$H_1 : r_{x_ix_j.x_1x_2\ldots x_k\neq 0}$$

Having estimated the partial correlation coefficients, we test their significance by computing for each of them the statistic:

$$t^* = \frac{(r_{x_i x_j . x_1 x_2 \dots x_k})\sqrt{n-k}}{\sqrt{1-r_{x_i x_j . x_1 x_2 \dots x_k}}} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(2.1.17)$$

Where $r_{x_i x_j . x_1 x_2 \dots x_k}$ denotes the partial correlation coefficient between x_i and x_j.

The observed value t* is compared with the theoretical t value (from the student's t-table) with v = (n-k) degrees of freedom (at a chosen level of significance).

If t* > t, we accept that partial correlation coefficient between the variables x_i and x_j are responsible for multicollinearity in the function.

If t* < t, we accept that x_i and x_j are the cause of multicollinearity in the function since their partial correlation coefficient is not statistically significant.

3.3.5 Solutions for Multicollinearity

The solutions which may be adopted if multicollinearity exist in a function varies depending on the severity of multicollinearity, unavailability of other sources of data, on the importance of factors which are collinear, on the purpose for which a function is being estimated etc.

1. Increase the size of the sample. Multicollinearity may be avoided or reduced if the size of the sample is increased by gathering more observations. Thus, by increasing the size of the sample, higher covariances among estimated parameters resulting from multicollinearity in an equation can be reduced because these covariances are inversely proportional to the sample size. However, this is possible if the source of multicollinearity comes from errors of measurement as well as the original sample size.

2. Introduction of additional equations in the model. To solve the problem of multicollinearity, we can introduce additional equations into the model to express meaningfully the relationship between the multicollinear X's.

3. Application of methods of incorporating extraneous quantitative information. This include the method of restricted least squares, pooling cross-section and time-series, Durbin version of generalised least squares, mixed estimation techniques as propounded by Theil and Goldberger.

4. Substitution of lagged variables for other explanatory variables in distributed lag models. Multicollinearity may be avoided by substituting for a single lagged value of the dependent variable as suggested by Koyck.

5. Application of principal component analysis.

SELF ASSESSMENT EXERCISE

Substantiate the meaning of multicollinearity, it causes, consequences, detection and remedial measures.

4.0 CONCLUSION

This sub-unit discussed the incidence where the explanatory variables collinear in a function. This is a violation of one of the assumptions of the least squares regression model. The consequence of multicollinearity is that, their regression coefficients are indeterminate and their standard errors are not defined. If multi co linearity is high but not perfect, estimation of regression coefficients is possible but their standard errors tend to be large.

5.0 SUMMARY

One of the assumptions of the least squares regression model is that there is no multicollinearity among the regressors, the X's. Broadly interpreted, refers to a situation where there is either an exact or approximately exact linear relationship among the

explanatory variables. The unit further threw light on the causes of multicollinearity, consequences, detection and solutions for multicollinearity.

6.0 TUTOR-MARKED ASSIGNMENT

1. The following table shows the annual consumption and disposable income of Nigeria for a given period.

Table 2.9: Income and consumption expenditure n Nigeria (N'000million)

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
C_t	26	29	35	39	42	46	50	54	60	64	69	73
Y_{dt}	38	43	53	60	66	71	77	86	94	102	109	115

(a) Estimate the savings function $S_t = f(Y_{d(t)})$.

(b) Test for heteroscedasticity using spearman's rank correlation coefficient

2. Table 2.10 below shows the annual consumption (C) and disposable income (Y_d) of Nigeria (in N million).

Table 2.10: Consumption and disposable (N' million)

Year	C	Y_d	Year	C	Y_d
2008	11,378	11,617	2014	20,074	21,512
2009	13,012	13,297	2015	21,439	23,124
2010	15,263	15,790	2016	22,833	24,724
2011	16,873	18,017	2017	24,205	26,175
2012	17.764	19,314	2018	25,307	27,219

2013	18,857	20,198	2019	27,020	28,915

Ordinary least squares (OLS) application on the data in Table 2.10 yields the following outcomes:

$$\hat{C} = 8,526 + 0.65Y_d \quad r^2 = 0.953$$

Find the residuals and test for autocorrelation.

3. The following table shows time-series on three variables, Y, X_1, X_2 in arbitrary units.

   Table 2.11: Calculations for the Test of Autocorrelation

Y	6	6	6.5	7.1	7.2	7.6	8	9	9	9.3
X_1	40.1	40.3	47.5	49.2	52.3	58	61.3	62.5	64.7	66.8
X_2	5.5	4.7	5.2	6.8	7.3	8.7	10.2	14.1	17.1	21.3

   (a) Test for multicollinearity with any appropriate method.

   (b) How does multicollinearity affect the parameter estimates?

7.0 REFERENCES/FURTHER READINGS

Astrios, D. & Hall, S.G. (2006). Applied econometrics: A modern Approach (Revised Edition), Palgrave Macmillan, New York.

Gujarati, D.N. (2006). Essentials of Econometrics (Third Edition). McGraw-Hill, New York.

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

Koutsoyannis, A. (1977). Theory of Econometrics (Second Edition). PALGRAVE. New York.

Mirer, T.W. (1995). Economic Statistics and Econometrics (Third Edition), Prentice-Hall Inc, London.

Ragnar, F. (1934). Statistical Confluence Analysis by Means of Complete Regresson Systems, Institute of Economics, Olso University, publ. no. 5.

UNIT TWO: BASIC IDEA OF IDENTIFICATION PROBLEM

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 Meaning of Identification problem
    3.2 Implications of Identification problem
    3.3 Formal Rules (Conditions) for Identification
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

The crux of the identification problem is seen in the famous demand-and-supply model discussed in the previous units. Suppose that we have time-series data on quantity and

price only and no additional information (such as income of the consumer, price prevailing in the previous period, and weather condition). The identification problem then consists in seeking an answer to this question: Given only the data on quantity and price, how do we know whether we are estimating the demand function or the supply function? Alternatively, if we think we are fitting a demand function, how do we guarantee that it is, in fact, the demand function that we are estimating and not something else? A moment of reflection will reveal that an answer to the preceding question is necessary before one proceeds to estimate the parameters of our demand function. In this unit, we shall show how the identification problem is resolved.

2.0 OJECTIVES

At the end of this unit, students should be able to:

- Define what identification problem is all about.
- State the implications of identification problems
- State the formal rules or conditions for identification.

3.0 MAIN CONTENT

3.1 Meaning of identification problem

Identification problem associated with model formulation. Identification problem arises if a given model is not correctly specified. This is sequel to the violation of the assumption that the relationship to be estimated must have a unique mathematical form. If this is violated truly, then the relationship to be estimated will contain the same variable(s) as any other equation one is estimating. In the face of unidentified model, the estimates of the parameters of the relationship between the variable(s) measured in samples relate to the model in question or to another model or to a mixture of models (Gujarati and Sangeetha, 2007).

Let us illustrate this concept of non-identifiability by making reference to a model of market equilibrium:

$D = b_0 + b_1P + u$……..(iii) Demand function

$S = a_0 + a_1P + v$……(iv) Supply function

$D = S$ ……(v) Clearance equation

Here, we cannot identify the two models in equation (iii) and (iv) since in both, quantities Q = f(P). Hence, we may not know exactly which function one is estimating, whether it is the demand or the supply function.

To be sure of the function one is estimating, there is the need to examine identification condition to enable us judge the precise equation.

To achieve this objective, we need further information on factors affecting the demand and the supply function separately (shift factors). This will help us to state clearly the function one is dealing with.

In the demand function, other determining factors such as consumers' income, taste, price of other commodities (substitute or complement), fashion, etc, should be included to differentiate this function from the supply function determined by factors such as weather, technological change, government policy etc.

The stability of each of these functions depends, to a large extent, on the aforementioned factors.

We can state here that a model is identified if it is in unique mathematical form such that its estimates can be uniquely obtained from the given data. As a corollary, a model is not identified, if its parameter estimates relate to the model in question or to another model or both.

Thus, for a set of simultaneous equations, the identification of the entire set requires that the model be complete.

The completeness of such models implies that it should contain at least as many independent equations as the endogenous variables.

Referring to the market equilibrium previously indicated, the given model is complete, because it contains three equations and three endogenous variables: D, S, and P. However, they are not identified, as previously highlighted with reasons.

We generalised by stating that a function belonging to a system of simultaneous equation is identified, if it has a unique statistical form, meaning that there is no other equation in the system which contains the same variable as the function in question Koutsoyiannis, 1977).

Thus, we can state that:

(i) The identification of a system requires that each of the equations in the system be identified.

(ii) The parameters of the given equation be identified, (Note that in the previous example, the parameters of D(b_0 and b_1) and S(a_0 and a_1) cannot be statistically since one is not really sure of the function one is dealing with.

SELF ASSESSMENT EXERCISE

Briefly discuss identification as a problem associated with model formulation.

3.2 Implications of Identification

Identification is closely related to the estimation of a model although it is a problem associated with model formulation.

(a) Once an equation (or model) is underidentified, it is impossible to estimate all its parameters with any econometric technique.

(b) If an equation is identified, its coefficients can, in general be estimated. In particular: (i) If the equation is exactly identified, the appropriate method to be used for its estimation is the method of indirect least squares (ILS). (ii) If the equation is over-identified, indirect least squares cannot be applied, because it will not yield unique

estimates of the structural parameters. There are various other methods which can be applied in this case, for example, two-stage least squares (2SLS), or maximum likelihood methods.

SELF ASSESSMENT EXERCISE

What are the implications of identification problem in a regression model (or equation)?

3.3 Formal Rules (Conditions) for Identification

To establish identification, therefore, two rules are adhered to:

(a) Order condition (b) Rank condition

In econometrics, there are two possible situations of identifiability of any possible situation of equation:

(i) Under-identification

(ii) Identification

An under indentified equation exists if the mathematical form of the equation is not unique. A system is under identified when one or more of its equations are under identified. For such an equation or model, it is not possible to estimate all its parameters using any econometric technique.

An identified equation on the other hand has a unique statistical form. An over identified equation may either come n the form of exact identification or over identification condition. A given system, therefore, is identified if all its equations are identified.

For an exact identification, we estimate such equation using the indirect least squares (ILS). In the case of over identification, we use the two stage least squares (2SLS) or maximum likelihood methods.

Note that identification problem is peculiar to only those equations which contained coefficient that needs to be estimated (From a given set of data).

Therefore, identification of a model can best be established by the examination of the specification of the structural model. In applying the identification rules, we either ignore the constant term or retain it, and include in the set of variables, a dummy variable which takes the values 1. In this context, we shall ignore the constant term.

For an equation (or model) to be identified the conditions are:

1. **Order Condition**

This condition is based on the number of variables included and those excluded from a particular equation. This is a necessary condition, although not sufficient for complete identification. This implies that the order condition for identification may be satisfied if the equation is not identified. The condition stated that: for an equation to be identified, the number of variables (endogenous and exogenous) excluded from it must be equal to or greater than the number of endogenous variables in the model less one.

Since the number of endogenous variables equals the number of equation in a complete model, the above condition may also be equivalently stated in this form: an equation is identified, if the total number of variables excluded from it but included in the other equations is at least as great as the number of equations of the system less one.

Symbolically, the order condition for identification is given as:

$$K - M \geq G - 1 \qquad (2.2.1)$$

Where, K = total number of variables (endogenous and exogenous) in the system

M = number of variables included in a particular equation

G = total number of equations in the system

So that K - M = excluded variables in a particular equation

G – 1 = total number of equations less one.

Note: the order condition for identification is necessary for aa relation to be identified but it is not sufficient for identification.

2. The Rank Condition

This condition maintains that in a system of G-equations, any particular equation is identified if and only if it is possible to construct at least one non-zero determinant of G – 1 from the coefficient of the variables excluded from that particular equation, but contained in the other equations of the model.

In order to identify a particular equation, therefore, the following should be followed:

(i) Write out the parameters of all the equations of the model in a separate table, bearing in mind to assign zero to each parameter of a variable excluded from a given equation. Make a table of coefficients of the equations which is being examined for identification.

(ii) Strike out the columns in which a non-zero coefficient of the equation being examined appears. (By deleting relevant row and columns, we are left with the coefficients of variables not in the other equations of the model).

(iii) From the determinant(s) of order (G – 1) and examine their values.

-If at least one of these determinants is non-zero, the equation is identified.

-If however, all the determinants of order G – 1) are zero, the equation is under identified and we make conclusion.

(iv) Once the equation is identified, we then move to the order condition to determine the nature of identification.

Example:

Examine the identification state of the following model.

$C_t = a_0 - a_1Y_t – a_2T_t + u_1$……..(i)

$I_t = b_0 + b_1Y_t + b_2Y_{t-1} + b_3R_t + u_2$……(ii)

$Y_t = C_t + I_t + G_t$……...(iii)

Solution:

Re-writing the above equations, we have:

$-C_t + a_0 - a_1Y_t - a_2T_t + u_1 = 0$

$-I_t + b_0 + b_1Y_t + b_2Y_{t-1} + b_3R_t + u_2 = 0$

$-Y_t\ C_t + I_t + G_t = 0$

We make table for the model as shown:

Table 2.12: structural parameters

Equation	C_t	Y_t	I_t	T_t	Y_{t-1}	R_t	G_t
(i)	-1	a_1	0	$-a_2$	0	0	0
(ii)	0	b_1	-1	0	b_2	b_3	0
(iii)	1	-1	1	0	0	0	1

To identify equation (i), delete the first row containing the coefficient of equation and also delete the column of the table with non-zero coefficients of the equation to be identified.

Table 2.13: identifying equation (i)

Equation	C_t	Y_t	I_t	T_t	Y_{t-1}	R_t	G_t
(i)	1	a_1	0	$-a_2$	0	0	0

(ii)	0	b_1	-1	0	b_2	b_3	0
(iii)	1	-1	1	0	0	0	1

Thus, we are left with:

I_t	Y_{t-1}	R_t	G_t
-1	b_2	b_3	0
1	0	0	1

i.e. coefficients of the excluded variables (in tabular form).

Determinant of G – 1 i.e. 3 -1 are:

$\Delta_1 = \begin{vmatrix} -1 & b_2 \\ 1 & 0 \end{vmatrix} = b_2 \quad \Delta_2 = \begin{vmatrix} -1 & b_3 \\ 1 & 0 \end{vmatrix} = -b_3 \quad \Delta_3 = \begin{vmatrix} -1 & 0 \\ 0 & 1 \end{vmatrix}$

Since we have at least a non-zero determinant, equation (i0 is identified.

For the nature of identification, we use order condition as follows:

$K - M \geq G - 1$; $K = 7$, $M = 3$, $G = 3$

$\therefore 7 - 3 \geq 3 - 1$, $4 > 2$ (a case of over identification)

Using similar procedure as in (i) above, equations (ii) and (iii) are respectively identified as follows:

Equation (i) coefficients of excluded variables are:

C_t	T_t	G_t

-1	$-a_2$	0
1	0	1

And G – 1 determinants are:

$$\Delta_1 = \begin{vmatrix} -1 & -a_2 \\ 1 & 0 \end{vmatrix} = a_2 \quad \Delta_2 = \begin{vmatrix} -1 & 0 \\ 1 & 1 \end{vmatrix} = -1 \quad \Delta_3 = \begin{vmatrix} -a_2 & 0 \\ 0 & 1 \end{vmatrix} = -a_2$$

Since we have at least a non-zero determinant, equation (ii0 is identified.

Using order condition: K – M ≥ G – 1; K = 7, M = 4, G = 3

∴ 7 – 4 > 3 – 1 ; 3 > 2 [equation (ii) is over-identified]

Similarly, in equation (iii), coefficients of excluded variables are:

T_t	Y_{t-1}	R_t
$-a_2$	0	0
0	b_2	b_3

So that determinants of G – 1 are:

$$\Delta_1 = \begin{vmatrix} -a_2 & 0 \\ 0 & b_2 \end{vmatrix} = a_2b_2 \quad \Delta_2 = \begin{vmatrix} -a_2 & 0 \\ 0 & b_3 \end{vmatrix} = -a_2b_3 \quad \Delta_3 = \begin{vmatrix} 0 & 0 \\ b_2 & b_3 \end{vmatrix} = 0$$

Again equation (iii) is identified.

Nature of identification is given as: K – M ≥ G – 1; K = 7, M = 4, G = 3

∴ 7 – 4 > 3 – 1, 3 > 2 [Equation (iii) is over identified].

SELF ASSESSMENT EXERCISE

1. Discuss the two conditions that must be satisfied for an equation in a system of equations to be identified.

4.0 CONCLUSION

The problem of identification precedes the problem of estimation. The identification problem asks whether one can obtain numerical estimates of structural coefficients from the estimated reduced form coefficients. If this can be done, an equation in a system of simultaneous equations is identified. If this cannot be done, that equation is un-or under-identified. Identification problem arises because the set of data may be compatible with different sets of structural coefficients, that is, different models. Thus, in regression of quantity on price only, it is difficult to tell whether one is estimating the supply function or the demand function, because price and quantity enter both equations. An equation can be just (exact) identified or over-identified. In the exact identification, unique values of the structural coefficients can be obtained using the indirect least squares (ILS); in the over-identification, there may be more than one value for one or more structural parameters [the parameters can be obtained using the two stage least squares (2SLS)]. To identify an equation, two conditions must be met which include the order condition and the rank condition.

5.0 SUMMARY

This unit discussed the idea of identification problem which has to do with model formulation rather than of model estimation or evaluation. The unit added that an equation from a system of simultaneous equation is identified if it is in a unique statistical form, guaranteeing unique estimates of its parameters to be subsequently made from sample data. An equation can be just (exact) identified or over-identified. In the exact identification, unique values of the structural coefficients can be obtained using the indirect least squares (ILS); in the over-identification, there may be more than one value for one or more structural parameters [the parameters can be obtained using the two stage least squares (2SLS)]. To identify an equation, two conditions must be met which include the order condition and the rank condition.

6.0 TUTOR-MARKED ASSIGNMENT

(1) Consider the following structural model and examine the identification state of the model:

$Y_1 = 2Y_2 – X_1 + 2X_3 – X_4 + Z_1 + u$……..(i)

$Y_2 = X_1 + 2Y_1 – 2X_1 – Y_1 + 2X_2 + v$……...(ii)

$Y_3 = Y_1 –Y_2 + 3Y_1 – 3X_2 – X_4 + w$…….(iii)

(2) State and explain the estimation techniques for an exact and over-identified equation in a system of simultaneous model.

7.0 REFERENCES/FURTHER READINGS

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

UNIT THREE: DUMMY VARIABLE AND DISTRIBUTED LAG MODELS

1.0 Introduction

2.0 Objectives

3.0 Main Contents
3.1 Dummy Variable in the Regressors
3.2 Dummy Variable in the Regressand
3.3 Lag Variables and Distributed Lag Models

4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

It sometimes is the case that the economic process under study leads to outcomes that are categorical, rather than measureable. For example, a person might taken private or public transportation to work, a firm might go bankrupt or not, and a high school senior might go to college or not. In each case the outcome variable Y can be coded as a binary

dummy variable: one outcome is assigned the value 0, and the other is assigned the value 1. Similarly, n regression analysis, the dependent variable, or the regressand, is frequently influenced not only by ratio scale variables (e.g. income, output, price, costs, height, temperature) but also by variables that are essentially qualitative, or nominal scale, in nature, such as sex, race, colour, religion, nationality, geographical region, political upheavals and party affiliations. In this unit, we shall begin our discussion with dummy variables in the regressors, and then followed with dummy variables in the regressand.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- State the nature of dummy variables.
- Compute ANOVA models.
- Estimate ANCOVA models.
- Analyse regression with a mixture of quantitative and qualitative regressors.
- Examine the use of dummy variables in seasonal analysis

3.0 MAIN CONTENT

3.1 Nature of Dummy Variable

In some regression analysis, the dependent variable is influenced by categorical or nominal scale, in nature, such as sex, race, colour, religion, nationality, geographical region, political upheavals and party affiliations. For example, holding all other factors constant, female workers are found to earn less than their male counterparts. This pattern may result from sex discrimination. But whatever the reason, qualitative variables such as sex seems to influence the regressand and clearly should be included among the explanatory variables, or the regressors.

Since such variables usually indicate the presence or absence of a "quality" or an attribute, such as male or female, APC or PDP party, North or South, they are essentially nominal scale variables. One way we could "quantify" such attributes is by constructing artificial variables that take on values of 1 or 0, 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute.

For example1 may indicate that a person is a female and 0, may designate a male. Variables that assume such 0 and 1 values are called **dummy variables**. Such variables are thus essentially a device to classify data into mutually exclusive categories such as male or female.

3.1.1 Regression with Dummy Variables as Regressors

Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact, a regression model may contain regressors that are all exclusively dummy, or qualitative, in nature. Such models are called **Analysis of Variance (ANOVA) models**.

To illustrate the ANOVA models, consider the following example:

We examine this by looking at public school teachers' salaries by geographical region. Table 1.0 in the appendix gives data on average salary (in naira) of public school teachers in 51 local government Areas (LGAs) in Nigeria. These 51 LGAs are classified into three geographical regions (i) North (21 states in all) (ii) South (17 states in all) and (iii) Middle Belt (13 states in all).

Suppose we want to find out the average annual salary (AAS) of public school teachers differs among the three geographical regions.

There are various statistical techniques to compare the two or more mean values of these categories, which generally go by the name of analysis of variance. By the same objective can be achieved by regression analysis.

Note: $D_2 = 1$ for states in the North; 0 otherwise.

D_3 = 1 for state in the South; 0 otherwise

To see this, consider the following model:

$$Y_i = b_1 + b_2D_{2i} + b_3D_{3i} + u_i \text{……..(2.3.1)}$$

Where Y_i = average salary of public school teachers in state i

D_{2i} = 1 if the state is in the North; 0 otherwise (i.e. in other regions of the country).

D_{3i} = 1 if the state is in the south; 0 otherwise (i.e. in other regions of the country).

Equation (2.3.1) shows a multiple regression where the explanatory variables are dummy variables. Assuming that the error term satisfies the usual OLS assumptions, on taking expectation of (2..3.20) on both sides, we obtain:

Mean salary of public school teachers in the North:

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = b_1 + b_2$$

Mean salary of public school teachers in the South:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = b_1 + b_3$$

The mean salary of public school teachers in the Middle Belt:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = b_1$$

This means that the mean salary of public school teachers in the Middle belt is given by the intercept b_1, in the multiple regression of (2.20), and the "slope" coefficient b_2 and b_3 tell by much the mean salaries of teachers in the North and n the South differ from the mean salary of teachers in the Middle Belt.

Assuming the regression equation of (2.3.20) is given as follows:

$$\hat{Y}_i = 26{,}158.62 - 1734.473D_{2i} - 3264.615D_{3i}$$

SEE = (1128.523) (1435.953) (1499.615)

t = [23.1759] [-1.2078] [-2.1776]

P-value = (0.0000) (0.2330) (0.0349)

We can deduce from the regression that the mean salary teachers in the Middle belt is N26,158, that of teachers in the North is lower by about N1734, and that of teachers in the South is lower by about N3264. The actual mean salaries in the last two regions can be obtained by adding these differential salaries to the mean salary of teachers in the Middle Belt. Doing this, we will find that the mean salaries in the North and South regions are about N24,424 and N22,894.

But how do we know that these mean salaries are statistically different from the mean salary of teachers in the Middle Belt, the bench mark category? All we need do is to check the significance of the respective slope coefficients. From the regression, the estimated coefficient for the North is not statistically significant, as the p=value is 0.2330, where as that of the south is statistically significant, as the p-value is 0.0349. Therefore the overall conclusion is that North is about the same with middle belt but the mean salary of teachers in the South is statistically lower by about N3265.

A word of caution:

i. If we introduce three dummies for the three regions, we will run into the problem of perfect multicollinearity. Therefore, if qualitative variables have m categories, introduce only (m-1) dummy variables. However if we introduce three variables for the three regions, then we should not include the intercept as a benchmark category.

ii. The category for which a dummy is not assigned (in our example, Middle Belt)is known as the base, benchmark, control, comparison, reference, or omitted category.

iii. The intercept value (b_1) represents the mean value of the benchmark category.

iv. The coefficient attached to dummy variables in (2.3.20) are known as the differential coefficients because they tell by how much the value of the intercept that receives the value of 1 differs from the intercept coefficient of the benchmark category.

v. The choice of a benchmark category is strictly up to the researcher.

3.1.2 Regression with a Mixture of Quantitative and Qualitative Regressors

ANOVA models of the type discussed in the preceding sub-unit, although common in fields such as sociology, psychology, education, and market research, are not common in economics. Typically, in most economic research a regression model contains some explanatory variables that are quantitative and qualitative. Regression models containing a mix of quantitative and qualitative variables are called **analysis covariance (ANCOVA) models**. They are an extension of ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called covariates or control variables, in the model that includes both quantitative and qualitative, or dummy, regressors.

As an illustration, let us consider, let us consider, the three regions teachers' salaries by maintaining that that the mean salary of public school teachers may not be different n the three regions if we take into consideration any variables that cannot be standardize across the regions. Consider, for example, the variable, expenditure on public schools by the state and local governments. Now, we developed the following model:

$$Y_i = b_1 + b_2D_{2i} + b_3D_{3i} + b_4X_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2.3.2)$$

Where Y_i = average salary of public school teachers in state i

X_i = spending on public schools per pupil (N)

D_{2i} = 1 if the state is in the North; 0 otherwise (i.e. in other regions of the country).

D_{3i} = 1 if the state is in the south; 0 otherwise (i.e. in other regions of the country).

The following regression results emerged after introducing X in the model:

$$\hat{Y}_i = 13269.11 - 1673.514D_{2i} - 1144.517D_{3i} + 3.2889Xi$$

SEE = (1395.056) (801.1703) (861.1182) (0.3176)

t = (9.5115)* (-2.0089)* (-1.3286)** (10.3539)*

where * indicates p-value less than 5 percent and ** indicates p-value greater than 5 percent. The result suggests that as public expenditure goes up by a naira, on average, a public school teacher's salary goes up by N3.29. Controlling for spending on education, we can now see that the differential intercept coefficient is significant for the North region but not for the south. These results are different from the previous one without X because we did not account for the covariate, differences in per pupil public spending on education.

3.2 Regression with Dummy variable as Regressand

So far we have considered regression models in which the regressand is quantitative while regressors are quantitative or qualitative or both. But there are situations where the regressand can be qualitative or dummy.

Suppose we are interested in studying labour force participation (LFP) decision of adult males. Since an adult is either in the labour force or not, LFP is a 'yes' or 'no' decision. Hence, the response variable, or regressand, can take only two values, say, 1 if the person is in the labour force and 0 if he is not. In other words, the regressand is a binary or dichotomous variable. In the theory of labour economics, LPF depends on unemployment rate, average wage rate, education, family income etc.

Another example of a binary variable is that, consider Nigerian presidential elections. Assume that there are only two political parties, ABC and XYZ. The regressand here is

vote choice between the two political parties. Suppose we let Y = 1, if the vote is for ABC candidate, and Y = 0, if the vote is for XYZ candidate. Some of the variables that can enter into the vote choice function are growth rate of GDP, unemployment and inflation rates, whether the candidate is running for re-election etc.

Other examples where the regressand is a qualitative in nature include: a family either owns a house or it does not, both husband and wife are n the labour force or one spouse is.

In addition, we do not have to restrict our response variable to yes/no or dichotonomous category only. Returning to our example on presidential election, suppose there are three parties: ABC, XYZ and Independent. The response variable here is trichotomous. In general, we can have polychotomous (or multiple) response variable.

Our emphasis now is to first consider the dichotomous regressand and then consider various extension of the basic model. It is worthy to note that in a model where the Y is quantitative, the objective is to estimate its expected or mean value given the values of the regressors, i.e. $E(Y_i \mid X_{1i}, X_{2i},\ldots X_{ki})$, where the X's are regressors both quantitative and qualitative. However, in models where Y is qualitative, the objective is to find the probability of something happening, such as voting ABC party, or owning a house, belong to a union etc. hence qualitative response regression models are often known as probability models.

Our study of qualitative response models begins with the binary response regression model. There are three approaches to developing a probability model for a binary response variable:

1. The linear probability model (LPM)
2. The logit model
3. The probit model

1. The Linear Probability Model (LPM)

Consider the following regression model:

$Y_i = b_1 + b_2X_i + u_i$……………………………………………….(2.3.3)

Where X = family income and Y = 1 if the family owns a house and 0 if it does not own a house.

Equation (2.22) looks like a typical linear model but because the regressand s binary, or dichotomous, it is called a linear probability model (LPM). This is because the conditional expectation of Y_i given X_i, $E(Y_i \mid X_i)$, can be interpreted as the conditional probability that the event will occur given X_i, $Pr(Y_i = 1 \mid X_i)$. thus, in our example, $E(Y_i \mid X_i)$ gives the probability of a family owning a house and whose income is the given amount of X_i.

Assuming $E(u_i)$, as usual (to obtain unbiased estimators, we obtain:

$E(Y_i \mid X_i) = b_1 + b_2X_i$…………………………………………….(2.3.4)

If P_i = probability that $Y_i = 1$ (that is, that 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 2.14: Probability distribution

Y_i	Probability
0	$1 - P_i$
1	P_i
Total	1

This implies that Y_i follows a ***Bernoulli probability distribution***.

Mathematically, we obtain:

$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i$……………………………………..(2.3.5)

Comparing (2.24) with (2.25), we can equate, thus:

$$E(Y_i \mid X_i) = b_1 + b_2X_i = P_i \text{……………………………………..(2.3.6)}$$

Equation (2.3.6) implies that, the conditional expectation of the model (2.22) 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. 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 the variance is *np*(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:

$$0 \leq (Y_i \mid X_i) \leq 1 \text{……………………………………………(2.3.7)}$$

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 possesses several problems, which are as follows:

i. Non-Normality of the disturbance u_i

ii. Heteroscedasticity variance of the disturbance

iii. Non-fulfilment of $0 \leq (Y_i \mid X_i) \leq 1$

iv. Questionable R^2 as a measure of goodness of fit

3.2.1 Alternative to Linear Probability Model (LPM)

As highlighted in the previous sub-unit, the LPM has several problems. However, these problems can be overcome. For instance, the weighted least squares (WLS) can be used

to resolve the problem of heteroscedasticity or one can increase the sample size to solve the problem of non-normality.

Even with these solutions, the major issue with LPM is that, it is not logically a very attractive model because it assumes that $P_i = E(Y = 1 \mid X)$ increases linearly with X, that is, the marginal increment remains constant throughout. This seems practically unrealistic. In reality, one would expect that P_i is nonlinearly related to X_i. at a 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 is needed is a probability with these two characteristic: (a) As X_i increases, $Pi = E(Y = 1 \mid X)$ increases but never steps outside the 0 – 1 interval (b) the relationship between P_i and X_i is nonlinear, that is, "one which approaches zero at slower rates X_i gets small and approaches one at slower and slower rates as Xi gets very large (Gujarati & Sangeetha, 2007).

Figure 2.2 below geometrically shows a model that the probability lies between 0 and 1, and that it varies nonlinearly with X. the s-shape curve in the figure very much resembles the **cumulative distribution function** (CDF) of random variable. Therefore, one can use the CDF to model regressions where the response is dichotomous.

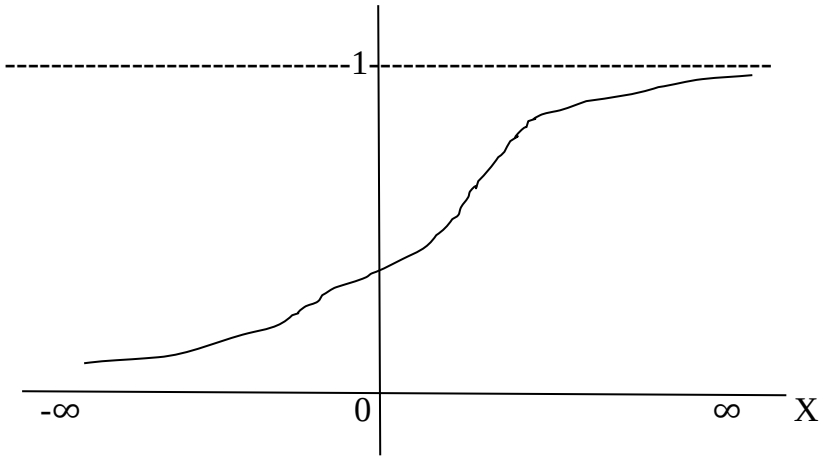


Figure 2.2: A cumulative distribution function (CDF)

The CDF normally chosen to represent the 0 – 1 response models are (1) the logistic (logit) and (2) the normal (probit) model.

(1) The Logit Model

Using our example of house ownership, we can explain the basic ideas underlying the logit model. Recall that in explaining home ownership n relation to income, the LPM was:

$$P_i = E(Y = 1 \mid X_i) = b_1 + b_2X_i \text{……..(2.3.8)}$$

Where X is income and Y = 1 means family owns a house. But not consider the following representation of home ownership:

$$P_i = E\left(Y = 1 \mid X_i\right) = \frac{1}{1+e^{-(b_1+b_2x_i)}} \text{ ……(2.3.9)}$$

This can be written as:

$$P_i = \frac{1}{1+e^{-z_i}} = \frac{e^z}{1+e^z} \text{ ……(2.3.10)}$$

Where $Z_i = b_1 + b_2X_i$.

Equation (2.300 is what is known as the cumulative distribution function (CDF) (Kramer, 1991).

It is easy to verify that as Z_i ranges from -∞ to +∞, P_i ranges from 0 – 1 and that P_i is nonlinearly related to Z_i (i.e. X_i), thus, satisfying the two conditions considered earlier. However, in satisfying these two requirements, we have created a problem of estimation because P_i is nonlinearly not only in X but in the b's as shown in (2.3.9). This further means that the OLS cannot be used to estimate the parameters. But this problem can be solved by linearizing (2.3.9).

If P_i, the probability of owning a house is given by (2.3.10), the (1 – P), the probability of not owning a house, is:

$$1-P=\frac{1}{1+e^{z_i}} \text{ ……(2.3.11)}$$

Therefore, we can write:

$$\frac{P_i}{1-P_i}=\frac{1+e^{z_i}}{1+e^{-z_i}}=e^{z_i} \text{ ……(2.3.12)}$$

The ration, $P_i/1\text{-}P_i$ is simple 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.

If we take the natural log of (2.3.12), we obtain a very interesting result, namely:

$$Li=\ln\left(\frac{P_i}{1-P_i}\right)=Z_i=b_1+b_2X_i \text{ ……..……(2.3.13)}$$

That is, L, the log of the odds ratio, is not only linear in X, but also linear in the parameters. L is called the logit, and hence the name logit models for models like (2.3.13).

(2) The Probit Model

As we have noted, to explain the behaviour of a dichotomous variable we will have to use a suitable chosen CDF. The logit model uses the cumulative logistic function as shown n (2.3.9). But this is not the only CDF that one can apply. In some applications, the normal CDF has been found useful. The estimating model emerges from the normal CDF is popularly called the probit model. In principle, one could substitute the normal CDF in place of the logistic CDF in (2.3.9). Instead of following this route, we will present probit model based on utility theory, or rational choice perspective on behaviour.

Assume that in our house ownership example, the decision to of the i^{th} family to own a house or not depends on an unobservable utility index, I_i (also called 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:

$$I_i = b_1 + b_2X_i \text{……..(2.3.14)}$$

Where X_i is the income of the i^{th} family. How is the (unobservable) index related to the acyual decision to own a house? As before, let Y = 1 if 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, called it I_i^*, like the I_i is not observable, but if we assume it is normally distributed with the same mean and variance, it is possible not only to get some information about the unobservable index itself.

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:

$$P_i = P(= 1 \mid X) = p(I_i^* \leq I_i) = P(Z_i \leq b_1 + b_2X_i) = F(b_1 + b_2X_i) \text{……(2.3.15)}$$

Where P(Y = 1 | X) means the probability that an event occurs given the values(s) of the X,. and where Z_i is the standard normal variable, i.e., $Z \sim N(0,\sigma^2)$. F is the standard normal CDF, which written explicitly in the present context is:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-z^2/2} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b_i + b_2X_i} e^{-z^2/2} dz \text{ ……(2.3.16)}$$

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 $-\infty$ to I_i as shown in Figure 2.3a.

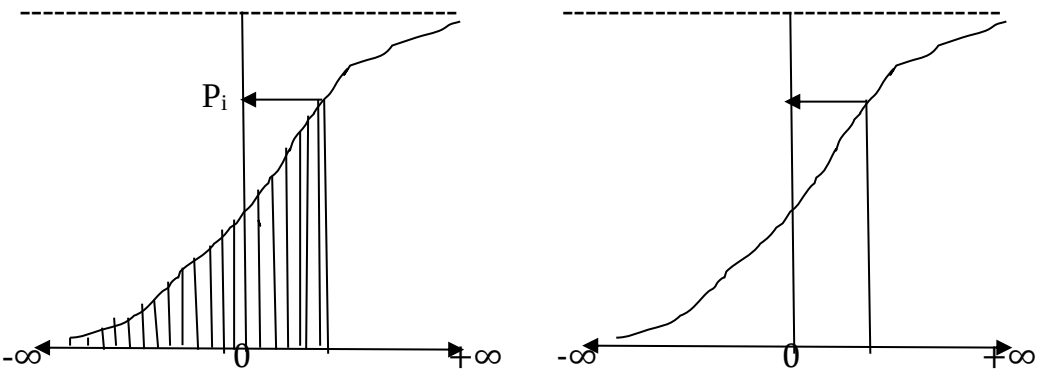
Now to obtain the information on I_i, the utility index, as well as on b_1 and b_2, we take the inverse of (2.35):

$I_i = F^{-1}(I_i) = F^{-1}(P_i)$

$= b_1 + b_2X_i$……(2.3.17)

Where F^{-1} is the inverse of normal CDF

$P_i = F(I_i)$ $P_i = F(I_i)$



(a) $I_i = b_1 + b_2X_i$ (b) $I_i = F^{-1}(P_i)$

SELF ASSESSMENT EXERCISE

(1) What are the shortcomings of the linear probability model (LPM)?

(2) Discuss logit and probit models.

3.3 Dynamic Econometric Models (Distributed Lag and Autoregressive (Models)

Although many econometric models are formulated in static terms, it is quite possible in time series models to have relationships where the concept of time plays a more crucial role. So, for example, we might find ourselves with a model that has the following form:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + u_t \text{ ……..(2.3.18)}$$

In this model we have Y_t is not depending on the current value of X_t only, but also on past (lagged) values of X_t. there are various reasons why lags might need to be

introduced in a model. Consider, for example, an exogenous shock stimulating the purchase of capital goods. It is unavoidable that some time will elapse from the moment the shock occurred till the firm's knowledge of the situation. This can be either because (a) it requires some time to get the relevant statistical information, (b) it takes time for the firm's managers to draw up plans for the new capital project, or (c) the firm might want to obtain different prices from competing suppliers of capital equipment, among various reasons. Therefore, lagged effects will occur and dynamic models which can capture the effects of the time paths of exogenous variables and/or disturbances on the time path of the endogenous variables are needed (Asteriou & Hall, 2007).

In general there are two types of dynamic models.

(1) Distributed lag models that include lagged terms of the independent (or explanatory variables), and

(2) Autoregressive models that include lagged terms of the dependent variable.

These two types of model are described in this unit.

- The Distributed Lag Models

Consider the model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + u_t$$

$$= \alpha + \sum_{i=0}^{p} \beta_i X_{t-i} + u_t \quad \text{.....(2.3.19)}$$

In which the βs are coefficients of the lagged X terms. With this model the reaction to Y_t after a change in X_t is distributed over a number of time periods. In the model, we have p lagged terms and the current X_t term, so it takes p+1 period for the full effect of a change in X_t to influence Y_t. it is interested to examine the effect of the βs.

(a) The coefficient β_0 is the weight attached to the current X (X_t) given by $\Delta Y_t/\Delta X_t$. it therefore, shows how much the average change in Y_t will be when X_t changes by one unit. β_0 is for this reason called the **impact multiplier**.

(b) β_i is similarly given by $\Delta Y_t/\Delta X_{t-i}$ and shows the average change in Y_t for a unit increase in X_{t-i}, i.e. for a unit increase in X made I periods prior to t. for this reason the β_{is} are called the interim multipliers of order i.

(c) The total effect s given by the sum of the effects on all periods:

$$\sum_{i=0}^{p} \beta_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_p \quad \text{.....(2.3.20)}$$

This is also called the long run equilibrium effect when the economy is at the steady state (equilibrium) level. In the long run:

$$X^* = X_t = X_{t-1} = \ldots = X_{t-p} \quad \text{.....(2.3.21)}$$

And therefore:

$$Y_t^* = \alpha + \beta_0 X^* + \beta_1 X^* + \beta_2 X^* + ... + \beta_p X^* + u_t \quad \text{.....(2.3.22)}$$

Distributed lag models can be estimated by simple OLS and the estimators of the βs are BLUE. The question here is how lags are required in order to have a correctly specified? Or, in other words, what is the optimal lag-length? One way to resolve this is to use a relatively large value for p, estimate the model for p, p-1, p-2,… lags and choose the model with the lowest value of Akaike Information Criteria (AIC), Schwarz Bayesian Criteria (SBC) or any other criterion. However, this approach generates two considerable problems:

(a) it can suffer from severe multicollinearity problems, because of close relationships between X_t, X_{t-1}, X_{t-2},… X_{t-p}; and

(b) a large number of p means a considerable loss of degree of freedom because we can use only the p+1 to n observations.

Therefore, and alternative approach s needed to provide methods that can resolve these difficulties. The typical approach is to impose restrictions regarding the structure of the βs and then reduce from P+1 to only a few the number of parameters to be estimated.

Two of the most popular methods to do this are the Koyck (geometric lag) and the Almon (polynomial lag) transformations which are presented below.

- **The Koyck Transformation**

Koyck (1954) proposed a geometrically declining scheme for the βs. To understand this, consider again the distributed lag model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_p X_{t-p} + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.23)$$

Two assumptions were made by Koyck:

(a) All the βs have the same sign; and

(b) The βs decline geometrically as in the following equation:

$$\beta_i = \beta_0^{\lambda^i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.24)$$

Where λ takes value among 0 and 1 and I = 1, 2, 3,…

It is easy to see that it is declining. Since λ is positive and less than one and all the β_i have the same sign, then $\beta_0^{\lambda^1} > \beta_0^{\lambda^2} > \beta_0^{\lambda^i}$ and son on; and therefore $\beta_1 > \beta_2 > \beta_3$ and so on.

Let us say we have an infinite distributed lag model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.25)$$

Substituting $\beta_i = \beta_0^{\lambda^i}$, we have:

$$Y_t = \alpha + \beta_0^{\lambda^0} X_t + \beta_0^{\lambda^1} X_{t-1} + \beta_0^{\lambda^2} X_{t-2} + ... + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2.3.26)$$

For this infinite lag model the immediate impact is given by β_0 (because $\lambda^0 = 1$), while the long run effect will be the sum of an infinite geometric series. Koyck transforms this model to a much simpler one as follows:

Step 1: lag both sides of equation (2.3.26)

$$Y_{t-1} = \alpha + \beta_0^{\lambda^0} X_{t-1} + \beta_0^{\lambda^1} X_{t-2} + \beta_0^{\lambda^2} X_{t-3} + ... + u_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.27)$$

Step 2: Multiply both sides of equation (2.3.27) one period to get:

$$\lambda Y_{t-1} = \lambda\alpha + \beta_0^{\lambda^1} X_{t-1} + \beta_0^{\lambda^2} X_{t-2} + \beta_0^{\lambda^3} X_{t-3} + ... + \lambda u_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.28)$$

Step 3: Subtract (2.3.28) from (2.3.26)

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + u_t - \lambda u_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.29)$$

Or

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + v_t \ldots\ldots\ldots\ldots\ldots..\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.30)$$

Where $v_t = u_t - u_{t-1}$. In this case the immediate effect is $\beta_0/(1 - \lambda)$ (considering again that in the long run we have $Y^* = Y_t = Y_{t-1} = \ldots$). So equation (2.3.30) is now enough to give us both the immediate and long run coefficients very easily (Ateriou & Hall, 2007).

- **The Almon Transformation**

An alternative procedure is provided by Almon (1965). Almon assumes that the coefficients β_i can be approximated by polynomials in I, such as:

$\beta_0 = f(0) = \alpha_0$

$\beta_1 = f(1) = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$

$\beta_2 = f(2) = \alpha_0 + 2\alpha_1 + 4\alpha_2 + 8\alpha_3$

$\beta_3 = f(3) = \alpha_0 + 3\alpha_1 + 9\alpha_2 + 27\alpha_3$

$\beta_1 = f(4) = \alpha_0 + 4\alpha_1 + 16\alpha_2 + 64\alpha_3$

Substituting these into the distributed lag model of order p = 4, we have:

$$Y_t = \alpha + (\alpha_0) X_t + (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3) X_{t-1}$$

$$+ (\alpha_0 + 2\alpha_1 + 4\alpha_2 + 8\alpha_3) X_{t-2}$$

$$+(\alpha_0 + 3\alpha_1 + 9\alpha_2 + 27\alpha_3)X_{t-3}$$

$$+(\alpha_0 + 4\alpha_1 + 16\alpha_2 + 64\alpha_3)X_{t-4} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.31)$$

And factorizing the α_is, we get:

$$Y_t = \alpha + \alpha_0(X_t + 4X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}$$

$$+\alpha_1(X_{t-1} + 2X_{t-2} + 3X_{t-3} + 4X_{t-4})$$

$$+\alpha_2(X_{t-1} + 4X_{t-2} + 9X_{t-3} + 16X_{t-4})$$

$$+\alpha_3(X_{t-1} + 8X_{t-2} + 27X_{t-3} + 64X_{t-4}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3.32)$$

Therefore, what is required is to apply appropriate transformation of the Xs such as the ones given in parentheses. If α_3 is not significant, then a second-degree polynomial might be preferable. If we want to include additional terms, we can easily do that as well. The best model will be either the one that maximizes R^2 (for different model combinations regarding r and p), or the one that minimizes AIC, SBC or any other criteria.

- **Other Models of lag Structure**

There are several other models for reducing the number of parameters in a distributed lag model. Some of the most important ones are the Pascal lag, the gamma lag, the LaGuerre lag and the Shiller lag (Kmenta, 1986).

- **Autoregressive Models**

Autoregressive models are models that simply include lagged dependent variables as regressors. In the Koyck transformation discussed in the previous section, we saw that Yt-1 appears as a regressor, so it can be considered as a case of a distributed lag model that was transformed to an autoregressive model. There are two more specifications involving lag-dependent variables:

(a) The partial adjustment model; and

(b) The adaptive expectations model.

We will examine these models below.

- **The Partial Adjustment Model**

Suppose that the adjustment of the actual value of a variable Y_t to its optimal (or desired) level (denoted by Y_t^*) needs to be modelled. One way to do this is through the partial adjustment model which assumes that the change in actual Y_t ($Y_t - Y_{t-1}$) will be equal to a proportion of the optimal change ($Y_t^* - Y_{t-1}$) or:

$$Y_t - Y_{t-1} = \lambda(Y_t^* - Y_{t-1}) \quad \text{(2.3.33)}$$

Where λ is the adjustment coefficient, which takes values from 0 t0 1, and $1/\lambda$ denotes the speed of adjustment.

Consider the two extreme cases (a) if $\lambda = 1$, then $Y_t = Y_{t-1}$ which means that there is no adjustment of the Y. Therefore, the closer λ is to unity, the faster the adjustment will be. To understand it better, we can use a model from economic theory. Suppose Y_t^* is the desired level of inventories for a firm I, and that this depends on the level of the sales of the firm X_t.

$$Y_t^* = \beta_1 + \beta_2 X_t \quad \text{(2.3.34)}$$

Because there are 'frictions' in the market, there is bound to be a gap among the actual level of inventories and the desired one. Suppose also that a part of the gap can be closed each period, and then the equation that will determine the actual level of inventories will be given by:

$$Y_t = Y_{t-1} + \lambda(Y_t^* - Y_{t-1}) + u_t \quad \text{(2.3.35)}$$

That is, the actual level of inventories is equal to that at time t-1 plus an adjustment factor and a random component.

Combining (2.3.44) and (2.3.35):

$$Y_t = Y_{t-1} + \lambda(\beta_2 X_t - Y_{t-1}) + u_t$$

$$= \beta_1\lambda + (1 - \lambda)Y_{t-1} + \beta_2\lambda X_t + u_t \text{…………………………………(2.3.36)}$$

From this model we have the following:

(a) The short run reaction of Y to a unit change in X is $\beta_2\lambda$;

(b) The long run reaction is given by β_1; and

(c) An estimate of β_1 can be obtained by dividing the estimate of $\beta_2\lambda$ by one minus the estimate of $(1 - \lambda)$, i.e. $\beta_1 = \beta_2\lambda/[1 - (1 - \lambda)]$.

Here, it is useful to note that error correction model (ECM) s also an adjustment model.

Example:

Consider the money demand function:

$$M_t^* = aY_t^{b1}R_t^{b2}e_t^{ut} \text{…………………………………………………..(2.3.36)}$$

where the usual notation applies. Taking logarithms of this equation we get:

$$\ln M_t^* = \ln a + b_1\ln Y_t + b_2\ln R_t + u_t \text{…………………………………(2.3.57)}$$

The partial adjustment hypothesis can be written as:

$$\frac{M_t}{M_{t-1}} = \left(\frac{M_t^*}{M_{t-1}}\right) \text{……………………………………………………..(2.3.38)}$$

Where if taking logarithms, gives:

$$\ln M_t - \ln M_{t-1} = \lambda((\ln M_t^* - \ln M_{t-1}) \text{………………………………(2.3.39)}$$

Substituting (2.3.37) into (2.3.39), we get:

$$\ln M_t - \ln M_{t-1} = \lambda(\ln a + b_1\ln Y_t + b_2\ln R_t + u_t - \ln M_{t-1})$$

$$\ln M_{t-1} = \lambda \ln a + \lambda b_1\ln Y_t + \lambda b_2\ln R_t + (1 - \lambda)\ln M_{t-1} + \lambda u_t \text{ …(2.3.40)}$$

or

$$\ln M_{t-1} = \gamma_1 + \gamma_2\ln Y_t + \gamma_3\ln R_t + \gamma_4\ln M_{t-1} + v_t \text{……………………(2.3.41)}$$

- **The Adaptive Expectations Model**

The second of the autoregressive models is the adaptive model, which is based on the adaptive expectations hypothesis formulated by Cagan (1956). Before understanding the model, it is crucial to have a clear picture of the adaptive expectations hypothesis. So, consider an agent who forms expectations of a variable X_t. If we denote by the subscript e expectations, then $X^e{}_{t-1}$ is the expectation formed at time t-1 for X in t.

The adaptive expectations hypothesis assumes that agents make errors in their expectations (given by X_t - X^e_{t-1}) and also that they revise their expectations by a constant proportion of the most recent error. Thus:

$$X^e_t - X^e_{t-1} = \theta(X_t - X^e_{t-1}) \quad 0 < \theta \leq 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2.3.42)$$

Where θ is the adjustment parameter

If we consider again the two extreme cases, we have that:

(a) If $\theta = 0$, then $X^e_t = X^e_{t-1}$ and no revision in the expectations is made; while

(b) If θ 1, the $X^e_t = X_t$ and we have an instantaneous in the expectations.

(c) The adaptive expectations hypothesis can now be incorporated in an econometric model. Suppose that we have the following model:

$$Y_t = \beta_1 + \beta_2 X^e_t + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.43)$$

Where, for example, we can think of Y_t as consumption of X^e_t as expected income. Assume, then, that for the specific model the expected income follows the adaptive expectations hypothesis, so that:

$$X^e_t - X^e_{t-1} = \theta(X_t - X^e_{t-1}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3.44)$$

If actual X in period t-1 exceeds the expectations, then we would expect agents to revise their expectations upwards. Equation (2.3.44) then becomes:

$$X^e_t = \theta X_t + (1 - \theta) X^e_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.3.45)$$

Substituting (2.3.45) into (2.3.43), we obtain:

$$Y_t = \beta_1 + \beta_2(\theta X_t + (1 - \theta) X^e_{t-1}) + u_t$$

$$= \beta_1 + \beta_2\theta X_t + \beta_2(1 - \theta) X^e_{t-1}) + u_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3.46)$$

In order to estimate the X^e_{t-1} variable from equation (2.3.46) to obtain an estimable econometric model, we need to follow the following procedure:

Lagging equation (2.3.63) one period, we get:

$$Y_{t-1} = \beta_1 + \beta_2 X^e_{t-1} + u_{t-1}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3.47)$$

Multiply both sides of (2.67) by (1-θ) we get:

$$(1-\theta)Y_{t-1} = (1-\theta)\beta_1 + (1-\theta)\beta_2X^e_{t-1} + (1-\theta)u_{t-1}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3.48)$$

Subtract (2.68) from (2.63) we get:

$$Y_t - (1-\theta)Y_{t-1} = \beta_1 - (1-\theta)\beta_1 + \beta_2 X^e_t - (1-\theta)\beta_2X^e_{t-1} + u_t - (1-\theta)u_{t-1}\ldots\ldots.(2.3.49)$$

Or

$$Y_t = \beta_1\theta + \beta_2\theta X_t + (1 - \theta)Y_{t-1} + u_t - (1-\theta)u_{t-1}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3.50)$$

And finally:

$$Y_t = \beta_1^* + \beta_2^* X_t + \beta_3^*Y_{t-1} + v_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(2.3.51)$$

Where $\beta_1^* = \beta_1\theta$, $\beta_2^* = \beta_2\theta$, $\beta_3^* = (1 - \theta)$ and $v_t = u_t - (1-\theta)u_{t-1}$. Once estimates of the β^*s have been obtained, β_1, β_2, and θ can be estimated as follows:

$$\theta = 1 - \beta_3^*, \beta_1 = \beta_1^*/\theta \text{ and } \beta_2 = \beta_2^*/\theta$$

here, it is interesting to mention that through this procedure we are able to obtain an estimate of the marginal propensity to consume out of expected income, although we do not have data for the expected income.

Note: it is very highly important to test for autocorrelation n models with lagged dependent variables. In such cases, the Durbin-Watson (DW) tetst statistic is not

appropriate and Durbin's h-test should be applied instead, or alternatively the Lagrangian Multiplier (LM) test for autocorrelation. The Durbin h-test is given as:

$$h = \hat{\rho}\sqrt{\frac{n}{1-n[\text{var}(\hat{\alpha})]}} \quad \text{.....(2.3.52)}$$

Where n is the sample size, $\text{var}(\hat{\alpha})$ is the variance of the lagged Y_t ($=Y_{t-1}$) coefficient, and $\hat{\rho}$ is an estimate of the first-order serial correlation ρ.

SELF ASSESSMENT EXERCISE

Show how we might obtain an estimate of the marginal propensity to consume out of expected income, although we do not have data for expected income, using the adaptive expectations autoregressive model.

4.0 CONCLUSION

Qualitative response regression models refer to models in which the response, or regressand, variable is not quantitative or an interval scale. The simplest possible qualitative response regression model is the binary model in which the regressand is of the yes/no or presence /absence type. Also the simplest possible binary regression model is the linear probability model (LPM) in which the binary regression model is regressed on the relevant explanatory variables by using the standard OLS methodology. However, the LPM suffers from several estimation problems. Even if some of the estimations problem can be overcome, the major shortcoming of the LPM is that it assumes that the probability of something happening increases linearly with the level of the regressor. This very restrictive assumption can be overcome with the use of logit and probit models. Furthermore, this unit elaborated on regression models that take into account time lags known as dynamic or lagged regression models (distributed lag and autoregressive).

5.0 SUMMARY

So far unit discussed qualitative response regression models in which the explanatory variable(s) and the response variable takes on qualitative variables. The unit further

examined the simple possible binary regression model known as the linear probability model (LPM) which has the shortcoming of estimation problems (assuming linearity between the probability and the regressor. To overcome this shortcoming, the unit introduced the logit and probit models. The unit further elaborated on on regression models that take into account time lags known as dynamic or lagged regression models (distributed lag and autoregressive). A purely distributed lag model can be estimated by OLS, but in that case there is the problem of multicollinearity since successive lagged values of a regressor tend to be correlated. As a result, some short cut methods have been deviced. These include the Koyck, the adaptive expectations, and partial adjustment mechanisms, the first being purely algebraic approach and the other two being based on economic principles.

6.0 TUTOR-MARKED ASSIGNMENT

(1) Consider the following model for home ownership, the maximum likely hood estimates of the logit model are as follows:

$$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = -493.54 + 32.96income$$

t = (-0.000008) (0.000008)

Comment on these results, bearing in mind that all values of income below 16 correspond to Y = 0. A priori, what would you expect in such a situation.

(2) Discuss the problem one encounters when using OLS to estimate the parameters of linear probability model.

7.0 REFERENCES/FURTHER READINGS

Almon, S. (1965). The Distributed Lag between Capital Appropriations and Net Expenditure, Econometrica, 33, 178 – 196.

Asteriou, D. & Hall, S. (2011). Applied Econometrics: A Modern Approach (Revised Edition), New York: Palgrave Macmillam.

Cagan, P. (1956). The Monetary Dynamics of Hypernflation. In Friedman, Milton (ed.). Studies in Quantity Theory of Money, Chiicago: University of Chicago Press.

Gujarati, D.N. (2006). Essentials of Econometrics (Third Edition). McGraw-Hill, New York.

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

Kmenta, J. (1986). Elements of Econometrics. New York: Macmillan.

Koutsoyannis, A. (1977). Theory of Econometrics (Second Edition). PALGRAVE. New York.

Koyck, L.M. (1954). Distributed Lag and Investment Analysis.. North-Holland, Amsterdam.

Kramer, J.S. (1991). The Logit Models for Economists, Edward Arnold Publishers, London.

UNIT FOUR: SIMULTANEOUS EQUATION ESTIMATION

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 Nature of Simultaneous Equation
    3.2 Consequences of Ignoring Simultaneity **(**Simultaneous Bas: Inconsistency of OLS Estimators)
    3.3 Recursive Models and the OLS
    3.4 Approaches to Estimation
    3.5 Estimation of exactly Identified and Over-identified Equations
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

In this unit, we discuss the simultaneous-equation models. In particular we discuss special features, their estimation, and some of the statistical problems associated with them.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- State the nature of simultaneous-equation model.
- Identify simultaneous-equation bias in a model and the inconsistency of the OLS estimators.
- Describe approaches to simultaneous-equation estimators.
- Examine recursive models and the OLS
- Analyse estimation of exactly identified and over-identified equations.

3.0 MAIN CONTENT

3.1 Nature of Simultaneous Equation Model

All econometric models covered so far have dealt with a single dependent variable and estimations of a single equation. However, in modern economics, interdependence is very commonly encountered. Several dependent variables are determined simultaneously, therefore appearing both as dependent and independent variables in a set of different equations. For example, in the single equation case that we have examined so far, we had equations as demand functions of the following form:

$$Q_t^d = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4.1)$$

Where Q_t^d is quantity demanded, P_t is the price of the commodity, and Y_t s income. However, economic analysis suggest that price and quantity are typically determined simultaneously by the market processes, and therefore a full market model is not captured by a single equation but consists of a set of three different equations: the demand function, the supply function, and the condition for equilibrium in the market of a product. So, we have:

$$Q_t^d = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_{1t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4.2)$$

$$Q_t^s = \gamma_1 + \gamma_2 P_t + u_{2t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4.3)$$

$$Q_t^d = Q_t^s \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4.4)$$

Where of course Q_t^s denotes the quantity supplied.

Equations (2.4.2), (2.4.3) and (2.4.4) are called structural equations of the simultaneous equations model, and the coefficients β and γ are called structural parameters.

Because price and quantity are jointly determined, they are both endogenous variables, and because income is not determined by specified model, income is characterized as an exogenous variable. Note, here, that in the single-equation models, we were using the terms exogenous variable and explanatory variable interchangeably, this is no longer

possible in simultaneous equation models. So, we have price as an explanatory variable but not as an exogenous variable as well.

Equating (2.4.2) to (2.4.3) and solving for P_t we get:

$$P_t = \frac{\beta_1 - \gamma_1}{\beta_2 - \gamma_2} + \frac{\beta_3}{\beta_2 - \gamma_2} + \frac{u_{1t} - u_{2t}}{\beta_2 - \gamma_2} \quad \text{.....(2.4.5)}$$

This can be written as:

$$P_t = \pi_1 + \pi_2 Y_t + \upsilon_{1t} \quad \text{.....(2.4.6)}$$

If we substitute (2.4.6) into 2.4.3) we get:

$$Q = \gamma_1 + \gamma_2(\pi_1 + \pi_2 Y_t + \upsilon_{1t}) + u_2 t$$

$$= \gamma_1 + \gamma_2\pi_1 + \gamma_2\pi_2 Y_t + \gamma_2\upsilon_{1t} + u_2 t$$

$$= \pi_3 + \pi_4 Y_t + \upsilon_{2t} \quad \text{.....(2.4.7)}$$

So now we have those equations (2.4.6) and (2.4.7) specify each of the endogenous variables in terms only of the exogenous variables, the parameter of the model and the stochastic error terms. These two equations are known as ***reduced form equations*** and the πs are known as ***reduced form parameters***. In general reduced form equations can be obtained by solving for each of the endogenous variables in terms of the exogenous variables, the unknown parameters and the error terms.

SELF ASSESSMENT EXERCISE

Interdependence is common in modern economics; therefore, many economic relationships cannot be captured with a single equation. Discuss.

3.2 Consequences of Ignoring Simultaneity (Simultaneous Bias: The Inconsistency of the OLS Estimators)

One of the assumptions of classical linear regression model (CLRM) states that error term of an equation should be uncorrelated with each explanatory variable in this equation. If such a correlation exists, then the OLS regression equation is biased. It should evident from the reduced form equations that in cases of simultaneous-equation models such a bias exists. Recall that the new error terms v_{1t} and v_{2t} depend on u_{1t} and u_{2t}. However, to show this more clearly, consider the following general form of a simultaneous-equation model:

$$Y_{1t} = \alpha_1 + \alpha_2 Y_{2t} + \alpha_3 X_{1t} + \alpha_4 X_{3t} + e_{1t} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.4.8)}$$

$$Y_{2t} = \beta_1 + \beta_2 Y_{1t} + \beta_3 X_{3t} + \beta_4 X_{2t} + e_{2t} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.4.9)}$$

In this model, we have two structural equations, with two endogenous variables (Y_{1t} and Y_{2t}) and three exogenous variables (X_{1t}, X_{2t} and X_{3t}). let us see what happens if one of the error terms increases, assuming everything else in the equations to be held constant:

(a) If e_{1t} increases, this cause Y_{1t} to increase due to equation (2.3.80), then

(b) If Y_{1t} increases (assuming that β_2 is positive) Y_{2t} will then also increase due to the relationship in equation (2.4.9), but

(c) If Y_{2t} increases in (2.4.9) it also increases (2.4.8) where it is an explanatory variable.

Therefore, an increase in the error term of one of the equation causes an increase in an explanatory variable in the same equation. So the assumption of no correlation among the error term and the explanatory variable is violated leading to biased estimates.

SELF ASSESSMENT EXERCISE

Describe how the OLS becomes inconsistent in the face of simultaneous bias.

3.3 Approaches to Estimation

If we consider the general M equations model in M endogenous variables, we may adopt two approaches to estimate the structural equations, namely, single-equation methods, also known as **limited information methods**, and system methods, also known as **full information methods**. In the single-equation methods, we estimate each equation in the system (of simultaneous equations) individually, taking into consideration any restrictions imposed on the equation (such as exclusion of some variables) without worrying about the restrictions on the other equations in the system, hence the name limited information methods. In the system methods, on the other hand, we estimate all the equations in the model simultaneously, taking into due consideration of all restrictions on such equations by the omission or absence of some variables, hence the name full information methods (Christ, 1966).

As an illustration, let us consider the following system of simultaneous-equation:

$$
\begin{aligned}
Y_{1t} &= \beta_{10} + \quad + \beta_{12}Y_{2t} + \beta_{13}Y_{3t} + \quad + \gamma_{11}X_{1t} + \qquad\qquad + u_{1t} \\
Y_{2t} &= \beta_{20} + \qquad\quad + \beta_{23}Y_{3t} + \quad + \gamma_{21}X_{1t} + \gamma_{22}X_{2t} \qquad + u_{2t} \\
Y_{3t} &= \beta_{30} + \beta_{31}Y_{1t} + \qquad\quad + \beta_{34}Y_{4t} + \gamma_{31}X_{1t} + \gamma_{32}X_{2t} + \quad + u_{3t} \\
Y_{4t} &= \beta_{40} + \quad + \beta_{42}Y_{2t} \qquad\qquad\qquad\qquad + \gamma_{43}X_{3t} + \quad + u_{1t\ldots}(2.4.10)
\end{aligned}
$$

Here, the Y's are the endogenous variables whereas the X's are the exogenous variables. Supposing our interest is to estimate the third equation in the system noting that the variables Y_2 and X_3 are excluded from it. In the system method, on the other hand, emphasis is placed on estimating all the four equations in the system simultaneously, considering all the restrictions placed on the various equations of the system.

To maintain the spirit of simultaneous-equation models, one basically should apply the systems method, such as the full information maximum likelihood (FIML) method. Despite the advantageous nature of the system methods, in practice, such methods are not commonly used for the following reasons: (a) the computational burden is enormous and stupendously tasking even in these days of high-speed computers, not to mention the cost involved. (b) the systems methods, such as FIML, lead to solutions that are

highly nonlinear in the parameters and are therefore often difficult to determine. (c) if there is specification error in one or more equations of the system, that error is transmitted to the rest of the system as a result, the systems methods become sensitive to specification errors (Klein, 1974).

SELF ASSESSMENT EXERCISE

Describe the two approaches to estimate the structural equations, namely, single-equation methods, also known as limited information methods, and system methods, also known as full information methods.

3.4 Recursive Models and OLS

We learned from our previous unit that due to the interdependence between the stochastic disturbance term and the endogenous explanatory variable(s), the OLS method is inappropriate for the estimation of an equation in a system of simultaneous-equations. However, if OLS is applied mistakenly to such an equation, the estimators are not only biased (in small samples) but also inconsistent; that is, the bias does not disappear no matter how large the sample size. There is however an exceptional situation where the OLS can be applied appropriately even in the context of simultaneous equations. This is a typical case of the recursive, triangular, or causal models. As an example, consider the following system of equations:

$$Y_{1t} = \beta_{10} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + u_{1t}$$

$$Y_{2t} = \beta_{20} + \beta_{21}Y_{1t} + \gamma_{21}X_{1t} + \gamma_{22}X_{2t} + u_{2t} \quad \text{.....(2.4.11)}$$

$$Y_{3t} = \beta_{30} + \beta_{31}Y_{1t} + \beta_{32}Y_{2t} + \gamma_{31}X_{1t} + \gamma_{32}X_{2t} + u_{3t}$$

Where the Y's are the endogenous while the X's are exogenous as usual. The disturbances are assumed as follows:

$$\text{cov}(u_{1t}, u_{2t})\ \text{cov}(u_{1t}, u_{3t}) = \text{cov}(u_{2t}, u_{3t}) = 0$$

that is, the same-period disturbances in different equations are uncorrelated (the assumption of zero contemporaneous correlation).

Let us consider the first equation of (2.4.11). Since it contains only the exogenous variable on the right hand side and since by assumption they are not correlated with the disturbance term, u_{1t}, this equation satisfies the critical assumption of the classical OLS (uncorrelatedness between the explanatory variable and the stochastic disturbances). Hence, OLS can be applied to this equation. Similarly, consider the second equation of (2.4.11), which contains the endogenous variable Y_1 as an explanatory variable along with non-stochastic X's. In this case, OLS can also be applied provided Y_{1t} and u_{2t} are uncorrelated. The reason for the un-correlation of Y_{1t} and u_{2t} is that u_1, which affect Y_1, is by assumption uncorrelated with u_2. Therefore, Y_1 is a predetermined variable in as so far Y_2 is concerned. The third equation can be estimated using OLS because both Y_1 and Y_2 are uncorrelated with u_3.

Thus, in a recursive model, OLS can be applied to each equation separately. Actually, there is no a simultaneous equation problem in this situation. From the structure of such systems, it is clear there is no interdependence among the endogenous variables. Thus, Y_1 affects Y_2, but Y_2 does not affect Y_1. In the same vein, Y_1 and Y_2 affect Y_3 without, in turn, being influenced by Y_3. In other words, each equation exhibits a unilateral causal dependence, hence, the name causal models (Zellner, 1962).

3.5 Estimation of Simultaneous Equation Models

The question of identification is closely related to the problem of estimating the structural parameters in a simultaneous equation model. Thus, when an equation is not identified, such estimation is not possible. N cases, though, of exact or overidentification, there are procedures that allow us to obtain estimates of the structural parameters. These procedures are different from simple OLS in order to avoid the simultaneity bias we discussed previously.

In general, in cases of exact identification, the appropriate method is the so-called method of indirect least squares (ILS), while in cases of overidentified equations, the two stage least squares (2SLS) method is the most commonly used.

3.5.1 Estimation of an Exactly Identified Equation: The Method of ILS

This method can be applied only when the equations of a simultaneous equation model is found to be exactly identified. The procedure of the ILS involves the following three steps:

Step 1: Find the reduced form equations. These reduced-form-equations are obtained from the structural equations in such a manner that the dependent variable in each equation is the only endogenous variable and is a function solely of the predetermined (exogenous or lagged endogenous) variables and the stochastic error term(s).

Step 2: Estimate the reduced form parameters by applying simple OLS to the reduced-form-equations, and

Step 3: Obtain unique estimates of the structural parameters from the estimates of the parameters of the reduced-form-equation in step 2.

As this three-step procedure indicates, the name ILS derives from the fact that structural coefficients are obtained indirectly from the OLS estimates of the reduced-form-coefficients Gujarati & Sangeetha, 2007).

Example:

Consider the demand-and-supply model which is given below:

Demand function: $Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 X_t + u_{1t}$ .....(2.4.12)

Supply function: $Q_t = \beta_0 + \beta_1 P_t + u_{2t}$ .....(2.4.13)

Where Q = quantity, P = Price; X = income

Now we assume that X is exogenous. As observed previously, the supply function is exactly identified whereas the demand function is not identified.

The reduced-form-equations corresponding to the preceding structural equations are:

$$P_t = \psi_0 + \psi_1 X_t + w_t \quad \text{.....(24.14)}$$

$$Q_t = \psi_2 + \psi_3 X_t + v_t \quad \text{.....(2.4.15)}$$

Where ψ's are the reduced-form coefficients and are (nonlinear) combinations of the structural coefficients, and where ω and v are linear combinations of the structural disturbances u_1 and u_2.

One plausible thing is that, each reduced-form equation contains only one endogenous variable and which is a function of the exogenous variable X (income) and the stochastic disturbances. Hence, the parameters of the preceding reduced-form equations may be estimated by OLS as follows:

$$\begin{aligned} \hat{\psi}_1 &= \frac{\sum p_t x_t}{\sum x_t^2}; \hat{\psi}_0 = \bar{P} - \hat{\pi}_1 \bar{X} \\ \hat{\psi}_3 &= \frac{\sum q_t x_t}{\sum x_t^2}; \hat{\psi}_2 = \bar{Q} - \hat{\pi}_1 \bar{X} \end{aligned} \quad \text{.....(24.16)}$$

Where the lower case letters denote deviations from the sample means and where $\bar{Q}$ and $\bar{P}$ are the sample mean values of Q and P. As observed previously, $\hat{\pi}_i$'s are consistent estimators and under appropriate assumptions is minimum variance unbiased or asymptotically efficient.

Our major concern now is to determine the structural coefficients from the reduced-form coefficients. The supply function is exactly identified; therefore its parameters can be estimated uniquely from the reduced-form coefficients as follows:

$$\beta_0 = \psi_2 - \beta_1 \psi_0 \text{ and } \beta_1 = \frac{\psi_3}{\psi_1} \quad \text{.....(2.4.17)}$$

Hence, the estimates of these parameters can be obtained from the estimates of the reduced-form coefficients as follows:

$$\hat{\beta}_0 = \hat{\psi}_2 - \hat{\beta}_1 \hat{\psi}_0 \text{ and } \hat{\beta}_1 = \frac{\hat{\psi}_3}{\hat{\psi}_1} \quad \text{.....(2.4.18)}$$

Which are the ILS estimators.

The ILS method is not commonly used for two reasons:

(1) Most simultaneous equations models tend to be over-identified.

(2) If the system has several equations, solving for the reduced form and then for the structural form can be very tedious. An alternative is the TSLS method.

3.5.2 Estimation of an Over Identified Equation: The Method of 2SLS

The basic idea behind the 2SLS method is to replace the stochastic endogenous variable (which is correlated with the error term and causes the bias) with the one that is non-stochastic and consequently independent of the error term. This involves two stages (hence two-stage least squares):

Stage 1: Regress each endogenous variable which is a regressor as well, on all of the exogenous and lagged endogenous variables in the entire system by using OLS (that is equivalent to estimating the reduced form equations) and obtain the fitted values of the endogenous variables of these regressions ($\widehat{Y}$).

Stage 2: Use the fitted values from stage 1 as proxies or instruments for the endogenous regressors in the original (structural form) equations.

Example:

Consider the following model:

Income function: $Y_{1t} = \beta_{10} + \beta_{11}Y_{2t} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + u_{1t}$…(2.4.19)

Money supply function $Y_{2t} = \beta_{20} + \beta_{21}Y_{1t} + u_{2t}$…(2.4.20)

Where Y_1 = income, Y_2 = stock of money; X_1 = investment expenditure;

X_2 = government expenditure on goods and services

The variables X_1 and X_2 are exogenous.

Equation (2.3.91) is the hybrid of quantity-theory-Keynesian approaches to the determination of income while (2.3.92) postulates that the stock of money is determined by the apex bank (CBN) on the basis of the level of income. It is pertinent that there is a simultaneous problem in the two equations.

If we apply the order condition of identification, we will realise that the income equation (2.3.91) is under identified whereas the money supply equation (2.3.92) is over identified. Here, there is nothing that can be done with the income equation rather than changing the specification form of the model. In addition, the over identified money supply function may not be estimated by ILS due to the presence of only two estimates of β_{21}.

However, if we apply OLS to the over identified money supply function; we may get estimates that are inconsistent given that there may be presence of correlation between the stochastic explanatory variable Y_1 and the stochastic disturbance term u2. Suppose, we find a proxy for the stochastic explanatory variable Y_1 such that, it s highly correlated with Y_1 but uncorrelated with u_2. Such proxy is also known as an **instrumental variable**. If such a variable is found, then one can use OLS to estimate the money supply function. To obtain such an instrumental variable, one has to use the two stage least squares (Theil, 1953 & Basmann, 1957). As the name implies, the method involves two successive applications of OLS. The process is as follows:

Stage 1: Regress first Y_1 on all the predetermined variables in the whole system, not just that equation in order to avoid the correlation between Y_1 and u_2. In this, we regress Y_1 on X_1 and X_2 as follows:

$$Y_{1t} = \hat{\psi}_0 + \psi_1 X_{1t} + \psi_2 X_{2t} + \hat{u}_t \quad \text{.....(2.4.21)}$$

Where $\hat{u}_t$ are the usual OLS residuals. From (24.21), we obtain:

$$\hat{Y}_{1t} = \hat{\psi}_0 + \psi_1 X_{1t} + \psi_2 X_{2t} \quad \text{.....(2.4.22)}$$

Where $\widehat{Y}_{1t}$ is an estimate of the mean value of Y conditional upon the fixed X's. Note that (2.4.21) is nothing but a reduced-form regression because only the exogenous variables appear on the right-hand-side. Equation (2.4.21) can be expressed as:

$$Y_{it} = \widehat{Y}_{1t} + \widehat{u}_t \quad \text{(2.4.23)}$$

Which shows that the stochastic Y_1 consist of two parts: $\widehat{Y}_{1t}$ which is a linear combination of the non-stochastic X's, and the random component $\widehat{u}_t$. Following the OLS theory, $\widehat{Y}_{1t}$ and $\widehat{u}_t$ are uncorrelated.

Stage 2: The over identified money supply equation can now be written as:

$$\widehat{Y}_{2t} = \beta_{20} + \beta_{21}(\widehat{Y}_{1t} + \widehat{u}_t) + u_{2t}$$

$$= \beta_{20} + \beta_{21}\widehat{Y}_{1t} + (u_{2t} + \beta_{21}\widehat{u}_t) \quad \text{(2.4.24)}$$

$$= \beta_{20} + \beta_{21}\widehat{Y}_{1t} + u^*_t$$

Where $u^*_t = u_{2t} + \beta_{21}\widehat{u}_t$

When we compare (2.4.24.) and (2.4.20), we discover that they are very similar; the only difference is that Y_1 is replaced by $\widehat{Y}_1$. It can be shown that although Y_1 in the original money supply equation is correlated with the disturbance term u_2 and rendering OLS inappropriate, Y_{1t} in (2.4.24) is uncorrelated with u_t^* asymptotically. As a result, OLS can be used to (2.4.24), which will give consistent estimates of the parameters of the money supply function.

The basic idea behind the 2SLS is to purify the stochastic explanatory variable Y_1 of the influence of influence of the stochastic disturbance, u_2. This objective is achieved by performing the reduced-form regression of Y_1 on all the exogenous variables in the system (i.e. stage 1), and obtain the estimates $\widehat{Y}_{1t}$ and replacing Y_{1t} in the original equation by estimated $\widehat{Y}_{1t}$, and then applying OLS to the equation thus transformed

(stage 2). The estimators thus obtained are consistent; that is they converge to their true values as the sample size increases indefinitely.

SELF ASSESSMENT EXERCISE

Explain the stages of estimating an exact and an over identified equations in a system of simultaneous equation model.

4.0 CONCLUSION

In contrast to single-equation models, in simultaneous-equation models, more than one dependent variable is involved, necessitating as many equations as the number of endogenous variables. A unique feature of simultaneous-equation models is that the endogenous variable in one equation may appear as an explanatory variable in another equation of the system. As a consequence, such endogenous explanatory variable becomes stochastic and is usually correlated with the disturbance term of the equation in which it appears as an explanatory variable. This makes the classical OLS estimators inconsistent. However, alternative methods have been developed to take care of the simultaneous bias of such equations in a system. Some of these methods include: the ILS for exact identification of an equation in a system and the 2SLS for overidentified equation.

5.0 SUMMARY

This unit discussed simultaneous-equation models as models where more than one endogenous variable is involved, which makes the models to have as many equations as the number of dependent variables. The unit further explained how OLS estimators become inconsistent in the presence of simultaneous bias of a system of equations. Although OLS is, in general, inappropriate in the context of simultaneous-equation models, it can be applied to the so-called recursive models where there is definite but unidirectional cause-and-effect relationship among the endogenous variables. Finally, the method of ILS is suited for exact identified equation whereas the 2SLS is designed for an over identified equation.

6.0 TUTOR-MARKED ASSIGNMENT

(1) J. Riti developed the following model for Nigerian economy:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha 2_{It} + u_{1t}$$

$$I_t = \alpha_3 + \alpha_4 Y_t + \alpha_5 Q_t + u_{2t}$$

$$C_t = \alpha_6 + \alpha_7 Y_t + \alpha_8 C_{t-1} + \alpha_9 P_t + u_{3t}$$

$$Q_t = \alpha_{10} + \alpha_1 1 Q_{t-1} + \alpha_1 2 R_t + u_{4t}$$

Where Y = national income; I = investment; C = consumption; Q = profits;

P = cost of living index; R = industrial productivity, t = time;

u = stochastic disturbances

(a) Which of the variables would you regard as endogenous and which as exogenous?

(b) Is there any equation in the system that can be estimated by a single-equation least squares method?

(2) Why is it unnecessary to apply the 2SLS method to an exact identified equation?

(3) Consider the following modified Keynesian model of income determination:

$$C_t = \alpha_{10} + \alpha_{11} Y_t + u_{1t}$$

$$I_t = \alpha_{20} + \alpha_{21} Y_t + \alpha_{22} Y_{t-1} + u_{2t}$$

$$Y_t = C_t + I_t + G_t$$

Where C = consumption; I = investment expenditure; Y = income;

G = government expenditure; G_t and Y_{t-1} are assumed predetermined

7.0 REFERENCES/FURTHER READINGS

Basmann, R.L. (1957). A generalised Classical Method of Linear Estimation of Coefficients in a structural Equation, Econometrica, vol. 25, pp. 77-83.

Christ, C.F. (1966). Econometric Models and methods, John Wiley & Sons, New York, pp. 395-401.

Gujarati, D.N. & Sangeetha (2007). Basic Econometrics. The MacGraw-Hill, New Dehi, India.

Klein, L.R. (1974). A Textbook of Econometrics, 2nd ed., Prentice Hall, Eaglewood Cliffs, N.J., p. 150.

Theil, H. (1953). Repeated least Squares Applied to Complete Equation Systems, The haque: The Central Planning Bureau, The Netherlands.

Zellner, A. (1962). "An Efficient Method of Estimating Seemingly Uncorrelated regressions and Tests for Aggregation Bias". Journal of the American Statistical Association, vol. 57, pp. 348-368.

MODULE THREE: MATRIX TREATMENT OF REGRESSION ANALYSIS AND TIME SERIES ECONOMETRICS

Unit 1: Matrix Treatment of Multiple Regressions

Unit 2: Vector Autoregressive (VAR) Models and Causality Tests

Unit 3: Non-Stationarity and Unit-Root Tests

Unit 4: Cointegration and Error-Correction Models

UNIT 1: MATRIX TREATMENT OF MULTIPLE REGRESSIONS

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 A matrix formulation of the regression model
    3.2 Leas Squares Estimates in Matrix Notation
    3.3 Further Matrix Results for Multiple Linear Regression
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses. Here, we review basic matrix algebra, as well as learn some of the more important multiple regression formulas in matrix form.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- Analyze matrix formulation of the regression model
- Estimate least squares estimate in matrix notation
- Analyze further matrix result for multiple regression

3.0 MAIN CONTENT

3.1 A Matrix Formulation of the Regression Model

As always, let us start with the simple case first. Consider the following simple linear regression function:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for i = 1,…,n} \ldots\ldots(3.1.1)$$

If we actually let i = 1, ..., n, we see that we obtain n equations:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned} \ldots\ldots(3.1.2)$$

Well, that is a pretty inefficient way of writing it all out! As you can see, there is a pattern that emerges. By taking advantage of this pattern, we can instead formulate the above simple linear regression function in matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \ldots\ldots(3.1.3)$$

That is, instead of writing out the n equations, using matrix notation, our simple linear regression function reduces to a short and simple statement:

$$Y = X\beta + \epsilon \ldots\ldots(3.1.4)$$

Now, what does this statement mean? Well, here is the answer:

- ***X*** is an $n \times 2$ **matrix**.
- ***Y*** is an $n \times 1$ **column vector**, $\boldsymbol{\beta}$ is a 2×1 column vector, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ column vector.
- The matrix ***X*** and vector $\boldsymbol{\beta}$ are multiplied together using the techniques of **matrix multiplication**.
- And, the vector $\boldsymbol{X\beta}$ is added to the vector $\boldsymbol{\varepsilon}$ using the techniques of **matrix addition**.

Now, that might not mean anything to you, if you have never studied matrix algebra — or if you have and you forgot it all! So, let us start with a quick and basic review.

(a) Definition of a Matrix

An $\boldsymbol{r \times c}$ **matrix** is a rectangular array of symbols or numbers arranged in r rows and c columns. A matrix is almost always denoted by a single capital letter in boldface type.

Here are three examples of simple matrices. The matrix ***A*** is a 2×2 **square matrix** containing numbers:

$$A = \begin{bmatrix} 1 & 2 \\ 6 & 3 \end{bmatrix}$$

The matrix ***B*** is a 5 × 3 matrix containing numbers:

$$B = \begin{bmatrix} 1,80, 3.4 \\ 1,92, 3.1 \\ 1,65, 2.5 \\ 1,71, 2.8 \\ 1,40, 1.9 \end{bmatrix}$$

And, the matrix ***X*** is a 6 × 3 matrix containing a column of 1's and two columns of various *x* variables:

$$X = \begin{bmatrix} 1, x_{11}, x_{12} \\ 1, x_{21}, x_{22} \\ 1, x_{31}, x_{32} \\ 1, x_{41}, x_{42} \\ 1, x_{51}, x_{52} \end{bmatrix} \text{.....(3.1.5)}$$

(b) Definition of a Vector and a Scalar

A **column vector** is an *r* × 1 matrix, that is, a matrix with only one column. A vector is almost often denoted by a single lowercase letter in boldface type. The following vector ***q*** is a 3 × 1 column vector containing numbers:

$$q = \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}$$

A **row vector** is an 1 × *c* matrix, that is, a matrix with only one row. The vector ***h*** is a 1 × 4 row vector containing numbers:

h = [21 46 32 90]

A 1 × 1 "matrix" is called a **scalar**, but it is just an ordinary number, such as 29 or σ^2.

(c) Matrix multiplication

Recall that ***Xβ*** that appears in the regression function: Y = Xβ + ϵ

is an example of matrix multiplication. Now, there are some restrictions — you cannot just multiply any two old matrices together. **Two matrices can be multiplied together only if** the number of columns of the first matrix equals the number of rows of the second matrix. Then, when we multiply the two matrices, we get: the number of

rows of the resulting matrix equals the number of rows of the first matrix, and the number of columns of the resulting matrix equals the number of columns of the second matrix.

For example, if ***A*** is a 2 × 3 matrix and ***B*** is a 3 × 5 matrix, then the matrix multiplication ***AB*** is possible. The resulting matrix ***C*** = ***AB*** has 2 rows and 5 columns. That is, ***C*** is a 2 × 5 matrix. Note that the matrix multiplication ***BA*** is not possible.

For another example, if ***X*** is an $n \times (k+1)$ matrix and $\boldsymbol{\beta}$ is a $(k+1) \times 1$ column vector, then the matrix multiplication $\boldsymbol{X\beta}$ is possible. The resulting matrix $\boldsymbol{X\beta}$ has n rows and 1 column. That is, $\boldsymbol{X\beta}$ is an $n \times 1$ column vector.

Okay, now that we know when we can multiply two matrices together, how do we do it? Here's the basic rule for multiplying ***A*** by ***B*** to get ***C = AB***:

The entry in the i^{th} row and j^{th} column of ***C*** is the **inner product** — that is, element-by-element products added together — of the i^{th} row of A with the j^{th} column of B.

For example:

$$C = AB = \begin{bmatrix} 1,9,7 \\ 8,1,2 \end{bmatrix} \begin{bmatrix} 3,2,1,5 \\ 5,4,7,3 \\ 6,9,6,8 \end{bmatrix} = \begin{bmatrix} 90,101,106,88 \\ 41,38,27,59 \end{bmatrix}$$

That is, the entry in the **first row** and **first column** of ***C***, denoted c_{11}, is obtained by:

$$c_{11} = 1(3) + 9\,(5) + 7(6) = 90$$

And, the entry in the **first row** and **second column** of ***C***, denoted c_{12}, is obtained by:

$$c_{12} = 1(2) + 9(4) + 7(9) = 101$$

And, the entry in the **second row** and **third column** of ***C***, denoted c_{23}, is obtained by:

$$c_{23} = 8(1) + 1(7) + 2(6) = 27$$

You might convince yourself that the remaining five elements of ***C*** have been obtained correctly.

(d) Matrix Addition

Recall that $\boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ that appears in the regression function: $Y = X\beta + \epsilon$

is an example of matrix addition. Again, there are some restrictions — you cannot just add any two old matrices together. **Two matrices can be added together only if** they have the same number of rows and columns. Then, to add two matrices, simply add the corresponding elements of the two matrices. That is:

- Add the entry in the first row, first column of the first matrix with the entry in the first row, first column of the second matrix.
- Add the entry in the first row, second column of the first matrix with the entry in the first row, second column of the second matrix.
- And, so on.

For example:

$$C = A + B = \begin{bmatrix} 2,4,-1 \\ 1,8,7 \\ 3,5,6 \end{bmatrix} \begin{bmatrix} 7,5,2 \\ 9,-3,1 \\ 2,1,8 \end{bmatrix} = \begin{bmatrix} 9,9,1 \\ 10,5,8 \\ 5,6,14 \end{bmatrix}$$

That is, the entry in the **first row** and **first column** of ***C***, denoted c_{11}, is obtained by:

$c_{11} = 2 + 7 = 9$

And, the entry in the **first row** and **second column** of ***C***, denoted c_{12}, is obtained by:

$c_{12} = 4 + 5 = 9$

You might convince yourself that the remaining seven elements of ***C*** have been obtained correctly.

SELF ASSESSMENT EXERCISE

Define matrix multiplication and addition. What condition must be fulfilled for two matrices to be multiplied and added?

3.2 Least squares estimates in matrix notation

Here is the punch-line: the $(k+1) \times 1$ vector containing the estimates of the $(k+1)$ parameters of the regression function can be shown to equal:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1} X'Y \text{ ………………………….(3.1.6)}$$

where: $(\boldsymbol{X'X})^{-1}$ is the inverse of the $\boldsymbol{X'X}$ matrix, and $\boldsymbol{X'}$ is the transpose of the $\boldsymbol{X}$ matrix.

As before, that might not mean anything to you, if you have never studied matrix algebra — or if you have and you forgot it all! So, let's go off and review inverses and transposes of matrices.

(a) Definition of the Transpose of a Matrix

The **transpose** of a matrix $\boldsymbol{A}$ is a matrix, denoted $\boldsymbol{A'}$ or $\boldsymbol{A}^{T}$, whose rows are the columns of $\boldsymbol{A}$ and whose columns are the rows of $\boldsymbol{A}$ — all in the same order. For example, the transpose of the 3 × 2 matrix $\boldsymbol{A}$:

$$A = \begin{bmatrix} 1 & 5 \\ 4 & 8 \\ 7 & 9 \end{bmatrix}$$

is the 2 × 3 matrix $\boldsymbol{A}'$:

$$A' = A^T = \begin{bmatrix} 1,4,7 \\ 5,8,9 \end{bmatrix}$$

And, since the $\boldsymbol{X}$ matrix in the simple linear regression setting is:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ ……………………………………………..……..(3.1.7)}$$

The $\boldsymbol{X'X}$ matrix in the simple linear regression setting must be:

$$X'X = \begin{bmatrix} 1,1\cdots 1 \\ x_1, x_2 \cdots x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \text{ ……………………..….(3.1.8)}$$

(b) Definition of the identity matrix

The square $n \times n$ identity matrix, denoted $\boldsymbol{I}_n$, is a matrix with 1's on the diagonal and 0's elsewhere. For example, the 2×2 identity matrix is:

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The identity matrix plays the same role as the number 1 in ordinary arithmetic:

$$\begin{bmatrix} 9 & 7 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 9 & 7 \\ 4 & 6 \end{bmatrix}$$

That is, when you multiply a matrix by the identity, you get the same matrix back.

(c) Definition of the inverse of a matrix

The **inverse** $\boldsymbol{A}^{-1}$ of a square (!!) matrix $\boldsymbol{A}$ is the unique matrix such that:

$$A^{-1}A = I = AA^{-1} \quad \text{.....(3.1.10)}$$

That is, the inverse of $\boldsymbol{A}$ is the matrix $\boldsymbol{A}^{-1}$ that you have to multiply $\boldsymbol{A}$ by in order to obtain the identity matrix $\boldsymbol{I}$. The inverse only exists for square matrices!

Now, finding inverses is a really messy venture. The good news is that we'll always let computers find the inverses for us. In fact, we won't even know that statistical software is finding inverses behind the scenes!

An example

Let us take a look at an example just to convince ourselves that, yes, indeed the least squares estimates are obtained by the following matrix formula:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (X'X)^{-1}X'Y \quad \text{.....(3.1.11)}$$

Let us consider the data in soapsuds.txt, in which the height of suds (y = *suds*) in a standard dishpan was recorded for various amounts of soap (x = *soap*, in grams) (Draper

and Smith, 1998, p. 108). Using statistical software to fit the simple linear regression model to these data, we obtain:

```
Regression Equation

suds = -2.68 + 9.500 soap
```

Now, let us see if we can obtain the same answer using the above matrix formula. We previously showed that:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \quad \text{.....(3.1.12)}$$

We can easily calculate some parts of this formula:

x_i soap	y_i suds	$x_i \times y_i$ so*su	x_i^2 soap^2
4.0	33	132.0	16.00
4.5	42	189.0	20.25
5.0	45	225.0	25.00
5.5	51	280.5	30.25
6.0	53	318.0	36.00
6.5	61	396.5	42.25
7.0	62	434.0	49.00
---	---	-----	-----
38.5	**347**	**1975.0**	**218.75**

That is, the 2 × 2 matrix $\boldsymbol{X'X}$ is:

$$X'X = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

And, the 2 × 1 column vector $\boldsymbol{X'Y}$ is:

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix}$$

So, we have determined $\boldsymbol{X'X}$ and $\boldsymbol{X'Y}$**.** Now, all we need to do is to find the inverse $(\boldsymbol{X'X})^{-1}$. As mentioned before, it is very messy to determine inverses by hand. Letting computer software do the dirty work for us, it can be shown that the inverse of $\boldsymbol{X'X}$ is:

$$(X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

And so, putting all of our work together, we obtain the least squares estimates:

$$(X'X)^{-1}X'Y = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

That is, the estimated intercept is b_0 = -2.67 and the estimated slope is b_1 = 9.51. Our estimates are the same as those reported above **(**within rounding error).

SELF ASSESSMENT EXERCISE

Define the following matrix's terms: transpose identity and inverse matrix

3.3 Further Matrix Results for Multiple Linear Regression

Matrix notation applies to other regression topics, including fitted values, residuals, sums of squares, and inferences about regression parameters. One important matrix that appears in many formulas is the so-called "hat matrix,"

$H = X(X'X)^{-1}X'$, since it puts the hat on Y.

(a) Linear Dependence

There is just one more really critical topic that we should address here, and that is linear dependence. We say that the columns of the matrix ***A***:

$$A = \begin{bmatrix} 1,2,4,1 \\ 2,1,8,6 \\ 3,6,12,3 \end{bmatrix}$$

are **linearly dependent**, since (at least) one of the columns can be written as a linear combination of another, namely the third column is 4 × the first column. If none of the columns can be written as a linear combination of the other columns, then we say the columns are **linearly independent**.

Unfortunately, linear dependence is not always obvious. For example, the columns in the following matrix ***A***:

$$A = \begin{bmatrix} 1,4,1 \\ 2,3,1 \\ 3,2,1 \end{bmatrix}$$

are linearly dependent, because the first column plus the second column equals 5 × the third column.

Now, why should we care about linear dependence? Because the inverse of a square matrix exists only if the columns are linearly independent. Since the vector of regression estimates $\boldsymbol{b}$ depends on $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the parameter estimates b_0, b_1, and so on cannot be uniquely determined if some of the columns of $\boldsymbol{X}$ are linearly dependent! That is, if the columns of your $\boldsymbol{X}$ matrix — that is, two or more of your predictor variables — are linearly dependent (or nearly so), you will run into trouble when trying to estimate the regression equation.

For example, suppose for some strange reason we multiplied the predictor variable *soap* by 2 in the dataset soapsuds.txt. That is, we would have two predictor variables, say *soap1* (which is the original *soap*) and *soap2* (which is 2 × the original soap):

soap1	soap2	suds
4.0	8	33
4.5	9	42
5.0	10	45
5.5	11	51
6.0	12	53
6.5	13	61
7.0	14	62

If we tried to regress y = *suds* on x_1 = *soap1* and x_2 = *soap2*, we see that statistical software spits out trouble:

```
* soap2 is highly correlated with other X variables
* soap2 has been removed from the equation

The regression equation is suds = - 2.68 + 9.50 soap1
```

In short, the first moral of the story is "do not collect your data in such a way that the predictor variables are perfectly correlated." And, the second moral of the story is "if your software package reports an error message concerning high correlation among your predictor variables, then think about linear dependence and how to get rid of it."

3.4 More on the Regression Model in Matrix Form

We will consider the linear regression model in matrix form.

For simple linear regression, meaning one predictor, the model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for } i = 1, 2, 3, \ldots, n \ldots\ldots\ldots\ldots\ldots\ldots (3.1.13)$$

This model includes the assumption that the εi 's are a sample from a population with mean zero and standard deviation σ. In most cases we also assume that this population is normally distributed.

The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \varepsilon_i \quad \text{for } i = 1, 2, 3, \ldots, n \ldots\ldots\ldots (3.1.14)$$

This model includes the assumption about the ε_i 's.

This requires building up our symbols into vectors. Thus:

$$\underset{nx1}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.15)$$

captures the entire dependent variable in a single symbol. The "$n \times 1$" part of the notation is just a shape reminder. These get dropped once the context is clear.

For simple linear regression, we will capture the independent variable through this $n \times 2$ matrix:

$$\underset{nx2}{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1.16)$$

The coefficient vector will be $\underset{2x1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and the noise will be $\underset{nx1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

The simpler linear regression model is written then as: $\underset{nx1}{Y} = \underset{nx2}{X}\ \underset{2x1}{\beta} + \underset{nx1}{\varepsilon}$

The product part, meaning $\underset{nx2}{X}\ \underset{2x1}{\beta}$, is found through the usual rule for matrix multiplication as:

$$\underset{nx2}{X}\ \underset{2x1}{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \beta_0 + \beta_1 x_3 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} \quad \text{.....(3.1.17)}$$

We usually write the model without the shape reminders as $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This is a short hand notation for:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \beta_0 + \beta_1 x_3 + \varepsilon_3 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix} \quad \text{.....(3.1.18)}$$

It is helpful that the multiple regression story with $K \geq 2$ predictors leads to the same model expression $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (just with different shapes). As a notational convenience, let $p = K + 1$. In the multiple regression case, we have:

$$\underset{nxp}{X} = \begin{bmatrix} 1x_{11} & x_{12} \cdots x_{1k} \\ 1x_{21} & x_{22} \cdots x_{2k} \\ 1, x_{31} & x_{32} \cdots x_{3k} \\ 1, x_{41} & x_{42} \cdots x_{4k} \\ 1, x_{51} & x_{52} \cdots x_{5k} \\ 1, x_{61} & x_{62} \cdots x_{6k} \\ \vdots, \vdots & \vdots \\ 1, x_{n1} & x_{n2} \cdots x_{nk} \end{bmatrix} \text{ and } \underset{px1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{.....(3.1.19)}$$

The detail shown here is to suggest that $\boldsymbol{X}$ is a tall, skinny matrix. We formally require $n \geq p$.

In most applications, n is much, much larger than p. The ratio n/p is often in the hundreds. If it happens that n/p is as small as 5, we will worry that we don't have enough data (reflected in n) to estimate the number of parameters in $\boldsymbol{\beta}$ (reflected in p).

The multiple regression model is now:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1\beta x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + ... + \beta_k x_{1k} + \varepsilon_1 \\ \beta_0 + \beta_1\beta x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + ... + \beta_k x_{2k} + \varepsilon_2 \\ \beta_0 + \beta_1\beta x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + ... + \beta_k x_{3k} + \varepsilon_3 \\ \vdots \\ \beta_0 + \beta_1\beta x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + ... + \beta_k x_{nk} + \varepsilon_n \end{bmatrix} \quad \text{.....(3.1.20)}$$

The model form $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is thus completely general.

The assumptions on the noise terms can be written as $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and Var $(\boldsymbol{\varepsilon}) = \sigma^2\,\mathbf{I}$. The **I** here is the $n \times n$ identity matrix. That is,

$$I = \begin{pmatrix} 1,0,0\cdots & 0 \\ 0,1,0\cdots & 0 \\ 0,0,1\cdots & 0 \\ \vdots & \vdots \\ 0,0,0\cdots & 1 \end{bmatrix} \quad \text{.....(3.1.21)}$$

The variance assumption can be written as var $(\varepsilon) = I = \begin{pmatrix} \sigma^2,0,0\cdots & 0 \\ 0,\sigma^2,0\cdots & 0 \\ 0,0,\sigma^2\cdots & 0 \\ \vdots & \vdots \\ 0,0,0\cdots & \sigma^2 \end{bmatrix}$. You may see this expressed as cov $(\varepsilon_i,\varepsilon_j) = \sigma^2\,\delta_{ij}$, where,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

We will call ***b*** as the estimate for unknown parameter vector β. You will also find the notation $\hat{\beta}$ as the estimate. Once we get ***b***, we can compute the *fitted* vector $\boldsymbol{Y}\hat{} = \boldsymbol{X}\,\boldsymbol{b}$. This fitted value represents annex-post guess at the expected value of ***Y***.

The estimate ***b*** is found so that the fitted vector $\boldsymbol{Y}\hat{}$ is close to the actual data vector ***Y***. Closeness is defined in the least squares sense, meaning that we want to minimize the criterion *Q*, where,

$$Q = \sum_{i=1}^{n} (Y_i - (Xb)_{i^{th}\ entry})^2 \quad \text{.....(3.1.22)}$$

This can be done by differentiating this quantity $p = K + 1$ times, once with respect to $b0$, once with respect to b_1, ….., and once with respect to bK . This is routine in simple regression ($K = 1$), and it's possible with a lot of messy work in general.

It happens that Q is the squared length of the vector difference $\boldsymbol{Y} - \boldsymbol{Xb}$. This means that we can write:

$$Q = \underbrace{(Y - Xb)'}_{1xn}\underbrace{(Y - Xb)}_{nx1} \quad \text{.....(3.1.23)}$$

This represents Q as a 1×1 matrix, and so we can think of Q as an ordinary number.

There are several ways to find the $\boldsymbol{b}$ that minimizes Q. The simple solution we will show here (alas) requires knowing the answer and working backward.

Define the matrix, $\underset{nxn}{H} = \underset{nxp}{X}\left(\underset{pxn}{X'}\ \underset{nxp}{X}\right)^{-1}\underset{pxn}{X'}$. We will call H as the "hat matrix", and it has some important uses. There are several technical comment about H:

(1) Finding $\boldsymbol{H}$ requires the ability to get $\left(\underset{pxn}{X'}\ \underset{nxp}{X}\right)^{-1}$ .This matrix inversion is possible if and only if $\boldsymbol{X}$ has full rank p. Things get very interesting when $\boldsymbol{X}$ almost has full rank p; that's a longer story for another time.

(2) The matrix $\boldsymbol{H}$ is *idempotent*. The defining condition for idempotence is this:

The matrix $\boldsymbol{C}$ is idempotent $\Leftrightarrow \boldsymbol{C}\,\boldsymbol{C} = \boldsymbol{C}$.

Only square matrices can be idempotent.

Since $\boldsymbol{H}$ is square (It is $n \times n$.), it can be checked for idempotence. You will indeed find that $\boldsymbol{H}\,\boldsymbol{H} = \boldsymbol{H}$.

(3) The ith diagonal entry, that in position (i, i), will be identified for later use as the ith leverage value. The notation is usually hi , but you'll also see hii .

Now write $\boldsymbol{Y}$ in the form $\boldsymbol{H}\,\boldsymbol{Y} + (\mathbf{I} - \boldsymbol{H})\,\boldsymbol{Y}$.

Now let's develop Q. This will require using the fact that $\boldsymbol{H}$ is symmetric, meaning $\boldsymbol{H}' = \boldsymbol{H}$.

This will also require using the transpose of a matrix product. Specifically, the property will be:

The second and third summands above are zero, as a consequence of:

$$(\mathbf{I}-H)X = X - H\ X = (\ )-1X - X\ X'X\ X'X = X - X = \mathbf{0}.$$
$$= \{H\ Y - Xb\}'\{H\ Y - Xb\} + ((\mathbf{I}-H)Y)'(\mathbf{I}-H)Y \quad \text{.....(3.1.24)}$$

If this is to be minimized over choices of ***b***, then the minimization can only be done with regard to the first summand, $\{H\ Y - Xb\}'\{H\ Y - Xb\}$. It is possible to make the vector $\boldsymbol{H\ Y - Xb}$ equal to **0** by selecting $b = (X'X)^{-1}X'Y$. This is very easy to see, as: $H = X(X'X)^{-1}X'$. This $b = (X'X)^{-1}X'Y$ is known as the least squares estimate of β.

For the simple linear regression case *K* = 1, the estimate $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ and be found with relative ease. The slope estimate is $b_1 = \frac{s_{xy}}{s_{xx}}$,

where $s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$

and where $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^{\,2} - n(\bar{x})^2$

For the multiple regression case $K \geq 2$, the calculation involves the inversion of the $p \times p$ matrix $\boldsymbol{X'\ X}$. This task is best left to computer software.

There is a computational trick, called "mean-centering," that converts the problem to a simpler one of inverting a $K \times K$ matrix.

The matrix notation will allow the proof of two very helpful facts:

* E (**b**) = **β**. This means that ***b*** is an unbiased estimate of **β**. This is a good thing, but there are circumstances in which biased estimates will work a little bit better.

***** Var (**b**) = σ^2 $(\boldsymbol{X'\ X})^{-1}$. This identifies the variances and covariances of the estimated coefficients. It's critical to note that the separate entries of ***b*** are not statistically independent.

SELF ASSESSMENT EXERCISE

What is linear dependence in matrix algebra? And happens if some of the columns of X are linearly dependent?

4.0 CONCLUSION

Instead of writing regression models in normal function (which is pretty inefficient), we can instead formulate the linear regression function in a matrix form. In a simple and short statement of a matrix regression formulation, it shows that X is an n x 2 matrix, β is a 2 x 1 column vector; Y is n x 1 column vector; and ε is an n x 1 column vector. The matrix X and vector β are multiplied together using the technique of matrix multiplication and added to the vector ε using the technique of matrix addition. The least squares estimates in matrix notation can be as: the inverse of X-transpose times X multiplied by X-transpose times Y [$(X'X)^{-1}X'Y$].

5.0 SUMMARY

This unit explained the basic matrix formulation of a regression model. Emphasis is laid on the use of matrix notation to formulate regression model due to its efficiency. The unit further stressed that, in a simple and short statement of a matrix regression formulation, it shows that X is an n x 2 matrix, β is a 2 x 1 column vector; Y is n x 1 column vector; and ε is an n x 1 column vector. The matrix X and vector β are multiplied together using the technique of matrix multiplication and added to the vector ε using the technique of matrix addition. The least squares estimates in matrix notation can be as: the inverse of X-transpose times X multiplied by X-transpose times Y [$(X'X)^{-1}X'Y$].

6.0 TUTOR-MARKED ASSIGNMENT

(1) The following table (Table 3.1) is students (Y), soap 1 (X_1) and soap 2 (X_2)

Table 3.1

Soap 1(x_1)	4	4.5	5	5.5	6	6.5	7
Soap 2(x_2)	8	9	10	11	12	13	14
Suds (y)	33	42	45	51	53	61	62

Note: all variables are deviation from their means

use matrix notation to show that there is linear dependency between the columns of the X's.

(2) Table 3.2 shows students and soap used

Soap 1(x_1)	4	4.5	5	5.5	6	6.5	7
Suds (y_i)	33	42	45	51	53	61	62

Use both normal regression estimation and matrix notation; prove that both methods produce the same estimates.

Note: $b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X'X)^{-1}X'Y$; $X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$ and $X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$

7.0 REFERENCES/FURTHER READINGS

Draper, N.R. & Smith, H. (1998). Applied Regression Analysis, Third Edition, John Wiley online; doi: 10.1002/9781118625590.

UNIT 2: VECTOR AUTOREGRESSIVE (VAR) MODELS AND CAUSALITY TESTS

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 The VAR Model
    3.2 Causality Test
    3.3 Computer Applications of VAR with Examples
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

It is quite common in economics to have models in which some variables are not only explanatory variables for a given dependent variable, but are also explained by the variables that they are used to determine. In these cases we have models of simultaneous equations, in which it is necessary to identify clearly which are the endogenous and which are the exogenous or predetermined variables. This means that in its general reduced form each equation has the same set of regressors, which leads to the development of VAR models.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- Differentiate between univariate and multivariate time series models.
- Understand Vector Autoregressive (VAR) models and discuss their advantages.
- Understand the concept of causality and its importance in economic applications.
- Use the Granger causality test procedure.
- Use the Sims causality test procedure.
- Estimate VAR models and test for Granger and Sims causality through the use of econometric software.

3.0 MAIN CONTENT

3.1 The VAR Model

When we are not confident that a variable really is exogenous, each variable has to be treated symmetrically. Take, for example, the time series yt that is affected by current and past values of xt and, simultaneously, the time series xt to be a series that is affected by current and past values of the yt series. In this case the simple bivariate model is given by:

$$y_t = \beta_{10} - \beta_{12}x_t + \gamma_{11}y_{t-1} + \gamma_{12}x_{t-1} + u_{yt} \quad \text{.....(3.2.1)}$$

$$x_t = \beta_{20} - \beta_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}x_{t-1} + u_{xt} \quad \text{.....(3.2.2)}$$

where we assume that both yt and xt are stationary, and u_{yt} and u_{xt} are uncorrelated white-noise error terms. Equations (3.2.1) and (3.2.2) constitute a first-order VAR model, because the longest lag length is unity. These equations are not reduced-form equations, since yt has a contemporaneous impact on xt (given by $-\beta_{21}$), and xt has a contemporaneous impact on yt (given by $-\beta_{12}$). Rewriting the system using matrix algebra, we get:

$$\begin{bmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{bmatrix}\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}\begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_{yt} \\ u_{xt} \end{bmatrix} \quad \text{.....(3.2.3)}$$

Or $Bz_t = \Gamma_0 + \Gamma_1 z_{t-1} + u_t$ .....(3.2.4)

Where: $B = \begin{bmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{bmatrix}, zt = \begin{bmatrix} y_t \\ x_t \end{bmatrix}, \Gamma_0 = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix}$

$\Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}, u_t = \begin{bmatrix} u_{yt} \\ u_{xt} \end{bmatrix}$

Multiplying both sides by B^{-1} we obtain:

$$z_t = A_0 + A_1 z_{t-1} + e_t \text{ ....(3.2.5)}$$

Where: $A_0 = B^{-1}\Gamma_0, A_1 = B^{-1}\Gamma_1, e_t = B^{-1}u_t,$

For purposes of notational simplification we can denote as a_{i0} the i^{th} element of the vector A_0; a_{ij} the element in row i and column j of the matrix A_1; and e_{it} as the i_{th} element of the vector e_t. Using this, we can rewrite the VAR model as:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}x_{t-1} + e_{1t} \quad \text{(3.2.6)}$$

$$x_t = a_{20} + a_{21}y_{t-1} + a_{22}x_{t-1} + e_{2t} \quad \text{(3.2.7)}$$

To distinguish between the original VAR model and the system we have just obtained, we call the first a structural or primitive VAR system and the second a VAR in standard (or reduced) form. It is important to note that the new error terms, e_{1t} and e_{2t}, are composites of the two shocks u_{yt} and u_{xt}. Since $e_t = B^{-1}ut$ we can obtain e_{1t} and e_{2t} as:

$$e_{1t} = (u_{yt} + \beta_{12}u_{xt}) / 1 - \beta_{12}\beta_{21}) \quad \text{(3.2.8)}$$

$$e_{2t} = (u_{xt} + \beta_{21}u_{yt}) / 1 - \beta_{12}\beta_{21}) \quad \text{(3.2.9)}$$

Since u_{yt} and u_{xt} are white-noise processes, it follows that both e_{1t} and e_{2t} are also white-noise processes.

SELF ASSESSMENT EXERCISE

If there is simultaneity among a number of variables, then all these variables should be treated in the same way, why?

3.1.1 Pros and Cons of the VAR Models

The VAR model approach has some very good characteristics:

(1) It is very simple. The econometrician does not have to worry about which variables are endogenous or exogenous.
(2) Estimation is also very simple, in the sense that each equation can be estimated separately with the usual OLS method.
(3) Forecasts obtained from VAR models are in most cases better than those obtained from the far more complex simultaneous equation models (see Mahmoud, 1984; McNees, 1986).

However, on the other hand, VAR models have faced severe criticism over various points.

(1) They are atheoretic, in that they are not based on any economic theory.
(2) A second criticism concerns the loss of degrees of freedom. If we suppose that we have a three-variable VAR model and decide to include 12 lags for each variable in each equation, this will entail the estimation of 36 parameters in each

equation plus the equation constant. If the sample size is not sufficiently large, estimating that great a number of parameters will consume many degrees of freedom, thus creating problems in estimation.

(3) Finally, the obtained coefficients of the VAR models are difficult to interpret because of their lack of any theoretical background. To overcome this criticism, the advocates of VAR models estimate so-called impulse response functions. The impulse response function examines the response of the dependent variable in the VAR to shocks in the error terms. The difficult issue here, however, is defining the shocks.

SELF ASSESSMENT QUESTION

Examine the merits and demerits of VAR models

3.2 Causality Tests

One of the good features of VAR models is that they allow us to test for the direction of causality. Causality in econometrics is somewhat different from the concept in everyday use; it refers more to the ability of one variable to predict (and therefore cause) the other. Suppose two variables, say y_t and x_t, affect each other with distributed lags. The relationship between these variables can be captured by a VAR model. In this case it is possible to state that (a) y_t causes x_t; (b) x_t causes y_t; (c) there is a bi-directional feedback (causality among the variables); and (d) the two variables are independent. The problem is to find an appropriate procedure that allows us to test and statistically detect the cause and effect relationship among the variables. Granger (1969) developed a relatively simple test that defined causality as follows: a variable y_t is said to Granger cause x_t if x_t can be predicted with greater accuracy by using past values of the y_t variable rather than not using such past values, all other terms remaining unchanged.

3.2.1 The Granger Causality Test

The Granger causality test for the case of two stationary variables y_t and x_t involves as a first step the estimation of the following VAR model:

$$y_t = \alpha_1 + \sum_{i=1}^{n} \beta_1 x_{t-i} + \sum_{j=1}^{m} \gamma_j y_{t-j} + e_{1t} \quad \text{.....(3.2.10)}$$

$$x_t = \alpha_2 + \sum_{i=1}^{n} \theta_1 x_{t-i} + \sum_{j=1}^{m} \delta_j y_{t-j} + e_{2t} \quad \text{.....(3.2.11)}$$

where it is assumed that both e_{yt} and e_{xt} are uncorrelated white-noise error terms. In this model we can have the following different cases:

Case 1: The lagged x terms in Equation (3.2.10) may be statistically different from zero as a group, and the lagged y terms in Equation (3.2.11) not statistically different from zero. In this case we see x_t causes y_t.

Case 2: The lagged y terms in Equation (3.3.11) may be statistically different from zero as a group, and the lagged x terms in Equation (3.2.10) not statistically different from zero. In this case we see that y_t causes x_t.

Case 3: Both sets of x and y terms are statistically different from zero in Equations (3.2.10) and (3.3.11), so that there is bi-directional causality.

Case 4: Both sets of x and y terms are not statistically different from zero in Equations (3.2.10) and (3.2.11), so that x_t is independent of y_t.

The Granger causality test, then, involves the following procedure. First, estimate the VAR model given by Equations (3.2.10) and (3.2.11). Then check the significance of the coefficients and apply variable deletion tests, first in the lagged x terms for Equation (3.2.10), and then in the lagged y terms for Equation (3.2.11). According to the result of the variable deletion tests we may come to a conclusion about the direction of causality based on the four cases mentioned above. More analytically, and for the case of one equation (we shall examine Equation (3.2.10), and it is intuitive to reverse the procedure to test for Equation (3.2.11)), we perform the following steps:

Step 1: Regress y_t on lagged y terms as in the following model:

$$y_t = \alpha_1 + \sum_{i=1}^{m} \gamma_j y_{t-j} + e_{1t} \quad \text{.....(3.2.12)}$$

and obtain the RSS of this regression (the restricted one) and label it as RSSR.

Step 2: Regress y_t on lagged y terms plus lagged x terms as in the following model:

$$y_t = \alpha_1 + \sum_{i=1}^{n} \beta_i x_{t-i} + \sum_{j=1}^{m} \gamma_j y_{t-j} + e_{1t} \quad \text{.....(3.2.13)}$$

and obtain the RSS of this regression (the unrestricted one) and label it as RSSU .

Step 3: Set the null and the alternative hypotheses as:

$H_0 = \sum_{i=1}^{n} \beta_i = 0$ or x_t does not cause y_t

$H_0 = \sum_{i=1}^{n} \beta_i \neq 0$ or x_t does cause y_t

Step 4: Calculate the F-statistic for the normal Wald test on coefficient restrictions given by:

$$F = \frac{(RSS_R - RSS_U)/m}{RSS_U/n-k)}$$

which follows the $F_{m,n-k}$ distribution. Here k = m + n + 1.

Step 5: If the computed F-value exceeds the F-critical value, reject the null hypothesis and conclude that x_t causes y_t.

SELF ASSESSMENT EXERCISE

1. Enumerate and explain the possible state of Granger causality
2. State and explain the steps or procedures to test for Granger causality in a single equation.

3.2.2 The Sim Causality Test

Sims (1980) proposed an alternative test for causality making use of the fact that in any general notion of causality it is not possible for the future to cause the present. Therefore, when we want to check whether a variable yt causes xt, Sims suggests estimating the following VAR model:

$$y_t = \alpha_1 + \sum_{i=1}^{n} \beta_i x_{t-i} + \sum_{j=1}^{m} \gamma_j y_{t-j} + \sum_{\rho=1}^{k} \varsigma_\rho x_{t+\rho} + e_{1t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.2.14)$$

$$x_t = \alpha_2 + \sum_{i=1}^{n} \theta_i x_{t-i} + \sum_{j=1}^{m} \delta_j y_{t-j} + \sum_{\rho=1}^{k} \xi_\rho x_{t+\rho} + e_{2t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.2.15)$$

The new approach here is that, apart from lagged values of x and y, there are also leading values of x included in the first equation (and similarly, leading values of y in the second equation). Examining only the first equation, if y_t causes x_t then we expect that there is some relationship between y and the leading values of x. Therefore, instead of testing for the lagged values of x_t we test for $\sum_{\rho=1}^{k} \varsigma_\rho = 0$. Note that if we reject the restriction then the causality runs from y_t to x_t, and not vice versa, since the future cannot cause the present. To carry out the test we simply estimate a model with no leading terms (the restricted version) and then the model as it appears in Equation (3.2.14) (the unrestricted model), and then obtain the F-statistic as in the Granger test above. It is unclear which version of the two tests is preferable, and most researchers use both. The Sims test,

however, using more regressors (because of the inclusion of the leading terms), leads to a greater loss of degrees of freedom.

SELF ASSESSMENT EXERCISE

Briefly explain Granger and Sims causality and state their differences

3.3 Computer Applications of VAR in Eviews

In EViews, to estimate a VAR model go to Quick\Estimate VAR. A new window opens that requires the model to be specified. First, we have to specify whether it is an unrestricted VAR (default case) or a cointegrating VAR (we shall discuss this in the next section). Leave this option as it is – that is, unrestricted VAR. Then the endogenous variables for our VAR model need to be defined by typing their names in the required box; the lag length (default is 1 2) by typing the start and end numbers of the lags we want to include; and the exogenous variable, if any (note that the constant is already included in the exogenous variables list).

As an example, we can use the data of manufacturing sector output (MSO) for growth and financial deepening (Private sector credit ratio to GDP [PSC] and stock market turnover ratio to GDP [STR]) variables in Nigeria from 1986 – 2017 given in the file VAR.wf1 (appendix 1). If we include as endogenous variables the series MSO, PSC and STR and estimate the VAR model for 2 lags, we obtain the results reported in Table 3.1. EViews can calculate very quickly the Granger causality test for all the series in the VAR model estimated below:

Vector Autoregression Estimates
Date: 08/29/20 Time: 09:07

Sample (adjusted): 1988 2017
Included observations: 30 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	MSO	PSC	STR
MSO(-1)	1.126280	-0.001745	0.001169
	(0.15254)	(0.00114)	(0.00536)
	[7.38368]	[-1.52931]	[0.21820]
MSO(-2)	-0.201119	0.001825	-0.000495
	(0.15075)	(0.00113)	(0.00530)
	[-1.33413]	[1.61862]	[-0.09347]
PSC(-1)	59.07464	0.537339	-0.613703
	(24.7179)	(0.18491)	(0.86852)
	[2.38996]	[2.90600]	[-0.70661]
PSC(-2)	42.15659	0.345108	0.291689
	(23.5602)	(0.17625)	(0.82784)
	[1.78932]	[1.95810]	[0.35235]
STR(-1)	20.04542	0.247529	0.708869
	(6.37529)	(0.04769)	(0.22401)
	[3.14424]	[5.19021]	[3.16444]
STR(-2)	-31.30572	-0.055663	0.035509
	(7.65655)	(0.05728)	(0.26903)
	[-4.08875]	[-0.97183]	[0.13199]
C	-624.7486	0.482719	4.914156
	(143.590)	(1.07415)	(5.04537)
	[-4.35093]	[0.44940]	[0.97399]
R-squared	0.995823	0.939400	0.523255
Adj. R-squared	0.994734	0.923592	0.398886
Sum sq. resids	1181654.	66.12615	1458.912
S.E. equation	226.6633	1.695597	7.964356
F-statistic	913.9983	59.42343	4.207298
Log likelihood	-201.2866	-54.42366	-100.8319
Akaike AIC	13.88577	4.094911	7.188793
Schwarz SC	14.21272	4.421857	7.515739
Mean dependent	2644.196	12.11333	9.572333
S.D. dependent	3123.485	6.134127	10.27241

Determinant resid covariance (dof adj.)	5964721.
Determinant resid covariance	2687880.
Log likelihood	-349.7684
Akaike information criterion	24.71789
Schwarz criterion	25.69873
Number of coefficients	21

Table 3.22: VAR Granger Causality Results from Eviews

VAR Granger Causality/Block Exogeneity Wald Tests
Date: 08/29/20 Time: 09:13
Sample: 1986 2017
Included observations: 30

Dependent variable: MSO

Excluded	Chi-sq	df	Prob.
PSC	28.25124	2	0.0000
STR	18.12155	2	0.0001
All	32.73529	4	0.0000

Dependent variable: PSC

Excluded	Chi-sq	df	Prob.
MSO	2.675969	2	0.2624
STR	31.08185	2	0.0000
All	32.82883	4	0.0000

Dependent variable: STR

Excluded	Chi-sq	df	Prob.
MSO	0.354166	2	0.8377
PSC	0.534053	2	0.7657
All	0.824661	4	0.9351

above. To do this, choose from the VAR window with the output View/Lag Structure/Granger Causality–Block Exogeneity Tests. The results of this Granger causality test are reported in Table 3.21 and show results for each equation of the VAR model, first for excluding the lagged regressors one by one and then all of them at once. EViews also quickly calculates Granger causality tests for different pairs of variables. This test is different from the one presented above because it assumes only the two variables that are being tested in the pair are endogenous in the VAR model. To do this very quick pairwise test, go to Quick/Group Statistics/Granger Causality Test, and in the window that appears define first the variables to be tested for causality (once again using MSO, PSC and STR) and then the number of lags (default 2) that are needed for the test. By clicking OK we get the results reported in Table 3.22. The results report the null hypothesis, the F-statistic and the probability limit value for all possible pairs of variables. From the probability limit values, it is clear that, at a 95% significance level, the only case for which we can reject the null (prob < 0.05) is for 'PSC does not cause MSO' and 'STR does not cause PSC', concluding that PSC does indeed Granger cause MSO and STR does indeed cause PSC. The null hypothesis cannot be rejected in any other case.

Table 3.23: Granger Pairwise Causality Results from Eviews

Pairwise Granger Causality Tests
Date: 08/29/2020 Time: 09:30
Sample: 1986 2017
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
PSC does not Granger Cause MSO	30	4.44224	0.0223
MSO does not Granger Cause PSC		0.40378	0.6721
STR does not Granger Cause MSO	30	1.09364	0.3505
MSO does not Granger Cause STR		0.15435	0.8578
STR does not Granger Cause PSC	30	14.6795	6.E-05
PSC does not Granger Cause STR		0.25183	0.7793

4.0 CONCLUSION

This unit discussed VAR models, estimates and causality analyses. It is obvious in economics to have simultaneous equations in which it is important to identify which are endogenous and which are exogenous variables. However, if we are not confident that a variable really is exogenous, then each variable has to be treated symmetrically. VAR technique can be applied in such a model and causality estimated.

5.0 SUMMARY

This unit explained VAR technique a model with variables that has to be treated symmetrically since one is not confident of whether the variables one is dealing with is really exogenous. The unit added that VAR model is simple and its estimation is simple because each equation is estimated separately. However, the disadvantage of VAR lies with it being atheoretic. The issue of causality was also discussed, a situation where the direction of cause-effect is detected. The two notable causality are the Granger and Sims. The unit ended with computer applications of VAR models estimations and causality using Eviews.

6.0 TUTORED MARKED ASSIGNMENT

1. Use the data of manufacturing sector output (MSO) for growth and financial deepening (Interest rate spread [IRS], Private sector credit ratio to GDP [PSC] and stock market turnover ratio to GDP [STR], Ratio of financial saving to GDP [RFS], and Stock market capitalization ratio to GDP [SMC]) variables in Nigeria from 1986 – 2017 given in the file VAR.wf1 (appendix 1).

   (1) Formulate a VAR model with the stated variables

   (2) Estimate the parameters and explain

   (3) Find the impulse response and explain

   (4) Estimate the pairwise Granger causality and explain whether causality between the variables exists or not.

7.0 REFERENCES

Asteriou, D. & Hall, S. (2011). Applied Econometrics: A Modern Approach (Revised Edition), New York: Palgrave Macmillam.

UNIT 3: NON STATIONARITY AND UNIT ROOT TEST

1.0 Introduction

2.0 Objective

3.0 Main Contents

3.1 Unit Roots and Spurious Regression

3.2 Testing for Unit Roots
3.3 Computer Applications of Unit Roots Tests in Eviews

4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

The point of this discussion is that formal tests for identifying non-stationarity (or, put differently, the presence of unit roots) are needed. The next section explains what a unit root is and discusses the problems regarding the existence of unit roots in regression models. Formal tests are then presented for the existence of unit roots, followed by a discussion of how results for the above tests can be obtained using EViews. Finally, results are presented from applications on various macroeconomic variables.

2.0 OBJECTIVES

At the end of this unit students should be able to:

- Understand the concept of stationarity.
- Explain the differences between stationary and non-stationary time series processes.
- Understand the importance of stationarity and the concept of spurious regressions.
- Understand the concept of unit roots in time series.
- Understand the meaning of the statement 'the series is integrated for order 1' or I(1).
- Learn the Dickey–Fuller (DF) test procedure for testing for unit roots.
- Differentiate among the three different DF models for unit-root testing.
- Learn the Augmented Dickey–Fuller (ADF) test.
- Learn the Philips-Perron (PP) test procedure.
- Estimate the DF, ADF and PP tests using appropriate software.

3.0 MAIN CONTENT

3.1 Unit Roots and Spurious Regression

3.1.1 Unit Roots

What is unit roots?

Consider AR(1) model:

$$y_t = \phi y_{t-1} + e_t \quad \text{.....(3.3.1)}$$

where e_t is a white-noise process and the stationarity condition is $| \phi | < 1$. In general, there are three possible cases:

Case 1: $| \phi | < 1$ and therefore the series is stationary. A graph of a stationary series for $\phi = 0.67$ is presented in Figure 3.1.

Case 2: $| \phi | > 1$ where the series explodes. A graph of a series for $\phi = 1.26$ is given in Figure 3.2.

Case: $| \phi | = 1$ where the series contains a unit root and is non-stationary. A graph of a series for $\phi = 1$ is given in Figure 3.3.

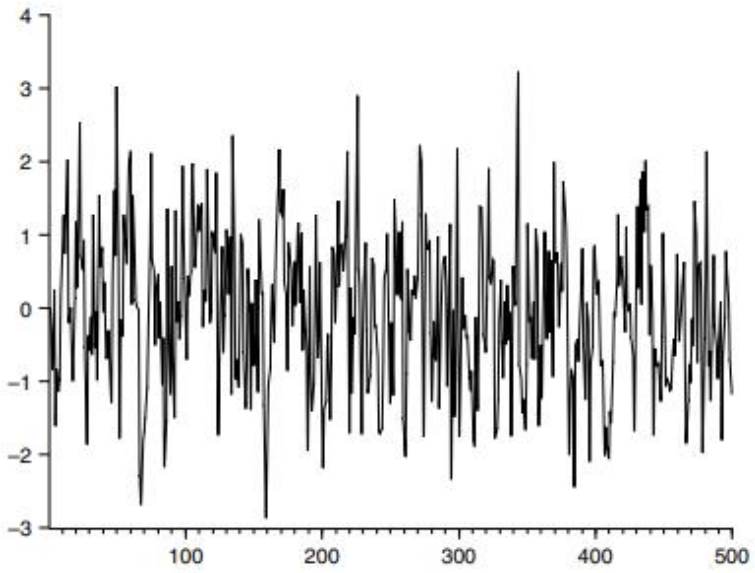


Figure 3.1: A Plot of Stationary AR(1) Model

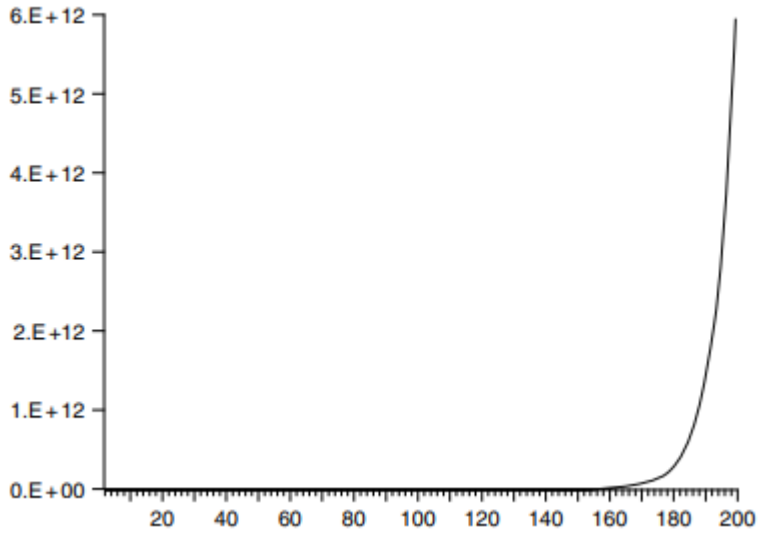


Figure 3.2: Plot of an exploding AR (1) model

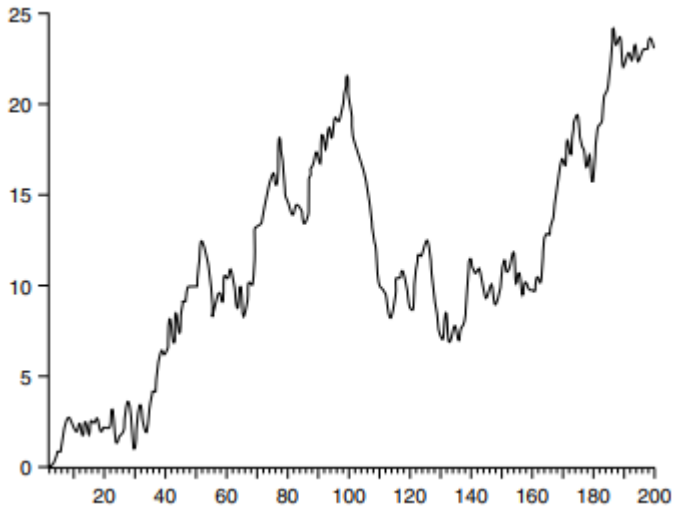


Figure 3.3: Plot of a non-stationary AR (1) model

To reproduce the graphs and the series that are stationary, exploding and nonstationary, type the following commands into EViews (or in a program file and run the program):

```
smpl @first @first+1

genr y=0

genr x=0

genr z=0

smpl @first+1 @last

genr z=0.67*z(-1)+nrnd

genr y=1.16*y(-1)+nrnd

genr x=x(-1)+nrnd

plot y

plot x

plot z
```

So if $\phi = 1$, then y_t contains a unit root. Having $\varphi = 1$ and subtracting y_{t-1} from both sides of Equation (3.3.2) we get:

$$y_t - y_{t-1} = e_t$$

$$\Delta y_t = e_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.3.2)$$

and because e_t is a white-noise process, so y_t is a stationary series. Therefore, after differencing y_t we obtain stationarity.

Definition 1: A series y_t is integrated of order one (denoted by $y_t \sim I(1)$) and contains a unit root if y_t is non-stationary but Δy_t is stationary.

Definition 2: A series y_t is integrated of order d (denoted by $y_t \sim I(d)$) if y_t is non-stationary but Δy_t is stationary; where $\Delta y_t = y_t - y_{t-1}$ and $\Delta^2 yt = \Delta (\Delta y_t) = \Delta y_t - \Delta y_{t-1}$ and so on.

We can summarize the above information under a general rule:

(order of integration of a series) ≡ (number of times the series needs to be differenced in order to become stationary) ≡ (number of unit roots)

SELF ASSESSMENT EXERCISE

Given an AR(1) model:

$$X_t = \varphi X_{t-1} + e_t$$

1. What are the possible cases of φ?
2. Use one of the series in appendix data and Eviews command to show the graph.

3.1.2 Spurious Regression

Most macroeconomic time series are trended and therefore in most cases are nonstationary (see, for example, time plots of the GDP, money supply and CPI for the Nigeria economy). The problem with non-stationary or trended data is that the standard OLS regression procedures can easily lead to incorrect conclusions. It can be shown that in these cases the norm is to get very high values of R^2 (sometimes even higher than 0.95) and very high values of t-ratios (sometimes even greater than 4) while the variables used in the analysis have no interrelationships. Such series are not stationary as the mean is continually rising; however, they are also not integrated, as no amount of differencing can make them stationary. This gives rise to one of the main reasons for taking the logarithm of data before subjecting it to formal econometric analysis. If we take the log of a series, which exhibits an average growth rate, we shall turn it into a

series that follows a linear trend and is integrated. This can easily be seen formally. Suppose we have a series x, which increases by 10% every period, thus:

$$x_t = 1.1x_{t-1}$$

Taking the log of this, we get:

$$\log x_t = \log 1.1 + \log x_{t-1}$$

Now the lagged dependent variable has a unit coefficient and in each period it increases by an absolute amount equal to log (1.1), which is, of course, constant. This series would now be I(1). More formally, consider the model:

$$y_t = \beta_1 + \beta_2 x_t + u_t \quad \text{.....(3.3.3)}$$

where u_t is the error term. The assumptions of the CLRM require both y_t and x_t to have a zero and constant variance (that is, to be stationary). In the presence of nonstationarity, the results obtained from a regression of this kind are totally spurious (using the expression introduced by Granger and Newbold, 1974) therefore these regressions are called spurious regressions.

A spurious regression usually has a very high R2 and t-statistics that appear to provide significant estimates, but the results may have no economic meaning at all. This is because the OLS estimates may not be consistent, and therefore the tests for statistical inference are not valid.

Granger and Newbold (1974) constructed a Monte Carlo analysis generating a large number of y_t and x_t series containing unit roots following the formulae:

$$y_t = y_{t-1} + e_{yt} \quad \text{.....(3.3.4)}$$

$$x_t = x_{t-1} + e_{xt} \quad \text{.....(3.3.5)}$$

where e_{yt} and e_{xt} are artificially generated normal random numbers. Since y_t and x_t are independent of each other, any regression between them should give insignificant results. However, when Granger and Newbold regressed the various y_t's to the x_t's, as shown in Equation (3.3.2), they were surprised to find that they were unable to reject the null hypothesis of $\beta_2 = 0$ for approximately 75% of their cases. They also found that their regressions had very high R^2's and very low values of DW statistics. To see the

spurious regression problem, we can type the following commands into EViews (or into a program file and run the file several times) to see how many times the null of $\beta_2 = 0$ can be rejected. The commands are:

```
smpl @first @first+1

genr y=0 genr x=0

smpl @first+1 @last

genr y=y(-1)+nrnd

genr x=x(-1)+nrnd

scat(r) y x

smpl @first @last

ls y c x
```

To understand the problem of spurious regression better, it might be useful to use an example with real economic data (1990 – 2015). Consider a regression of the logarithm of real GDP (y) to the logarithm of real domestic credit of private sector (x) and a constant. The results obtained from such a regression are the following:

$y_t = 5.941 + 0.560$; $R^2 = 0.926$, DW = 0.196

(0.4820) (0.1763)

Here we see very good t-ratios, with coefficients that have the right signs and more or less plausible magnitudes. The coefficient of determination is very high ($R^2 = 0.926$), but there is also a high degree of autocorrelation (DW = 0.196). This indicates the possible existence of spurious regression. In fact, this regression is totally meaningless because the domestic private sector credit data are for the Benin Republic economy and the real GDP figures are for the Nigerian economy. Therefore, while there should not be any significant relationship, the regression seems to fit the data very well, and this happens because the variables used in this example are, simply, trended (non-stationary). So, the final point is that econometricians should be very careful when working with trended variables.

SELF ASSESSMENT EXERCISE

1. What do you understand by spurious regression?

2. Use the data some of the selected series in appendix 1 and test whether the regressions are spurious or not.

3.2 Testing for Unit Roots

- **Testing for the order of integration**

A test for the order of integration is a test for the number of unit roots, and follows these steps:

Step 1: Test y_t to see if it is stationary. If yes, then $y_t \sim I(0)$; if no, then $y_t \sim I(n)$; $n > 0$.

Step 2: Take first differences of y_t as $\Delta y_t = y_t - y_{t-1}$, and test y_t to see if it is stationary. If yes, then $y_t \sim I(1)$; if no, then $y_t \sim I(n)$; $n > 0$.

Step 3: Take second differences of $\Delta^2 yt$ as $\Delta^2 yt = \Delta yt - \Delta yt{-}1$, and test $\Delta^2 yt$ to see if it is stationary. If yes, then $y_t \sim I(2)$; if no, then $y_t \sim I(n)$; $n > 0$ and so on until it is found to be stationary, and then stop. So, for example, if $\Delta^3 y_t \sim I(0)$, then $\Delta^2 y_t \sim I(1)$, and $y_t \sim I(2)$, and finally $y^t \sim I(3)$; which means that y_t needs to be differenced three times to become stationary.

- **The simple Dickey–Fuller (DF) test for unit roots**

Dickey and Fuller (1979, 1981) devised a formal procedure to test for non-stationarity. The key insight of their test is that testing for non-stationarity is equivalent to testing for the existence of a unit root. Thus the obvious test is the following, which is based on the simple AR(1) model of the form:

$$y_t = \phi y_{t-1} + e_{xt} \quad \text{.....(3.3.6)}$$

What we need to examine here is whether ϕ is equal to 1 (unity and hence 'unit root'). Obviously, the null hypothesis H_0: $\phi = 1$, and the alternative hypothesis H_1: $\varphi < 1$. A different (more convenient) version of the test can be obtained by subtracting y_{t-1} from both sides of Equation (3.3.6):

$$y_t - y_{t-1} = (\phi - 1)y_{t-1} + u_t$$

$$y_t = (\phi - 1)y_{t-1} + u_t$$

$$y_t = \gamma y_{t-1} + u_t \quad \text{.....(3.3.7)}$$

where of course $\gamma = (\phi - 1)$. Now the null hypothesis is H_0: $\gamma = 0$ and the alternative hypothesis Ha: $\gamma < 0$, where if $\gamma = 0$ then y_t follows a pure random-walk model.

Dickey and Fuller (1979) also proposed two alternative regression equations that can be used for testing for the presence of a unit root. The first contains a constant in the random-walk process, as in the following equation:

$$y_t = \alpha_0 + \gamma y_{t-1} + u_t \quad \text{.....(3.3.8)}$$

This is an extremely important case, because such processes exhibit a definite trend in the series when $\gamma = 0$ which is often the case for macroeconomic variables.

The second case is also to allow a non-stochastic time trend in the model, to obtain:

$$y_t = \alpha_0 + a_2 t + \gamma y_{t-1} + u_t \quad \text{.....(3.3.9)}$$

The DF test for stationarity is then simply the normal t-test on the coefficient of the lagged dependent variable y_{t-1} from one of the three models (Equations (3.3.7), (3.3.8) or (3.3.9). This test does not, however, have a conventional t-distribution and so we must use special critical values originally calculated by Dickey and Fuller. MacKinnon (1991) tabulated appropriate critical values for each of the three models discussed above and these are presented in Table 3.21. In all cases, the test focuses on whether $\gamma = 0$. The DF test statistic is the t-statistic for the lagged dependent variable. If the DF statistical value is smaller than the critical value then the null hypothesis of a unit root is rejected and we conclude that y_t is a stationary process.

Table 3.21: Critical values for the Dickey-Fuller test

Model	*1%*	*5%*	*10%*
$\Delta y_{t-1} = \gamma y_{t-1} + u_t$	−2.56	−1.94	−1.62
$\Delta y_{t-1} = \alpha_0 + \gamma y_{t-1} + u_t$	−3.43	−2.86	−2.57
$\Delta y_{t-1} = \alpha_0 + a_2 t + \gamma y_{t-1} + u_t$	−3.96	−3.41	−3.13
Standard critical values	−2.33	−1.65	−1.28

Note: Critical values are taken from MacKinnon (1991).

- **The Augmented Dickey–Fuller (ADF) Test for Unit Roots**

As the error term is unlikely to be white noise, Dickey and Fuller extended their test procedure by suggesting an augmented version of the test that includes extra lagged terms of the dependent variable in order to eliminate autocorrelation. The lag length on these extra terms is either determined by the Akaike information criterion (AIC) or the Schwartz Bayesian criterion (SBC), or more usefully by the lag length necessary to whiten the residuals (that is after each case we check whether the residuals of the ADF regression are autocorrelated or not through LM tests rather than the DW test). The three possible forms of the ADF test are given by the following equations:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_{xt} \quad \text{.....(3.3.10)}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_{xt} \quad \text{.....(3.3.11)}$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \alpha_2 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + u_{xt} \quad \text{.....(3.3.12)}$$

The difference between the three regressions again concerns the presence of the deterministic elements a_0 and a_2t. The critical values for the ADF tests are the same as those given in Table 16.1 for the DF test. Unless the econometrician knows the actual data-generating process, there is a question concerning whether it is most appropriate to estimate Equations (3.3.10), (3.3.11) or (3.3.12). Doldado et al. (1990) suggest a procedure which starts from the estimation of the most general model given by Equation (3.3.12), answering a set of questions regarding the appropriateness of each model and then moving to the next model. This procedure is illustrated in Figure 3.4 (below). It needs to be stressed here that, despite being useful, this procedure is not designed to be applied in a mechanical fashion. Plotting the data and observing the graph is sometimes very useful because it can indicate clearly the presence or not of deterministic regressors. However, this procedure is the most sensible way to test for unit roots when the form of the data-generating process is unknown.

- **The Phillips–Perron (PP) Test**

The distribution theory supporting the DF and ADF tests is based on the assumption that the error terms are statistically independent and have a constant variance. So, when using the ADF methodology, one has to make sure that the error terms are uncorrelated and that they really do have a constant variance. Phillips and Perron (1988) developed

a generalization of the ADF test procedure that allows for fairly mild assumptions concerning the distribution of errors. The test regression for the PP test is the AR(1) process:

$$\Delta y_{t-1} = \alpha_0 + \gamma y_{t-1} + e_t \quad \text{.....(3.3.13)}$$

While the ADF test corrects for higher-order serial correlation by adding lagged differenced terms on the right-hand side, the PP test makes a correction to the t-statistic of the coefficient γ from the AR(1) regression to account for the serial correlation in et. So the PP statistics are only modifications of the ADF t-statistics that take into account the less restrictive nature of the error process. The expressions are extremely complex to derive and are beyond the scope of this text. However, since many statistical packages (one of them is EViews) have routines available to calculate these statistics, it is good for the researcher to test the order of integration of a series by also performing the PP test. The asymptotic distribution of the PP t-statistic is the same as the ADF t-statistic and therefore the MacKinnon (1991) critical values are still applicable. As with the ADF test, the PP test can be performed with the inclusion of a constant, a constant and a linear trend, or neither in the test regression.

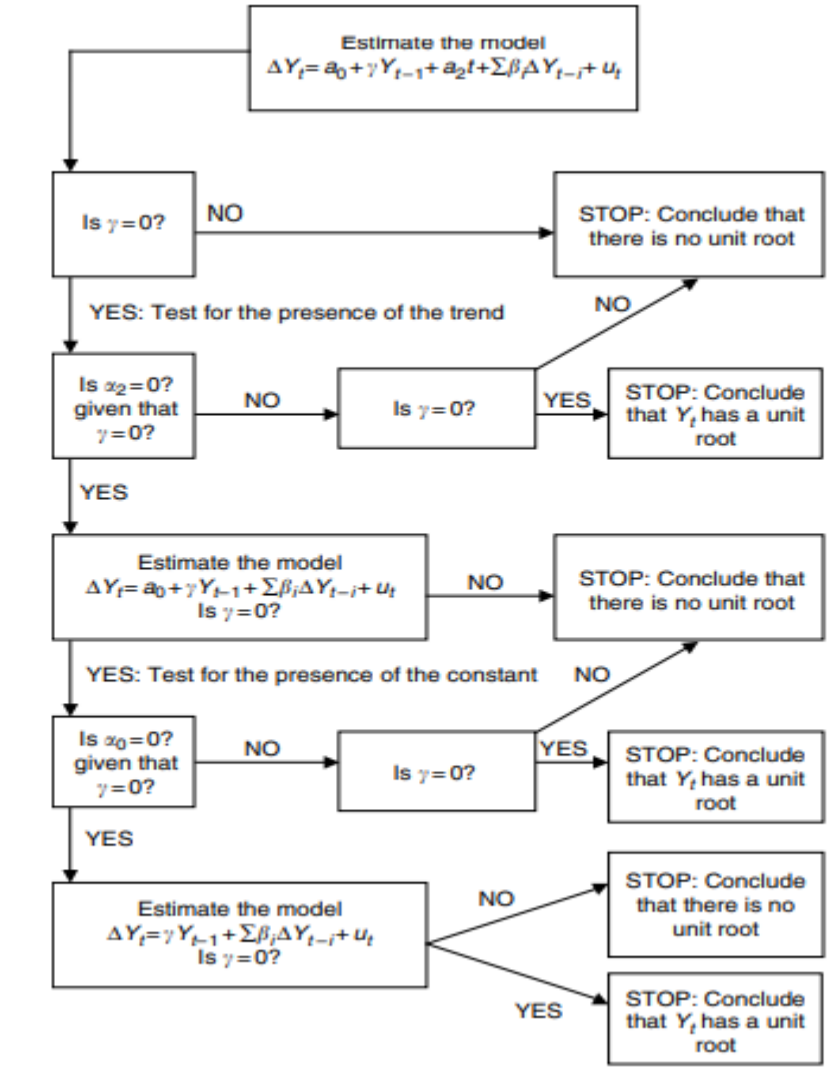


Figure 3.4: Procedure for testing for unit-root tests

Source: Enders (1995).

3.3 Computer Applications of Unit-Root Tests

- **Performing unit-root tests in EViews**

The DF and ADF Test

Step 1: Open the file gdp_uk.wf1 (appendix 1) in EViews by clicking File/Open/Workfile and then choosing the file name from the appropriate path.

Step 2: Let us assume that we want to examine whether the series named MSO contains a unit root. Double-click on the series named 'MSO' to open the series window and choose View/Unit-Root Test ... In the unit-root test dialog box that appears, choose the

type of test (that is the Augmented Dickey–Fuller test, which is the default) by choosing it from the Test Type drop-down menu.

Step 3: We then have to specify whether we want to test for a unit root in the level, first difference or second difference of the series. We can use this option to determine the number of unit roots in the series. As was noted in the theory section, we first start with the level and if we fail to reject the test there we continue with testing for the first differences and so on. So here we first click on levels in the dialog box to see what happens in the levels of the series and then continue, if appropriate, with the first and second differences.

Step 4: We also have to specify which model of the three ADF models we wish to use (that is whether to include a constant, a constant and a linear trend, or neither in the test regression). For the model given by Equation (3.3.10) click on none in the dialog box; for the model given by Equation (3.3.11) click on intercept; and for the model given by Equation (3.3.12) click on intercept and trend. The choice of the model is very important, since the distribution of the test statistic under the null hypothesis differs among these three cases.

Step 5: Finally, we have to specify the number of lagged dependent variables to be included in the model – or the number of augmented terms – to correct for the presence of serial correlation. EViews provides two choices: one is User Specified, which is used only in the event that we want to test for a predetermined specific lag length. If this is the case, we choose this option and enter the number of lags in the box next to it. The second choice is Automatic Selection, which is the default in EViews. If this option is chosen we need to specify from a drop-down menu the criterion we want EViews to use to find the optimal lag length. We have discussed the theory of the AIC and SBC criteria, which are referred to as the Akaike Info Criterion and the Schwarz Info Criterion, respectively, in EViews. We recommend choosing one of the two criteria before going on to the next step. EViews will present the results only for the optimal lag length determined from the criterion you have chosen.

Step 6: Having specified these options, click OK to carry out the test. EViews reports the test statistic together with the estimated test regression.

Step 7: We reject the null hypothesis of a unit root against the one-sided alternative if the ADF-statistic is less than (lies to the left of) the critical value, and conclude that the series is stationary.

Step 8: After running a unit-root test researchers should examine the estimated test regression reported by EViews, especially if unsure about the lag structure or

deterministic trend in the series. You may want to rerun the test equation with a different selection of right-hand variables (add or delete the constant, trend or lagged differences) or lag order.

- **The PP test**

Step 1: Open the file '(appendix 1)' in EViews by clicking File/Open/Workfile and then choosing the file name from the appropriate path.

Step 2: Let us assume that we want to examine whether the series RSF contains a unit root. Double-click on the series named rsf to open the series window and choose View/Unit-Root Test ... In the unit-root test dialog box that appears, choose the type of test (that is the Phillips–Perron test) by selecting it from the Test Type drop-down menu.

Step 3: We then have to specify whether we want to test for a unit root in the level, first difference or second difference of the series. We can use this option to determine the number of unit roots in the series. As was stated in the theory section, first start with the level and if the test is not rejected in the level continue with testing for the first differences and so on. So here we first click on levels to see what happens in the levels of the series, and then continue, if appropriate, with the first and second differences.

Step 4: We also have to specify which model of the three to be used (that is whether to include a constant, a constant and a linear trend or neither in the test regression). For the random-walk model, click on none in the dialog box; for the random walk with drift model click on intercept; and for the random walk with drift and with deterministic trend model click on intercept and trend.

Step 5: Finally, for the PP test specify the lag truncation to compute the Newey–West heteroskedasticity and autocorrelation (HAC) consistent estimate of the spectrum at zero frequency.

Step 6: Having specified these options, click OK to carry out the test. EViews reports the test statistic together with the estimated test regression.

Step 7: We reject the null hypothesis of a unit root against the one-sided alternative if the ADF-statistic is less than (lies to the left of) the critical value.

Computer example: unit-root tests for the financial development and economic growth.

Consider again the data we described in the computer example of the previous chapter for the Granger causality tests. Here we report results of tests for unit roots and orders of integration of all the variables (see file finance.wf1, appendix 1).

We begin the ADF test procedure by examining the optimal lag length using Akaike's FPE criteria; we then proceed to identify the probable order of stationarity. The results of the tests for all the variables and for the three alternative models are presented in Table 3.4, first for their logarithmic levels and then (in cases where we found that the series contain a unit root) for their first differences and so on. The results indicate that each of the series is non-stationary when the variables are defined in levels. First differencing the series removes the non-stationary components in all cases and the null hypothesis of non-stationarity is clearly rejected at the 5% significance level, suggesting that all our variables are integrated of order one, as was expected. The results of the PP tests are reported in Table 3.5, and are not fundamentally different from the respective ADF results. (The lag truncations for the Bartlett kernel were chosen according to Newey and West's (1987) suggestions.) Analytically, the results from the tests on the levels of the variables point clearly to the presence of a unit root in all cases apart from the claims ratio, which appears to be integrated of order zero. The results after first differencing the series robustly reject the null hypothesis of the presence of a unit root, suggesting therefore that the series are integrated of order one.

4.0 CONCLUSION

In this unit emphasis is laid on unit root. In stationary time series, shocks will be temporary, and over time their effects will be eliminated as the series revert to their long-run mean values. On the other hand, non-stationary time series will necessarily contain permanent components. Therefore, the mean and/or the variance of a non-stationary time series will depend on time, which leads to cases where a series (a) has no long-run mean to which the series returns; and (b) the variance will depend on time and will approach infinity as time goes to infinity.

5.0 SUMMARY

This unit explained the concept of stationarity and the differences between stationary and non-stationary time series processes. The importance of stationarity and the concept of spurious regressions were highlighted. The unit further discussed the concept of unit roots in time series and the meaning of the statement 'the series is integrated for order 1' or I(1). The different test for unit roots such as Dickey–Fuller (DF) test procedure for testing for unit roots, the three different DF models for unit-root testing, the Augmented Dickey–Fuller (ADF) test, the Philips-Perron (PP) test procedure were

discussed while the estimation of the DF, ADF and PP tests using appropriate software were explained.

6.0 TUTORED MARKED ASSIGNMENT

1. Explain why it is important to test for stationarity.

2. Describe how a researcher can test for stationarity.

3. Explain the term spurious regression and provide an example from economic time series data.

4. The file (Appendix 1) contains data from various macroeconomic indicators of the Nigerian economy. Check for the order of integration of all the variables using both the ADF and PP tests. Summarize your results in a table and comment on them.

UNIT 4: COINTEGRATION AND ERROR-CORRECTION MODELS

1.0 Introduction
2.0 Objectives
3.0 Main Contents
    3.1 What is Cointegration
    3.2 Cointegration and the Error Correction Mechanism (ECM)
    3.3 Testing for Cointegration
    3.4 Computer Examples of Cointegration
4.0 Conclusion
5.0 Summary
6.0 Tutor Marked Assignment
7.0 References/Further Readings

1.0 INTRODUCTION

The basic idea of this unit follows from our explanation of spurious regression in the previous unit, which showed that if the two variables are non-stationary we can represent the error as a combination of two cumulated error processes. These cumulated error processes are often called stochastic trends and normally we would expect them to combine to produce another non-stationary process. However, in the special case that X and Y are in fact related we would expect them to move together so the two stochastic trends would be very similar. When we put them together it should be possible to find a combination of them that eliminates the nonstationarity. In this special case we say that the variables are cointegrated. In theory, this should only happen when there is truly a relationship linking the two variables, so cointegration becomes a very powerful way of detecting the presence of economic structures.

2.0 OBJECTIVES

At the end of this unit, students should be able to:

- Understand the concept of cointegration in time series.
- Appreciate the importance of cointegration and long-run solutions in econometric applications.
- Understand the error-correction mechanism and its advantages.
- Test for cointegration using the Engle–Granger approach.
- Test for cointegration using the Johansen approach.
- Obtain results of cointegration tests using appropriate econometric software.
- Estimate error-correction models using appropriate econometric software.

3.0 MAIN CONTENT

3.1 What is Cointegration?

The main message from previous unit was that trended time series can potentially create major problems in empirical econometrics because of spurious regressions. We also made the point that most macroeconomic variables are trended and therefore the spurious regression problem is highly likely to be present in most macroeconometric models. One way of resolving this is to difference the series successively until stationarity is achieved and then use the stationary series for regression analysis. However, this solution is not ideal. There are two main problems with using first differences. If the model is correctly specified as a relationship between y and x (for example) and we difference both variables, then implicitly we are also differencing the error process in the regression. This would then produce a non-invertible moving average error process and would present serious estimation difficulties. The second problem is that if we difference the variables the model can no longer give a unique long-run solution. By this we mean that if we pick a particular value for x then regardless of the initial value for y the dynamic solution for y will eventually converge on a unique value. So, for example, if $y = 0.5x$ and we set $x = 10$, then $y = 5$. But if we have the model in differences, $y_t - y_{t-1} = 0.5(x_t - x_{t-1})$ then even if we know that $x = 10$ we cannot solve for y without knowing the past value of y and x, and so the solution for y is not unique, given x. The desire to have models that combine both short-run and long-run properties, and at the same time maintain stationarity in all of the variables, has led to a reconsideration of the problem of regression using variables that are measured in their levels.

The basic thrust of this unit follows from our explanation of spurious regression in the previous unit, which indicated that if the two variables are non-stationary we can represent the error as a combination of two cumulated error processes. These cumulated error processes are often called stochastic trends and normally we would expect them to combine to produce another non-stationary process. However, in the special case that X and Y are in fact related we would expect them to move together so the two stochastic trends would be very similar. When we put them together it should be possible to find a combination of them that eliminates the non-stationarity. In this special case we say that the variables are cointegrated. In theory, this should only happen when there is truly a relationship linking the two variables, so cointegration becomes a very powerful way of detecting the presence of economic structures. Cointegration then becomes an overriding requirement for any economic model using non-stationary time series data. If the variables do not cointegrate we have problems of spurious regression and econometric work becomes almost meaningless. On the other hand, if the stochastic trends do cancel then we have cointegration and, as we shall see later, everything works

even more effectively than we previously might have thought. The key point here is that, if there really is a genuine long-run relationship between Y_t and X_t, then despite the variables rising over time (because they are trended), there will be a common trend that links them together. For an equilibrium or long-run relationship to exist, what we require, then, is a linear combination of Y_t and X_t that is a stationary variable (an I(0) variable). A linear combination of Y_t and X_t can be taken directly from estimating the following regression:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \text{.....(3.4.1)}$$

And taking the residuals:

$$\hat{u}_t = Y_t - \hat{\beta}_1 + \hat{\beta}_2 X_t \quad \text{.....(3.4.2)}$$

If $\hat{u}_t \sim I(0)$ then Y_t and X_t are said to be cointegrated.

3.1.1 Cointegration: A more Mathematical Approach

To put it differently, consider a set of two variables {Y, X} that are integrated of order 1 (that is {Y, X} ~ I(1)) and suppose that there is a vector {θ1, θ2} that gives a linear combination of {Y, X} which is stationary, denoted by:

$$\theta_1 Y_t + \theta_2 X_t = u_t \sim I(0) \quad \text{.....(3.31)}$$

then the variable set {Y, X} is called the cointegration set, and the coefficients vector $\{\theta_1, \theta_2\}$ is called the cointegration vector. What we are interested in is the long-run relationship, which for Y_t is:

$$Y_t^* = \beta X_t \quad \text{.....(3.4.3)}$$

To see how this comes from the cointegration method, we can normalize Equation (3.31) for Y_t to give:

$$Y_t = \frac{\theta_2}{\theta_1} X_t + e_t \quad \text{.....(3.4.4)}$$

where now $Y^* = -(\theta_2/\theta_1)X_t$, which can be interpreted as the long-run or equilibrium value of Y_t (conditional on the values of X_t). We shall return to this point when discussing the error-correction mechanism later in the chapter. For bivariate economic I(1) time series processes, cointegration often manifests itself by more or less parallel

plots of the series involved. As noted earlier, we are interested in detecting long-run or equilibrium relationships and this is mainly what the concept of cointegration allows. The concept of cointegration was first introduced by Granger (1981) and elaborated further by Phillips (1986, 1987), Engle and Granger (1987), Engle and Yoo (1987), Johansen (1988, 1991, 1995a), Stock and Watson (1988), Phillips and Ouliaris (1990), among others. Working in the context of a bi-variate system with at most one cointegrating vector, Engle and Granger (1987) give the formal definition of cointegration between two variables as follows: **Definition 1:** Time series Y_t and X_t are said to be cointegrated of order d, b where $d \geq b \geq 0$, written as Yt, Xt ~ CI(d, b), if (a) both series are integrated of order d, and (b) there exists a linear combination of these variables, say $\beta_1 Y_t + \beta_2 X_t$ which is integrated of order d −b. The vector $\{\beta_1, \beta_2\}$ is called the cointegrating vector. A straightforward generalization of the above definition can be made for the case of n variables, as follows:

Definition 2: If Zt denotes an n×1 vector of series Z_{1t}, Z_{2t}, Z_{3t}, ... , Z_{nt} and (a) each Z_{it} is I(d); and (b) there exists an n × 1 vector β such that $Z_t'\beta \sim I(d-b)$, then $Z_t \sim CI(d,b)$. For empirical econometrics, the most interesting case is where the series transformed with the use of the cointegrating vector become stationary; that is, when d = b, and the cointegrating coefficients can be identified as parameters in the long-run relationship between the variables. The next sections of this unit will deal with these cases.

SELF ASSESSMENT EXERCISE

1. Explain the meaning of cointegration. Why is it so important for economic analysis?

2. Why is it necessary to have series that are integrated of the same order to make cointegration possible? Give examples.

3.2 Cointegration and the Error Correction Mechanism (ECM): A General Approach

3.2.1 The problem

As noted earlier, when there are non-stationary variables in a regression model we may get results that are spurious. So if Y_t and X_t are both I(1), if we regress:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \text{.....(3.4.5)}$$

We will not generally get satisfactory estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$. One way of resolving this is to difference the data to ensure stationarity of our variables. After doing this, $Y_t \sim I(0)$ and $X_t \sim I(0)$, and the regression model will be:

$$\Delta Y_t = \alpha_1 + \alpha_2 \Delta X_t + \Delta u_t \quad \text{.....(3.4.6)}$$

In this case, the regression model may give us correct estimates of the $\hat{\alpha}_1$ and $\hat{\alpha}_2$ parameters and the spurious equation problem has been resolved. However, what we have from Equation (3.4.6) is only the short-run relationship between the two variables. Remember that, in the long-run relationship:

$$Y_t^* = \beta_1 + \beta_2 X_t \quad \text{.....(3.4.7)}$$

so Y_t is bound to give us no information about the long-run behaviour of our model. Knowing that economists are interested mainly in long-run relationships, this constitutes a big problem, and the concept of cointegration and the ECM are very useful to resolve this.

- **Cointegration (again)**

We noted earlier that Y_t and X_t are both I(1). In the special case that there is a linear combination of Y_t and X_t (that is, I(0)), then Y_t and X_t are cointegrated. Thus, if this is the case, the regression of Equation (3.4.7) is no longer spurious, and it also provides us with the linear combination:

$$\hat{u}_t = Y_t - \hat{\beta}_1 + \hat{\beta}_2 X_t \quad \text{.....(3.4.8)}$$

which connects Y_t and X_t in the long run.

3.2.2 The error-correction model (ECM)

If, then, Y_t and X_t are cointegrated, by definitio $\hat{u}_t \square I(0)$. Thus we can express the relationship between Y_t and X_t with an ECM specification as:

$$\Delta Y_t = a_0 + b_1 \Delta X_t - \pi \hat{u}_{t-1} + e_t \quad \text{.....(3.4.9)}$$

which will now have the advantage of including both long-run and short-run information. In this model, b_1 is the impact multiplier (the short-run effect) that measures the immediate impact a change in X_t will have on a change in Y_t. On the other hand, π is the feedback effect, or the adjustment effect, and shows how much of the

disequilibrium is being corrected – that is the extent to which any disequilibrium in the previous period affects any adjustment in Y_t. Of course $\hat{u}_{t-1} = Y_{t-1} - \hat{\beta}_1 - \beta_2 X_{t-1}$, and therefore from this equation β_2 is also the long-run response (note that it is estimated by Equation (3.4.6)). Equation (3.38) now emphasizes the basic approach of the cointegration and error-correction models. The spurious regression problem arises because we are using non-stationary data, but in Equation (3.4.9) everything is stationary, the change in X and Y is stationary because they are assumed to be I(1) variables, and the residual from the levels regression (3.4.8) is also stationary, by the assumption of cointegration. So Equation (3.4.9) fully conforms to our set of assumptions about the classic linear regression model and OLS should perform well.

- **Advantages of the ECM**

The ECM is important and popular for many reasons:

1. First, it is a convenient model measuring the correction from disequilibrium of the previous period, which has a very good economic implication.

2. Second, if we have cointegration, ECMs are formulated in terms of first differences, which typically eliminate trends from the variables involved, and they resolve the problem of spurious regressions.

3. A third, very important, advantage of ECMs is the ease with which they can fit into the general to specific approach to econometric modelling, which is in fact a search for the most parsimonious ECM model that best fits the given data sets.

4. Finally, the fourth and most important feature of the ECM comes from the fact that the disequilibrium error term is a stationary variable (by definition of cointegration). Because of this, the ECM has important implications: the fact that the two variables are cointegrated implies that there is some adjustment process preventing the errors in the long-run relationship from becoming larger and larger.

SELF ASSESSMENT EXERCISE

1. Why is it that when a model is differenced and the regression of such model gives short run estimates?
2. Why is cointegration and ECM useful in resolving the problem of regression estimates of differenced model?
3. What are the advantages of ECM in a model?

3.2.3 Cointegration and the error-correction mechanism: a more mathematical approach

- **A simple model for only one lagged term of X and Y**

The concepts of cointegration and the error-correction mechanism (ECM) are very closely related. To understand the ECM it is better to think of it first as a convenient re-parameterization of the general linear autoregressive distributed lag (ARDL) model. Consider the very simple dynamic ARDL model describing the behaviour of Y in terms of X, as follows:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \gamma_0 X_t + \gamma_1 X_{t-1} + u_t \quad \text{.....(3.4.10)}$$

where the residual ut ~ iid(0, σ^2). In this model the parameter γ_0 denotes the short-run reaction of Y_t after a change in X_t. The long-run effect is given when the model is in equilibrium, where:

$$Y_t^* = \beta_0 + \beta_1 X_t^* \quad \text{.....(3.4.11)}$$

And for simplicity, we assume that:

$$X_t^* = X_t = X_{t-1} = ... = X_{t-p} \quad \text{.....(3.4.12)}$$

Thus, it is given by:

$$Y_t^* = \alpha_0 + \alpha_1 Y_t^* + \gamma_0 X_t^* + \gamma_1 X_t^* + u_t$$

$$Y_t^*(1-\alpha_1) = \alpha_0 + (\gamma_0 + \gamma_1) X_t^* + u_t$$

$$Y_t^* = \frac{\alpha_0}{1-\alpha_1} + \frac{\gamma_0 + \gamma_1}{1-\alpha_1} X_t^* + u_t$$

$$Y_t^* = \beta_0 + \beta_1 X_t + u_t \quad \text{.....(3.4.13)}$$

So the long-run elasticity between Y and X is captured by $\beta_1 = (\gamma_0 + \gamma_1)/(1 - \alpha_1)$. Here, we need to make the assumption that $\alpha_1 < 1$ so that the short-run model in Equation (3.4.10) converges to a long-run solution.

We can then derive the ECM, which is a re-parameterization of the original Equation (3.4.10) model:

$$\Delta Y_t = \gamma_0 \Delta X_t - (1-\alpha)[Y_{t-1} - \beta_0 - \beta_1 X_{t-1}] + u_t \quad \text{.....(3.4.14)}$$

$$\Delta Y_t = \gamma_0 \Delta X_t - \pi[Y_{t-1} - \beta_0 - \beta_1 X_{t-1}] + u_t \quad \text{.....(3.4.15)}$$

What is of importance here is that when the two variables Y and X are cointegrated, the ECM incorporates not only short-run but also long-run effects. This is because the long-run equilibrium $Y_{t-1} - \beta_0 - \beta_1 X_{t-1}$ is included in the model together with the short-run dynamics captured by the differenced term. Another important advantage is that all the terms in the ECM model are stationary, and standard OLS is therefore valid. This is because if Y and X are I(1), then Y and X are I(0), and by definition if Y and X are cointegrated then their linear combination $(Y_{t-1} - \beta_0 - \beta_1 X_{t-1}) \sim I(0)$. A final, very important, point is that the coefficient $\pi = (1 - a_1)$ provides us with information about the speed of adjustment in cases of disequilibrium. To understand this better, consider the long-run condition. When equilibrium holds, then $(Y_{t-1} - \beta_0 - \beta_1 X_{t-1}) = 0$. However, during periods of disequilibrium, this term will no longer be zero and measures the distance the system is away from equilibrium. For example, suppose that because of a series of negative shocks in the economy (captured by the error term u_t) Y_t increases less rapidly than is consistent with Equation (3.4.13). This causes $(Y_{t-1} - \beta_0 - \beta_1 X_{t-1})$ to be negative, because Y_{t-1} has moved below its long-run steady-state growth path. However, since $\pi = (1 - a_1)$ is positive (and because of the minus sign in front of π) the overall effect is to boost Y_t back towards its long-run path as determined by X_t in Equation (3.4.13). The speed of this adjustment to equilibrium is dependent on the magnitude of $(1 - a_1)$. The magnitude of π will be discussed in the next unit.

- **A More General Model for Large Numbers of Lagged Terms**

Consider the following two-variable Yt and Xt ARDL:

$$Y_t = \mu + \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{i=0}^{m} \gamma_i X_{t-i} + u_t \quad \text{.....(3.4.16)}$$

$$Y_t = \mu + + \alpha_1 Y_{t-1} + ... + \alpha_n Y_{t-n} + \gamma_0 X_t + \gamma_1 X_{t-1} + ... + \gamma_m X_{t-m} + u_t \quad \text{.....(3.4.17)}$$

We want to obtain a long-run solution of the model, which would be defined as the point where Y_t and X_t settle down to constant steady-state levels $Y*$ and $X*$, or more simply when:

$$Y_t^* = \beta_0 + \beta_1 X_t^* \quad \text{.....(3.4.18)}$$

And again assume X_t^* is constant:

$$X_t^* = X_t = X_{t-1} = ... = X_{t-m}$$

So, putting this condition into Equation (3.4.3), we get the long-run solution, as:

$$Y^* = \frac{\mu}{1-\sum \alpha_i} + \frac{\sum \gamma_i}{1-\sum \alpha_i} X^* \quad \text{.....(3.4.19)}$$

$$Y^* = \frac{\mu}{1-\alpha_1-\alpha_2-...-\alpha_n} + \frac{(\gamma 1+\gamma 2+...+\gamma)}{1-\alpha_1-\alpha_2-...-\alpha_n} X^* \quad \text{.....(3.4.20)}$$

$$Y^* = \beta_0 + \beta_1 X^* \quad \text{.....(3.4.21)}$$

which means we can define Y^* conditional on a constant value of X at time t as:

$$Y^* = \beta_0 + \beta_1 X_t \quad \text{.....(3.4.22)}$$

Here there is an obvious link to the discussion of cointegration in the previous section. Defining e_t as the equilibrium error as in Equation (3.4.3), we get:

$$e_t = Y_t - Y_t^* = Y_t - B_0 - B_1 X_t \quad \text{.....(3.4.23)}$$

Therefore, what we need is to be able to estimate the parameters B_0 and B_1. Clearly, B0 and B1 can be derived by estimating Equation (3.4.16) by OLS and then calculating $A = \mu / (1-\sum 1-\alpha_i)$ and $B = \sum \gamma_i / (1-\sum 1-\alpha_i)$. However, the results obtained by this method are not transparent, and calculating the standard errors will be very difficult. However, the ECM specification cuts through all these difficulties. Take the following model, which (although it looks quite different) is a re-parameterization of Equation (3.4.3):

$$\Delta Y_t = \mu + \sum_{i=1}^{n-1} \alpha_i \Delta Y_{t-i} + \sum_{i=0}^{m-1} \gamma_i \Delta X_{t-i} + \theta_1 Y_{t-1} + \theta_2 X_{t-1} + u_t \quad \text{.....(3.4.24)}$$

Note: for n = 1 the second term on the left-hand side of Equation (3.4.24) disappears. From this equation we can see, with a bit of mathematics, that:

$$\theta_2 = \sum_{i=1}^{m} \gamma_i \quad \text{.....(3.4.25)}$$

Which is the numerator of the long-run parameter, B_1, and that:

$$\theta_1 = -\left(1 - \sum_{i=1}^{n} \alpha_i\right) \text{...(3.4.26)}$$

So the long-run parameter β_0 is given by $\beta_0 = 1/\theta_1$ and the long-run parameter $\beta_1 = -\theta_2/\theta_1$. Therefore the level terms of Y_t and X_t in the ECM tell us exclusively about the long-run parameters. Given this, the most informative way to write the ECM is as follows:

$$\Delta Y_t = \mu + \sum_{i=1}^{n-1} \alpha_i \Delta Y_{t-i} + \sum_{i=0}^{m-1} \gamma_i \Delta X_{t-i} + \theta_1 (Y_{t-1} - \frac{1}{\theta_1} - \frac{\theta_2}{\theta_1} X_{t-1}) + u_t \text{ ....(3.4.27)}$$

$$\Delta Y_t = \mu + \sum_{i=1}^{n-1} \alpha_i \Delta Y_{t-i} + \sum_{i=0}^{m-1} \gamma_i \Delta X_{t-i} + \theta_1 (Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 X_{t-1}) + u_t \text{ ....(3.4.28)}$$

where $\theta_1 = 0$. Furthermore, knowing that $Y_{t-1} - \hat{\beta}_0 - \hat{\beta}_1 X_{t-1} = e_t$, our equilibrium error, we can rewrite Equation (3.3.56) as:

$$\Delta Y_t = \mu + \sum_{i=1}^{n-1} \alpha_i \Delta Y_{t-i} + \sum_{i=0}^{m-1} \gamma_i \Delta X_{t-i} - \pi \hat{e}_{t-1} + \varepsilon_t \text{ ....(3.4.29)}$$

What is of major importance here is the interpretation of π. π is the error-correction coefficient and is also called the adjustment coefficient. In fact, π tells us how much of the adjustment to equilibrium takes place in each period, or how much of the equilibrium error is corrected. Consider the following cases:

(a) If $\pi = 1$ then 100% of the adjustment takes place within a given period, or the adjustment is instantaneous and full.

(b) If $\pi = 0.5$ then 50% of the adjustment takes place in each period.

(c) If $\pi = 0$ then there is no adjustment, and to claim that Y_t^* is the long-run part of Y_t no longer makes sense. We need to connect this with the concept of cointegration. Because of cointegration, $\hat{e}_t \sim I(0)$ and therefore also $\hat{e}_{t-1} \sim I(0)$. Thus, in Equation (3.4.29), which is the ECM representation, we have a regression that contains only I(0) variables and allows us to use both long-run information and short-run disequilibrium dynamics, which is the most important feature of the ECM.

3.3 Testing for Cointegration

3.3.1 Cointegration in Single Equations: The Engle–Granger Approach

Granger (1981) introduced a remarkable link between non-stationary processes and the concept of long-run equilibrium; this link is the concept of cointegration defined above. Engle and Granger (1987) further formalized this concept by introducing a very simple test for the existence of cointegrating (that is long-run equilibrium) relationships. To understand this approach (which is often called the EG approach) consider the following two series, X_t and Y_t, and the following cases:

(a) If $Y_t \sim I(0)$ and $X_t \sim I(1)$, then every linear combination of those two series

$$\theta_1 Y_t + \theta_2 X_t \text{ ……..(3.4.30)}$$

will result in a series that will always be I(1) or non-stationary. This will happen because the behaviour of the non-stationary I(1) series will dominate the behaviour of the I(0) one.

(b) If we have that both X_t and Y_t are I(1), then in general any linear combination of the two series, say

$$\theta_1 Y_t + \theta_2 X_t \text{ ……..(3.4.31)}$$

will also be I(1). However, though this is the more likely case, there are exceptions to this rule, and we might find in rare cases that there is a unique combination of the series, as in Equation (3.4.31) above, that is I(0). If this is the case, we say that X_t and Y_t are cointegrated of order (1, 1).

Now the problem is how to estimate the parameters of the long-run equilibrium relationship and check whether or not we have cointegration. Engle and Granger proposed a straightforward method involving four steps.

Step 1: test the variables for their order of integration. By definition, cointegration necessitates that the variables be integrated of the same order. Thus the first step is to test each variable to determine its order of integration. The DF and ADF tests can be applied in order to infer the number of unit roots (if any) in each of the variables. We can differentiate three cases which will either lead us to the next step or will suggest stopping:

(a) if both variables are stationary (I(0)), it is not necessary to proceed, since standard time series methods apply to stationary variables (in other words, we can apply classical

regression analysis); (b) if the variables are integrated of different order, it is possible to conclude that they are not cointegrated; and

(c) if both variables are integrated of the same order we proceed with step 2.

Step 2: estimate the long-run (possible cointegrating) relationship If the results of step 1 indicate that both X_t and Y_t are integrated of the same order (usually in economics, I(1)), the next step is to estimate the long-run equilibrium relationship of the form:

$$Y_t = \beta_1 + \beta_2 X_t + e_t \quad \text{.....(3.4.32)}$$

and obtain the residuals of this equation. If there is no cointegration, the results obtained will be spurious. However, if the variables are cointegrated, then OLS regression yields 'super-consistent' estimators for the cointegrating parameter $\hat{\beta}_2$.

Step 3: check for (cointegration) the order of integration of the residuals To determine if the variables are in fact cointegrated, denote the estimated residual sequence from this equation by $\hat{e}_t$. Thus, $\hat{e}_t$ is the series of the estimated residuals of the long-run relationship. If these deviations from long-run equilibrium are found to be stationary, then X_t and Y_t are cointegrated. We perform a DF test on the residual series to determine their order of integration. The form of this DF test is:

$$\Delta\hat{e}_t = \alpha_1 \hat{e}_{t-1} + \sum_{i=1}^{n} \delta_i \Delta\hat{e}_{t-i} + v_t \quad \text{.....(3.4.33)}$$

Note that because $\hat{e}_t$ is a residual we do not include a constant or a time trend. The critical values differ from the standard ADF values, being more negative (typically around −3.5). Critical values are provided in Table 3.3.25. Obviously, if we find that $\hat{e}_t \sim I(0)$, we can reject the null that the variables X_t and Y_t are not cointegrated; similarly, if we have a single equation with more than just one explanatory variable.

Step 4: estimate the ECM If the variables are cointegrated, the residuals from the equilibrium regression can be used to estimate the ECM and to analyse the long-run and short-run effects of the variables as well as to see the adjustment coefficient, which is the coefficient of the lagged residual terms of the long-run relationship identified in step 2. At the end, the adequacy of the model must always be checked by performing diagnostic tests.

Table 3.3.25: Critical values for the null of no cointegration

	1%	5%	10%
No lags	-4.07	-3.37	-3.3
Lags	-3.73	-3.17	-2.91

Source: Engel and Granger (1987)

- **Drawbacks of the EG Approach**

One of the best features of the EG approach is that it very easy both to understand and to implement. However, there are important shortcomings in the Engle–Granger methodology:

1. One very important issue is related to the order of the variables. When estimating the long-run relationship, one has to place one variable in the left-hand side and use the others as regressors. The test does not say anything about which of the variables can be used as a regressor and why. Consider, for example, the case of just two variables, X_t and Y_t. One can either regress Y_t on X_t (that is $Y_t = a + \beta X_t + u_{1t}$) or choose to reverse the order and regress X_t on Y_t (that is $X_t = a + \beta Y_t +$ u2t). It can be shown, with asymptotic theory, that as the sample goes to infinity, the test for cointegration on the residuals of those two regressions is equivalent (that is there is no difference in testing for unit roots in u1t and u2t). However, in practice in economics, there are rarely very big samples and it is therefore possible to find that one regression exhibits cointegration while the other does not. This is obviously a very undesirable feature of the EG approach, and the problem becomes far more complicated when there are more than two variables to test.
2. A second problem is that when there are more than two variables there may be more than one cointegrating relationship, and the Engle–Granger procedure using residuals from a single relationship cannot treat this possibility. So a most important point is that it does not give us the number of cointegrating vectors.
3. A third problem is that it relies on a two-step estimator. The first step is to generate the residual series and the second is to estimate a regression for this series to see whether the series is stationary or not. Hence, any error introduced in the first step is carried into the second.

All these problems are resolved with the use of the Johansen approach that will be examined later.

3.3.2 Engel-Granger Apporach in Econometric Softwares

- **The Engel-Granger Approach in EViews**

The EG test is very easy to perform and does not require any more knowledge regarding the use of EViews. For the first step, ADF and PP tests on all variables are needed to determine the order of integration of the variables. If the variables (let's say X and Y) are found to be integrated of the same order, then the second step involves estimating the long-run relationship with simple OLS. So the command here is simply:

ls X c Y or ls Y c X

depending on the relationship of the variables (see the list of drawbacks of the EG approach in the section above). You need to obtain the residuals of this relationship, which are given by:

genr res_000 = resid

where instead of 000 a different alphanumeric name can be entered to identify the residuals in question. The third step (the actual test for cointegration) is a unit-root test on the residuals, for which the command is:

adf res_000

for no lags; or:

adf (4) res_000

for 4 lags in the augmentation term, and so on. A crucial point here is that the critical values for this test are not those reported in EViews, but the ones given in Table 3.3.25 in this text.

- **The Engel-Granger Approach in Stata**

The commands for Stata are:

regress y x

predict res_000 , residuals

dfuller res_000 , noconstant

for no lags or the simple DF test; or:

dfuller res_000 , noconstant lags(4)

to include 4 lags in the augmentation term, and so on.

3.3.3 Cointegration in Multiple Equations and the Johansen Approach

It was mentioned earlier that if there are more than two variables in the model, there is a possibility of having more than one cointegrating vector. This means that the variables in the model might form several equilibrium relationships governing the joint evolution of all the variables. In general, for n number of variables there can be only up to n − 1 cointegrating vectors. Therefore, when n = 2, which is the simplest case, if cointegration exists then the cointegrating vector is unique.

Having n > 2 and assuming that only one cointegrating relationship exists where there are actually more than one is a serious problem that cannot be resolved by the EG single-equation approach. Therefore an alternative to the EG approach is needed, and this is the Johansen approach for multiple equations.

To present this approach, it is useful to extend the single-equation error-correction model to a multivariate one. Let us assume that we have three variables, Y_t, X_t and W_t which can all be endogenous; that is we have it that (using matrix notation for $Z_t = [Y_t, X_t, W_t]$)

$$Z_t = A_1 Z_{t-1} + A_2 Z_{t-2} + ... + A_k Z_{t-k} + u_t \quad \text{.....(3.4.34)}$$

which is comparable to the single-equation dynamic model for two variables Yt and X_t given in Equation (3.4.14). Thus it can be reformulated in a vector error-correction model (VECM) as follows:

$$\Delta Z_t = \Gamma_1 \Delta Z_{t-1} + \Gamma_2 \Delta Z_{t-2} + ... + \Gamma_k Z_{t-k} + \Pi Z_{t-1} + u_t \quad \text{.....(3.4.35)}$$

where $\Gamma_i = (I - A_1 - A_2 - \cdots - A_k)$ $(i = 1, 2, ..., k-1)$ and $\Pi_i = -(I - A_1 - A_2 - \cdots - A_k)$. Here we need to examine carefully the 3 × 3 Π matrix. (The Π matrix is 3 × 3 because we assume three variables in $Z_t = [Y_t, X_t, W_t]$.) The matrix contains information regarding the long-run relationships. We can decompose $\Pi = \alpha\beta'$ where α will include the speed of adjustment to equilibrium coefficients while β' will be the long-run matrix of coefficients.

Therefore the $\beta' Z_{t-1}$ term is equivalent to the error-correction term ($Y_{t-1} - \beta_0 - \beta_1 X_{t-1}$) in the single-equation case, except that now $\beta' Z_{t-1}$ contains up to (n − 1) vectors in a multivariate framework. For simplicity, we assume that k = 2, so that we have only two lagged terms, and the model is then the following:

$$\begin{pmatrix} \Delta Y_t \\ \Delta X_t \\ \Delta w_t \end{pmatrix} = \Gamma_1 \begin{pmatrix} \Delta Y_{t-1} \\ \Delta X_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + \Pi \begin{pmatrix} \Delta Y_{t-1} \\ \Delta X_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + e_t \quad \text{.....(3.4.36)}$$

Or

$$\begin{pmatrix} \Delta Y_t \\ \Delta X_t \\ \Delta w_t \end{pmatrix} = \Gamma_1 \begin{pmatrix} \Delta Y_{t-1} \\ \Delta X_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{11}\ \alpha_{12} \\ \alpha_{21}\ \alpha_{22} \\ \alpha_{31}\ \alpha_{32} \end{pmatrix} \begin{pmatrix} \beta_{11}\beta_{21}\beta_{31} \\ \beta_{12}\beta_{22}\beta_{32} \end{pmatrix} + e_t \quad \text{.....(3.4.37)}$$

Let us now analyse only the error-correction part of the first equation (that is for Yt on the left-hand side), which gives:

$$\Pi_1 Z_{t-1} = ([\alpha_{11}\beta_{11} + \alpha_{12}\beta_{12}][\alpha_{11}\beta_{21} + \alpha_{12}\beta_{22}]$$

$$[\alpha_{11}\beta_{31} + \alpha_{12}\beta_{32}]) \begin{pmatrix} Y_{t-1} \\ X_{t-1} \\ W_{t-1} \end{pmatrix} \quad \text{.....(3.4.38)}$$

where Π_1 is the first row of the matrix. Equation (3.3.67) can be rewritten as:

$$\Pi_1 Z_{t-1} = \alpha_{11}(\beta_{11} Y_{t-1} + \beta_{21} X_{t-1} + \beta_{31} w_{t-1})$$

$$+\alpha_{12}(\beta_{12} Y_{t-1} + \beta_{22} X_{t-1} + \beta_{32} w_{t-1}) \quad \text{.....(3.4.39)}$$

which shows clearly the two cointegrating vectors with their respective speed of adjustment terms α_{11} and α_{12}.

- **Advantages of the Multiple-equation Approach**

So, from the multiple-equation approach we can obtain estimates for both cointegrating vectors from Equation (3.3.68), while with the simple equation we have only a linear combination of the two long-run relationships.

Also, even if there is only one cointegrating relationship (for example the first only) rather than two, with the multiple-equation approach we can calculate all three differing speeds of adjustment coefficients (α_{11} α_{21} α_{31}).

Only when $\alpha_{21} = \alpha_{31} = 0$, and only one cointegrating relationship exists, can we then say that the multiple-equation method is the same (reduces to the same) as the single-equation approach, and therefore there is no loss from not modelling the determinants of X_t and W_t. Here, it is good to mention too that when $\alpha_{21} = \alpha_{31} = 0$, this is equivalent to X_t and W_t being weakly exogenous.

- **The Johansen approach (again)**

Let us now go back and examine the behaviour of the matrix under different circumstances. Given that Z_t is a vector of non-stationary I(1) variables, then ΔZ_{t-1} are I(0) and ΠZ_{t-1} must also be I(0) in order to have that $u_t \sim$ I(0) and therefore to have a well-behaved system.

In general, there are three cases for ΠZ_{t-1} to be I(0):

Case 1: When all the variables in Zt are stationary. Of course, this case is totally uninteresting since it implies that there is no problem of spurious regression and the simple VAR in levels model can be used to model this case.

Case 2: When there is no cointegration at all and therefore the matrix is an n × n matrix of zeros because there are no linear relationships among the variables in Z_t. In this case the appropriate strategy is to use a VAR model in first differences with no long-run elements as a result of the non-existence of long-run relationships.

Case 3: When there exist up to (n−1) cointegrating relationships of the form $\beta' Z_{t-1} \sim I(0)$. In this particular case, r ≤ (n − 1) cointegrating vectors exist in β. This simply means that r columns of β form r linearly independent combinations of the variables in Z_t, each of which is stationary. Of course, there will also be (n − r) common stochastic trends underlying Z_t.

Recall that $\Pi = \alpha\beta'$ and so in case 3 above, while the matrix will always be dimensioned $n \times n$, the α and β matrices will be dimensioned n × r. This therefore imposes a rank of r on the matrix, which also imposes only r linearly independent rows in this matrix. So underlying the full size matrix is a restricted set of only r cointegrating vectors given by βZ_{t-1}. Reduced rank regression, of this type, has been available in the statistics literature for many years, but it was introduced into modern econometrics and linked with the

analysis of non-stationary data by Johansen (1988). Going back to the three different cases considered above regarding the rank of the matrix we have: Case 1: When has a full rank (that is there are r = n linearly independent columns) then the variables in Zt are I(0).

Case 2: When the rank of is zero (that is there are no linearly independent columns) then there are no cointegrating relationships.

Case 3: When has a reduced rank (that is there are $r \leq (n - 1)$ linearly independent columns) and therefore there are $r \leq (n - 1)$ cointegrating relationships.

- **The steps of the Johansen approach in practice**

Step 1: testing the order of integration of the variables As with the EG approach, the first step in the Johansen approach is to test for the order of integration of the variables under examination. It was noted earlier that most economic time series are non-stationary and therefore integrated. Indeed, the issue here is to have non-stationary variables in order to detect among them stationary cointegrating relationship(s) and avoid the problem of spurious regressions. It is clear that the most desirable case is when all the variables are integrated of the same order, and then to proceed with the cointegration test. However, it is important to stress that this is not always the case, and that even in cases where a mix of I(0), I(1) and I(2) variables are present in the model, cointegrating relationships might well exist. The inclusion of these variables, though, will massively affect researchers' results and more consideration should be applied in such cases. Consider, for example, the inclusion of an I(0) variable. In a multivariate framework, for every I(0) variable included in the model the number of cointegrating relationships will increase correspondingly. We stated earlier that the Johansen approach amounts to testing for the rank of (that is finding the number of linearly independent columns in), and since each I(0) variable is stationary by itself, it forms a cointegrating relationship by itself and therefore forms a linearly independent vector in . Matters become more complicated when we include I(2) variables. Consider, for example, a model with the inclusion of two I(1) and two I(2) variables. There is a possibility that the two I(2) variables cointegrate down to an I(1) relationship, and then this relationship may further cointegrate with one of the two I(1) variables to form another cointegrating vector. In general, situations with variables in differing orders of integration are quite complicated, though the positive thing is that it is quite common in macroeconomics to have I(1) variables. Those who are interested in further details regarding the inclusion of I(2) variables can refer to Johansen's (1995b) paper, which develops an approach to treat I(2) models.

Step 2: setting the appropriate lag length of the model The issue of finding the appropriate (optimal) lag length is very important because we want to have Gaussian error terms (that is standard normal error terms that do not suffer from non-normality, autocorrelation, heteroskedasticity and so on). Setting the value of the lag length is affected by the omission of variables that might affect only the shortrun behaviour of the model. This is because omitted variables instantly become part of the error term. Therefore very careful inspection of the data and the functional relationship is necessary before proceeding with estimation, to decide whether to include additional variables. It is quite common to use dummy variables to take into account short-run 'shocks' to the system, such as political events that had important effects on macroeconomic conditions. The most common procedure in choosing the optimal lag length is to estimate a VAR model including all our variables in levels (non-differenced data). This VAR model should be estimated for a large number of lags, then reducing down by re-estimating the model for one lag less until zero lags are reached (that is we estimate the model for 12 lags, then 11, then 10 and so on until we reach 0 lags). In each of these models we inspect the values of the AIC and the SBC criteria, as well as the diagnostics concerning autocorrelation, heteroskedasticity, possible ARCH effects and normality of the residuals. In general the model that minimizes AIC and SBC is selected as the one with the optimal lag length. This model should also pass all the diagnostic checks.

Step 3: choosing the appropriate model regarding the deterministic components in the multivariate system. Another important aspect in the formulation of the dynamic model is whether an intercept and/or a trend should enter either the short-run or the long-run model, or both models. The general case of the VECM, including all the various options that can possibly arise, is given by the following equation:

$$\Delta Z_t = \Gamma_1 \Delta Z_{t-1} + ... + \Gamma_{k-1} + \alpha(\beta Z_{t-1} \mu_1 1 \delta_1 t + \mu_2 + \delta_2 t + u_t \quad \text{.....(3.4.40)}$$

And for this equation we can see the possible cases. We can have a constant (with coefficient μ_1) and/or a trend (with coefficient δ_1) in the long-run model (the cointegrating equation (CE)), and a constant (with coefficient μ_2) and/or a trend (with coefficient δ_2) in the short-run model (the VAR model). In general, five distinct models can be considered. While the first and the fifth models are not that realistic, all of them are presented for reasons of complementarity.

Model 1: No intercept or trend in CE or VAR ($\delta_1 = \delta 2 = \mu_1 = \mu 2 = 0$). In this case there are no deterministic components in the data or in the cointegrating relations. However, this is quite unlikely to occur in practice, especially as the intercept is generally needed to account for adjustments in the units of measurements of the variables in (Z_{t-1} 1 t).

Model 2: Intercept (no trend) in CE, no intercept or trend in VAR ($\delta_1 = \delta_2 = \mu_2 = 0$). This is the case where there are no linear trends in the data, and therefore the first differenced series have a zero mean. In this case, the intercept is restricted to the long-run model (that is the cointegrating equation) to account for the unit of measurement of the variables in (Z_{t-1} 1 t).

Model 3: Intercept in CE and VAR, no trends in CE and VAR ($\delta_1 = \delta_2 = 0$). In this case there are no linear trends in the levels of the data, but both specifications are allowed to drift around an intercept. In this case, it is assumed that the intercept in the CE is cancelled out by the intercept in the VAR, leaving just one intercept in the short-run model.

Model 4: Intercept in CE and VAR, linear trend in CE, no trend in VAR ($\delta_2 = 0$). In this model a trend is included in the CE as a trend-stationary variable, to take into account exogenous growth (that is technical progress). We also allow for intercepts in both specifications while there is no trend in the short-run relationship.

Model 5: Intercept and quadratic trend in the CE intercept and linear trend in VAR. This model allows for linear trends in the short-run model and thus quadratic trends in the CE. Therefore, in this final model, everything is unrestricted. However, this model is very difficult to interpret from an economics point of view, especially since the variables are entered as logs, because a model like this would imply an implausible ever-increasing or ever-decreasing rate of change.

So the problem is, which of the five different models is appropriate in testing for cointegration. It was noted earlier that model 1 and model 5 are not that likely to happen, and that they are also implausible in terms of economic theory, therefore the problem reduces to a choice of one of the three remaining models (models 2, 3 and 4). Johansen (1992) suggests that the joint hypothesis of both the rank order and the deterministic components need to be tested, applying the so-called Pantula principle. The Pantula principle involves the estimation of all three models and the presentation of the results from the most restrictive hypothesis (that is r = number of cointegrating relations = 0 and model 2) to the least restrictive hypothesis (that is r = number of variables entering the VAR − 1 = n − 1 and model 4). The model-selection procedure then comprises moving from the most restrictive model, at each stage comparing the trace test statistic to its critical value, and stopping only when it is concluded for the first time that the null hypothesis of no cointegration is not rejected.

Step 4: determining the rank of Π or the number of cointegrating vectors According to Johansen (1988) and Johansen and Juselius (1990), there are two methods (and

corresponding test statistics) for determining the number of cointegrating relations, and both involve estimation of the matrix Π. This is a k × k matrix with rank r. The procedures are based on propositions about eigenvalues.

(a) One method tests the null hypothesis, that Rank (Π) = r against the hypothesis that the rank is r + 1. So the null in this case is that there are cointegrating vectors and there are up to r cointegrating relationships, with the alternative suggesting that there are (r + 1) vectors. The test statistics are based on the characteristic roots (also called eigenvalues) obtained from the estimation procedure. The test consists of ordering the largest eigenvalues in descending order and considering whether they are significantly different from zero. To understand the test procedure, suppose we obtained n characteristic roots denoted by $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_n$. If the variables under examination are not cointegrated, the rank of is zero and all the characteristic roots will equal zero. Therefore $(1-\hat{\lambda}_i)$ will be equal to 1 and, since ln(1) = 0, each of the expressions will be equal to zero for no cointegration. On the other hand, if the rank of is equal to 1, then $0 < \lambda_1 < 1$ so that the first expression is $(1-\hat{\lambda}_i) < 0$, while all the rest will be equal to zero. To test how many of the numbers of the characteristic roots are significantly different from zero this test uses the following statistic:

$$\lambda_{\max}(r, r+1) = -T\ln(1-\hat{\lambda}_{r+1}) \quad \text{.....(3.4.41)}$$

As noted above, the test statistic is based on the maximum eigenvalue and is thus called the maximal eigenvalue statistic (denoted by λmax). (b) The second method is based on a likelihood ratio test for the trace of the matrix (and because of that it is called the trace statistic). The trace statistic considers whether the trace is increased by adding more eigenvalues beyond the rth. The null hypothesis in this case is that the number of cointegrating vectors is less than or equal to r. From the previous analysis it should be clear that when all $\hat{\lambda}_i = 0$, then the trace statistic is also equal to zero. On the other hand, the closer the characteristic roots are to unity, the more negative is the $\ln(1-\hat{\lambda}_i)$ term and therefore the larger the trace statistic. This statistic is calculated by:

$$\lambda_{trace}(r) = -T\sum_{i=r+t}^{n}\ln(1-\hat{\lambda}_{r+1}) \quad \text{.....(3.4.42)}$$

The usual procedure is to work downwards and stop at the value of r, which is associated with a test statistic that exceeds the displayed critical value. Critical values for both statistics are provided by Johansen and Juselius (1990) (these critical values are directly provided from EViews after conducting a test for cointegration using the Johansen approach).

Step 5: Testing for Weak Exogeneity.

After determining the number of cointegrating vectors we proceed with tests of weak exogeneity. Remember that the matrix contains information about the long-run relationships, and that $\Pi = \alpha\beta'$, where α represents the speed of adjustment coefficients and β is the matrix of the long-run coefficients. A very useful feature of the Johansen approach for cointegration is that it allows us to test for restricted forms of the cointegrating vectors. Consider the case given by Equation (3.4.36), and from this the following equation:

$$\begin{pmatrix} \Delta Y_t \\ \Delta X_t \\ \Delta w_t \end{pmatrix} = \Gamma_1 \begin{pmatrix} \Delta Y_{t-1} \\ \Delta X_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{11}\ \alpha_{12} \\ \alpha_{21}\ \alpha_{22} \\ \alpha_{31}\ \alpha_{32} \end{pmatrix} \begin{pmatrix} \beta_{11}\beta_{21}\beta_{31} \\ \beta_{12}\beta_{22}\beta_{32} \end{pmatrix} \begin{pmatrix} \Delta Y_{t-1} \\ \Delta X_{t-1} \\ \Delta w_{t-1} \end{pmatrix} + e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots (3.4.43)$$

In this equation it can be seen that testing for weak exogeneity with respect to the long-run parameters is equivalent to testing which of the rows of α are equal to zero. A variable Z is weakly exogenous if it is only a function of lagged variables, and the parameters of the equation generating Z are independent of the parameters generating the other variables in the system. If we think of the variable Y in Equation (3.4.43), it is clearly a function of only lagged variables but in the general form above the parameters of the cointegrating vectors (β) are clearly common to all equations and so the parameters generating Y cannot be independent of those generating X and W as they are the same parameters. However, if the first row of the α matrix were all zeros then the βs would drop out of the Y equation and it would be weakly exogenous. So a joint test that a particular row of α is zero is a test of the weak exogeneity of the corresponding variable. If a variable is found to be weakly exogenous it can be dropped as an endogenous part of the system. This means that the whole equation for that variable can also be dropped, though it will continue to feature on the right-hand side of the other equations.

Step 6: testing for linear restrictions in the cointegrating vectors. An important feature of the Johansen approach is that it allows us to obtain estimates of the coefficients of the matrices α and β, and then test for possible linear restrictions regarding those matrices. Especially for matrix β, the matrix that contains the long run parameters, this is very important because it allows us to test specific hypotheses regarding various theoretical predictions from an economic theory point of view. So, for example, if we examine a money–demand relationship, we might be interested in testing restrictions regarding the long-run proportionality between money and prices, or the relative size of income and interest-rate elasticities of demand for money and so on.

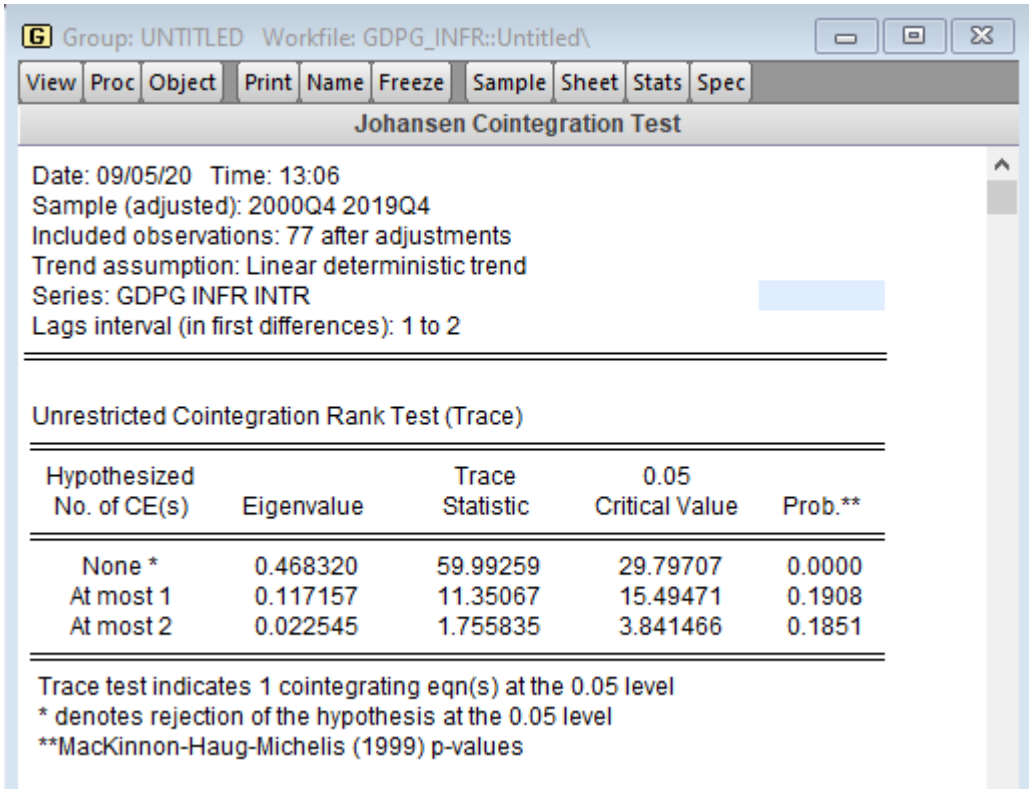
3.4 The Johansen Approach in EViews

The Johansen approach in EViews EViews has a specific command for testing for cointegration using the Johansen approach under group statistics.

Consider the file in appendix 2, which has quarterly data from 2000q1 to 2019q4 for the Nigerian economy and for the following variables: GDPg = the growth rate of GDP; INFL = the rate of inflation and INTR = the interest rate representing the opportunity cost of holding money.

The first step is to determine the order of integration of the variables. To do this, apply unit-root tests on all three variables that are to be tested for cointegration. Apply the Doldado et al. (1990) procedure to choose the appropriate model and determine the number of lags according to the SBC criterion. This model was found to be appropriate and we concluded from that model that there is a unit root in the series (because the ADF-statistic was bigger than the 5% critical value). The results of all tests for levels and first differences are presented in Table 3.3.26

The next step is to go to workfile in Eviews and highlight the variables starting with the dependent variable, in our own case: GDPG INFR INTR. The go to quick then Johansen cointegration. Depending on the model whether 2, 3 or 4, then click on ok.



Group: UNTITLED Workfile: GDPG_INFR::Untitled\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Johansen Cointegration Test

Date: 09/05/20 Time: 13:06
Sample (adjusted): 2000Q4 2019Q4
Included observations: 77 after adjustments
Trend assumption: Linear deterministic trend
Series: GDPG INFR INTR
Lags interval (in first differences): 1 to 2

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.468320	59.99259	29.79707	0.0000
At most 1	0.117157	11.35067	15.49471	0.1908
At most 2	0.022545	1.755835	3.841466	0.1851

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.468320	48.64192	21.13162	0.0000
At most 1	0.117157	9.594834	14.26460	0.2399
At most 2	0.022545	1.755835	3.841466	0.1851

Max-eigenvalue test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

For illustrative purposes for the use of EViews only, we consider the results from model 2 where only one cointegrating vector was found to exist. From the full results (reported in Table 3.3.27) we see that both the trace and the maximal eigenvalue statistics suggest the existence of one cointegrating vectors.

After establishing the number of cointegrating vectors, we proceed with the estimation of the ECM by clicking on Procs/Make Vector Autoregression. EViews here gives us two choices of VAR types; first, if there is no evidence of cointegration we can estimate the unrestricted VAR (by clicking on the corresponding button), or, if there is cointegration we can estimate the VECM. If we estimate the VECM we need to specify (by clicking on the Cointegration menu), which model we want and how many numbers of cointegrating vectors we wish to have (determined from the previous step), and to

impose restrictions on the elements of the α and β matrices by clicking on the VEC restrictions menu. The restrictions are entered as b(1, 1) = 0 for the $\beta_{11} = 0$ restriction. More than one restriction can be entered and they should be separated by commas.

4.0 CONCLUSION

This unit laid emphasis on cointegration and error correction model (ECM). When two variables are non-stationary, the errors can be represented as a combination of two cumulated error processes which are often referred to as stochastic trends. However, in the special case that two variables X and Y are in fact related, one expects them to move together so the stochastic trends would be very similar. In other words, although the two series are individually nonstationary, a linear combination of them is stationary. In the language of econometrics, the two series are cointegrated.if the two variables are cointegrated, the relationship can be expressed with an ECM which includes both longrun and short run information.

5.0 SUMMARY

In this unit, we discussed the concept of cointegration in time series econometrics, the importance of cointegration and the long run solutions in econometric applications. The unit also discussed error correction mechanism (ECM) model and its advantages. When two variables are cointegrated although individually non-stationary, the relationship can be expressed with an ECM which includes both long run and short run information. The unit further talked about Engle Granger and Johansen approaches to cointegration with some computer softwares applications.

6.0 TUTORED-MARKED ASSIGNMENT

1. Explain the meaning of cointegration. Why is it so important for economic analysis?

2. Why is it necessary to have series that are integrated of the same order to make cointegration possible? Give examples.

3. What is the error-correction model? Prove that the ECM is a reparametrization of the ARDL model.

4. What are the features of the ECM that make it so popular in modern econometric analysis?

5. Explain step by step how can one test for cointegration using the Engle–Granger (EG) approach.

6. State the drawbacks of the EG approach, and discuss these with reference to its alternative (that is the Johansen approach).

7. Is it possible to have two I(1) variables and two I(2) variables in a Johansen test for cointegration, and to find that the I(2) variables are cointegrated with the I(1)? Explain analytically.

8. The file (appendix 2) contains data for GDP growth rate and unemployment for the Nigerian economy. Test for cointegration between the two variables using the EG approach and comment on the validity of the Phillips curve theory for the Korean economy.

9. The file (appendix 2) contains data on three variables (GDPG, INFR and INTR). Test the variables for their order of integration and then apply the EG approach to the three different pairs of variables. In which of the pairs do you find cointegration?

7.0 REFERENCES

Engle, R.F. and C.W.J. Granger (1987) 'Co-integration and Error Correction: Representation, Estimation, and Testing', Econometrica, 55, pp. 251–76.

Engle, R.F. and B. Yoo (1987) 'Forecasting and Testing in Cointegrated Systems', Journal of Econometrics, 35, pp. 143–59.

Johansen, S. (1991) 'Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models', Econometrica, 59, pp. 1551–80.

Johansen, S. (1992) 'Determination of Cointegration Rank in the Presence of a Linear Trend', Oxford Bulletin of Economics and Statistics, 54, pp. 383–97.

Johansen, S. (1995a) Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford: Oxford University Press. Johansen, S. (1995b) ‘A Statistical Analysis of I(2) Variables’, Econometric Theory, 11, pp. 25–59.

Johansen, S. and K. Juselius (1990) ‘The Maximum Likelihood Estimation and Inference on Cointegration – with Application to Demand for Money’, Oxford Bulletin of Economics and Statistics, 52, pp. 169–210