NATIONAL OPEN UNIVERSITY OF NIGERIA

FACULTY OF SOCIAL SCIENCES

COURSE CODE: ECO 453

COURSE TITLE: APPLIED ECONOMETRICS 1

NATIONAL OPEN UNIVERSITY OF NIGERIA

APPLIED ECONOMETRICS I
ECO 453

FACULTY OF SOCIAL SCIENCES COURSE GUIDE

Course Writer/Developer: Dr. Likita J. Ogba
Department of Economics Faculty of Social Sciences University of Jos
likiogba@gmail.com

Course Content Editor: Dr. Ganiyat A. Adesina-Uthman, acma,mnes,fifp
Department of Economics Faculty of Social Sciences,
National Open University of Nigeria

Course Reviewer: Dr. Obumneke Ezie
Department of Economics, Bingham University,
Karu, Nasarawa State.
eobumneke@yahoo.com

NATIONAL OPEN UNIVERSITY OF NIGERIA
National Open University of Nigeria Headquarters
University Village, 91 Cadastral Zone
Nnamdi Azikiwe Expressway, Jabi
Abuja

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng
Re-Printed 2023
ISBN:

For
National Open University of Nigeria *Multimedia Technology in Teaching and Learning*

CONTENT

Introduction

Welcome to ECO 453 Applied Econometrics I. The course is available for students in undergraduate Economics. The Course Applied Econometrics I (ECO453) is a core course which carries two credit units. It is prepared and made available to all the students who are taking Economics; a programme tenable in the Faculty of Social Sciences. The course provides an opportunity for students to acquire a detailed knowledge and understanding of theory and applications of econometrics in data for policy interpretations. The course is a useful material to in your academic pursuit as well as in your workplace as economists, managers and administrators. This Course Guide is meant to provide you with the necessary information about the use of data run regression and provide policy interpretations. The course demonstrates the nature of the materials you will be using and how to make the best use of the materials towards ensuring adequate success in your programme as well as the practice of policy analysis. Also included in this course guide are information on how to make use of your time and information on how to tackle the tutor-marked assignment (TMA) questions. There will be tutorial sessions during which your instructional facilitator will take you through your difficult areas and at the same time have meaningful interaction with your fellow learners. Overall, this module will fill an important niche in the study of applied economics which has been missing on the pathway of Economics Students will acquire an understanding of the method of practical data estimation and the skills to evaluate and discuss econometric literature.

What You Will Learn in this Course

Applied Econometrics provides you with the opportunity to gain mastery and an in - depth understanding of application of economics quantitatively. If you devote yourself to practice of the software you will become used to quantitative analysis using different types of data. The data used will form part of the practical focus of real application implementation.

The course is made up of 12 units, covering areas such as:

Meaning of Applied Econometric Research

Time Series and Its Components

Simple Linear Regression Model

Time Series Data Analysis

Multicollinearity

Autoregressive Process

Concept of Stationarity

Cointegration Analysis

Autoregressive Distributed Lag (ARDL) Model

ARDL Post Estimation Tests

Panel Data Regression Model

Fixed Versus Random Effects Panel Data

Testing Fixed and Random Effects

Course Aims

The overall aims of this course include:

i. To introduce you to the major aspects of applied econometrics
ii. To know the basic assumptions of econometric variables that will be estimated
iii. To Know the practical estimation of models using real life data and identify deviations from models
iv. To give you an opportunity to determine model variables stationarity and the use of other options in estimation of econometric models
v. To learn about advanced applications of econometrics and how to use such in policy analysis.
vi. To know how to evaluate and discuss other methods of model estimation

Course Objectives

By the end of this course, you should be able to:

i. Explain simple and multiple regression with respect to economics policy and theory.
ii. Evaluate Nonlinear Regression Models.
iii. Discuss Panel Data Regression Models

iv. Examine Econometric Models

v. Evaluate Autoregressive and Distributed-Lag Models and their applications in the Economy.

vi. Evaluate and discuss other methods of model estimation

Working Through the Course

To complete this course, you are required to read the study units, read the set text books and read other materials that would be provided to you by the National Open University of Nigeria (NOUN). You will also need to undertake practical exercise using Econometric Eviews software this require that you have access to personal computer, purchase and install Eviews for practical. Each unit contains self-assessment exercise; and at certain points during the course, you will be expected to submit assignments. At the end of the course, you will be expected to write a final examination. The course will take you about 12 weeks to complete. Below are the components of the course. What you should do and how to allocate your time to each unit so as to complete the course successfully and on time.

Course Materials

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 as follows:

1. Course Guide

2. Study Units
3. Textbooks
4. Assignment File
5. Presentation schedule

Study Units

There are twelve units in this course, which should be studied carefully. Such units are as follows:

MODULE 1 INTRODUCTION TO ECONOMETRIC RESEARCH USING SOFTWARE

Unit 1 Meaning of Applied Econometric Research

Unit 2: Time Series and Its Components

Unit 3: Simple Linear Regression Model

Unit 4: Time Series Data Analysis

Unit 5: Multicollinearity

MODULE 2 STATIONARITY AND AUTOREGRESSIVE PROCESS

Unit 1. Autoregressive Process

Unit 2: Concept of Stationarity

Unit 3: Cointegration Analysis

Unit 4: Autoregressive Distributed Lag (ARDL) Model

Unit 5: ARDL Post Estimation Tests

MODULE 3: PANEL DATA ESTIMATION

Unit 1: Panel Data Regression Model

Unit 2: Fixed Versus Random Effects Panel Data

Unit 3: Testing Fixed and Random Effects

References and Other Resources

Every unit contains a list of references and further reading. Try to get as many as possible of those textbooks and materials listed. The textbooks and materials are meant to deepen your knowledge of the course.

Assignment File

There are many assignments on this course and you are expected to do all of them by following the schedule prescribed for them in terms of when to attempt them and submit same for grading by your tutor. The marks you obtain for these assignments will count towards the final score.

Presentation Schedule

The Presentation Schedule included in your course materials gives you the important dates for the completion of tutor-marked assignments and attending tutorials. Remember, you are required to submit all your assignments by the due date. You should guard against falling behind in your work.

Assessment

Your assessment will be based on tutor-marked assignments (TMAs) and a final examination which you will write at the end of the course.

Tutor-Marked Assignment

In doing the tutor-marked assignment, you are to apply your transfer knowledge and what you have learnt in the contents of the study units. These assignments which are many in number are expected to be turned in to your Tutor for grading. They constitute 30% of the total score for the course.

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

You should use the time between finishing the last unit and sitting for the examination to revise the entire course material. 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.

Assessment	**Marks**

Assignment (Best three assignments out of the four marked)	30%
Final Examination	70%
Total	100%

Course Overview

The table below brings together the units and the number of weeks you should take to complete them and the assignment that follow them.

Unit	Title of Work	Weekly Activity	Assessment End of Unit
1	Meaning of Applied Econometric Research	1	
2	Simple Linear Regression Model.		
3	How To Run Time Series		1ST Assignment
4	Autoregressive Process		
5	Stationarity		
6	Panel Data Regression Model		
7	Fixed Versus Random Effects Panel Data		2ND Assignment
8	Testing Fixed and Random Effects		
9	Panel Data Estimation in EViews		
10	Dynamic Models		3RD Assignment
11	Autoregressive Distributed Lag (ARDL) Model		
12	ARDL level Relation		4TH Assignment

How To Get the Most from This Course

In distance learning, the study units replace the lecturer. There is the advantage of reading and working through the course material at the pace that suits

the learner best. You are advised to think of it as reading the lecture as against listening to the lecturer. The study units provide exercises for you to do at appropriate periods instead of receiving exercises in the class. Each unit has common features which are designed, purposely, to facilitate your reading. The first feature being an introduction to the unit, the manner in which each unit is integrated with other units and the entire course. The second feature is a set of learning objectives. These objectives should guide your study. After completing the unit, you should go back and check whether you have achieved the objectives or not. The next feature is self-assessment exercises, study questions which are found throughout each unit. The exercises are designed basically to help you recall what you have studied and to assess your learning by yourself. You should do each self-assessment exercise and features are conclusion and summary at the end of each unit. These help you to recall all the main topics discussed in the main content of each unit. These are also tutor-marked assignments at the end of appropriate units. Working on these questions will help you to achieve the objectives of the unit and to prepare for the assignments which you will submit and the final examination. It should take you a couple of hours to complete a study unit, including the exercises and assignments. Upon completion of the first unit, you are advised to note the length of time it took you, and then use this information to draw up a timetable to guide your study of the remaining units. The margins on either sides of each page are meant for you to make notes on main ideas or key points for your usage when revising the course. These features are for your usage to significantly increase your chances of passing the course.

Tutors And Tutorials

There are 13 hours of tutorials provided in support of this course. You will be notified of the dates, times and location of these tutorials, together with the names 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 may encounter as this will be of help 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 may be circumstances in which you would find help necessary - when:

- You do not understand any part of the study units or the assigned readings.
- You have difficulty with the self-assessment with your tutor's comment on an assignment or with the grading of an assignment. You should try your best to attend tutorials. This is the only chance to have face-to-face contact with your tutor and to ask question which are course of your study. To gain maximum benefit from course tutorials, prepare your list of questions ahead of time. You will learn a lot from participating in the discussions.

Conclusion

The course, Applied Econometrics I (ECO453) exposes you to time series data utilization, model estimation and the issues involved in Applied Economics, such as

application of simple and multiple regression to solve economics policy and theory, Nonlinear Regression Models, Panel Data Regression Models, Dynamic Econometric Models: Autoregressive and Distributed-Lag Models and their applications in the Economy. On the successful completion of the course, you would have been armed with the materials necessary for efficient and effective management of applications of econometric related matters in any organization, policy institution and the country.

TABLE OF CONTENTS

PANEL DATA ESTIMATION

Unit 1: Panel Data Regression Model

Unit 2: Fixed Versus Random Effects Panel Data

Unit 3: Testing Fixed and Random Effects

MODULE 1

INTRODUCTION TO ECONOMETRIC RESEARCH USING SOFTWARE

Unit 1 Meaning of Applied Econometric Research

Unit 2: Simple Linear Regression Model.

Unit 3: How to Run Time Series Data Using EViews Software

Unit 4: Nonlinear Regression

Unit 5: Qualitative Response Regressions

UNIT 1: Meaning of Applied Econometric Research Contents

Unit Structure

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

1.1 Introduction

Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analyzing economic history, and forecasting. In econometric research there are different stages and one stage of the research leads to another, you need to learn these stages and know them in a chronological order. The stages of econometric research will help to give the basic foundational knowledge of what it takes to carry out applied research. In this unit you will be exposed to the basic assumptions with respect to the independent variables to be estimated.

1.2 Learning Outcomes

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

- Discuss meaning of Econometrics
- Analyze the basic tools for econometrics
- State methodology of econometrics
- Discuss the basic stages of econometric research
- Explain the assumptions of econometric models
- Make a critique of econometrics research

1.3 Meaning of Econometrics

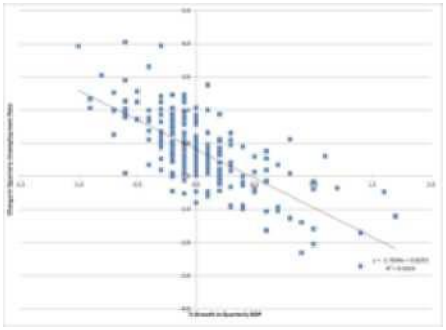
Econometrics is the application of statistical methods to economic data and is described as the branch of economics that aims to give empirical content to economic relations. More precisely, it is "the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference". An introductory economics textbook describes econometrics as allowing economists "to sift through mountains of data to extract simple relationships". The first known use of the term "econometrics" (in cognate form) was by Polish economist Pawel Ciompa in 1910. Jan Tinbergen is considered by many to be one of the founding fathers of econometrics. Ragnar Frisch is credited with coining the term in the sense in which it is used today.

The basic tool for econometrics is the multiple linear regression model. Econometric theory uses statistical theory and mathematical statistics to evaluate and develop econometric methods. Econometricians try to find estimators that have desirable statistical properties including unbiasedness, efficiency, and consistency. Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analyzing economic history, and forecasting.

1.4 The Basic Tool for Econometrics

The basic tool for econometrics is the multiple linear regression model. In modern econometrics, other statistical tools are frequently used, but linear regression is still the most frequently used starting point for an analysis. Estimating a linear regression on two

variables can be visualised as fitting a line through data points representing paired values of the independent and dependent variables.



Okun's law representing the relationship between GDP growth and the unemployment rate. The fitted line is found using regression analysis.

The basic tool for econometrics is the multiple linear regression model. In modern econometrics, other statistical tools are frequently used, but linear regression is still the most frequently used starting point for an analysis. Estimating a linear regression on two variables can be visualised as fitting a line through data points representing paired values of the independent and dependent variables.

For example, consider Okun's law, which relates GDP growth to the unemployment rate. This relationship is represented in a linear regression where the change in unemployment rate (Unemployment) is a function of an intercept (b0) a given value of GDP growth multiplied by a slope coefficient bi and an error term, U:

Unemployment = b0 + b1 Growth + U.

The unknown parameters b0 and b1 can be estimated. Here p1 is estimated to be -1.77 and b0 is estimated to be 0.83. This means that if GDP growth increased by one percentage point, the unemployment rate would be predicted to drop by 1.77 points. The model could then be tested for statistical significance as to whether an increase in growth is associated with a decrease in the unemployment, as hypothesized. If the estimate of b1 were not significantly different from 0, the test would fail to find evidence that changes in the growth rate and unemployment rate were related. The variance in a prediction of the dependent variable (unemployment) as a function of the independent variable (GDP growth) is given in polynomial least squares.

1.5 Methodology of Econometrics

Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting. Econometrics may use standard statistical models to study economic questions, but most often they are with observational data, rather than in controlled experiments. In this, the design of observational studies in econometrics is similar to the design of studies in other observational disciplines, such as astronomy, epidemiology, sociology and political science. Analysis of data from an observational study is guided by the study protocol, although exploratory data analysis may be useful for generating new hypotheses. Economics often analyses systems of equations and inequalities, such as supply and demand hypothesized to be in equilibrium. Consequently, the field of econometrics has developed methods for identification and estimation of simultaneous-

equation models. These methods are analogous to methods used in other areas of science, such as the field of system identification in systems analysis and control theory. Such methods may allow researchers to estimate models and investigate their empirical consequences, without directly manipulating the system.

One of the fundamental statistical methods used by econometricians is regression analysis. Regression methods are important in econometrics because economists typically cannot use controlled experiments. Econometricians often seek illuminating natural experiments in the absence of evidence from controlled experiments. Observational data may be subject to omitted-variable bias and a list of other problems that must be addressed using causal analysis of simultaneous-equation models.

1.5.1 Stages of Applied Econometric Research

In econometric research there are four main stages. These stages follow a chronological order, a good knowledge of economic theory will assist in identifying the econometric research structure and the characteristics associated with it, some of these basic stages of applied econometric research include:

(i) Model Formulation: 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 a prior 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 such that the model is theoretically meaningful and mathematically

useful. Model specification or formulation therefore, presupposes knowledge of economic theory and the familiarity with the particular phenomenon under investigation, the theoretical knowledge allows the researcher to have an idea of the interdependence of the variables under study.

This stage involves choosing a suitable econometric model to test the hypotheses. Model specification depends on the nature of the research question and the data. In econometrics, the chosen model often takes the form of a regression equation, where a dependent variable is expressed as a function of one or more independent variables. The model should capture all relevant variables and account for potential interaction effects and non-linear relationships.

(ii) Model Estimation: Model estimation 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.

• Data collection based on the variables included in the model.

• Examining the identification conditions of the model to ensure that the function that is being estimated is the real function in question.

• Examining aggregation problems of the function to avoid biased estimates.

• Ensuring that the explanatory variables are not collinear, the situation which always results in misleading results.

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

(iii) Evaluation of Model 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 the econometric research. Once the model is estimated, the researcher conducts hypothesis tests to determine whether the results support the initial hypotheses. These tests often involve examining the statistical significance of the estimated coefficients using t-tests, F-tests, or other appropriate tests. After the final model is estimated and the hypotheses are tested, the researcher interprets the results. This involves making inferences about the relationships between the variables and drawing conclusions about the economic phenomena under investigation. The interpretation should be consistent with the data and the underlying economic theory.

(iv) Testing 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 on the stability of the estimates in term of their sensitivity to changes in the size of the model even outside the given sample data within the period.

Communication of Findings: The final stage of econometric research is to communicate the findings. This usually involves writing a research report or academic paper that presents the research question, the econometric model, the data, the estimation results, the hypothesis tests, and the interpretation of the findings. The researcher should also discuss the implications of the findings for economic theory, policy, or future research.

Self-Assessment Exercise 1

Itemize the basic stages of an econometric research

1.5.2 Concepts of Economic and Econometric Model

A model is a simplified representation of a real-world process. That is, it is a prototype of reality, and so describes the way in which variables are interrelated. 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 the various relationships between the given variables In examining these concepts Bergstron (1966) defined model as any set of assumption and relationships which approximately describe the behaviour of an economy or a sector of an economy. In this way, an economic model guides economic analysis. 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 economic model in the following ways:

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

ii. In formulating 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 random disturbance term. This will enable random element that are not accounted for to be taken care of in the sample.

1.5.3 Properties of a Good Econometric Model

The "goodness" of an econometric model is judged on the basis of some basic fundamental properties that are universal in nature the following are some of these 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 co-efficient should be accurate. They should approximate as best as possible the true parameters 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. Example a model should explain price, demand, supply exchange rates, market behaviour, and any other practical situation.

iv. Prediction. The model should be able to correctly predict future values of the dependent variable. Example a model should predict with accuracy price, demand, supply exchange rates, market behaviour, and any other practical situation. This feature often help strengthen the validity of an econometric model within a given period.

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

vi. Identification. The equations of the model should be easily identified that is, it must have a unique mathematical form. This means the model should either be exactly identified, over identified or under identified.

Self-Assessment Exercise 2

Discuss the properties of a good econometric model

1.5.4 Limitations and criticisms of Econometrics Research

Like other forms of statistical analysis, badly specified econometric models may show a spurious relationship where two variables are correlated but causally unrelated. In a study of the use of econometrics in major economics journals, McCloskey concluded that some economists report p-values (following the Fisherian tradition of tests of significance of point null-hypotheses) and neglect concerns of type II errors; some economists fail to report estimates of the size of effects (apart from statistical significance) and to discuss their economic importance. She also argues that some economists also fail to use economic reasoning for model selection, especially for deciding which variables to include in a regression.

In some cases, economic variables cannot be experimentally manipulated as treatments randomly assigned to subjects. In such cases, economists rely on observational studies,

often using data sets with many strongly associated covariates, resulting in enormous numbers of models with similar explanatory ability but different covariates and regression estimates. Regarding the plurality of models compatible with observational data-sets, Edward Leamer urged that "professionals ... properly withhold belief until an inference can be shown to be adequately insensitive to the choice of assumptions".

1.6 Summary

In this unit which is the first unit in Module 1 of this course, you learnt the meaning and basic stages of applied econometric research. These stages of econometric model research are necessary for effective analysis of any model building. You need to thoroughly master these steps because their violation leads to several econometric problems that we shall study in this course.

Econometric theory uses statistical theory and mathematical statistics to evaluate and develop econometric methods. Econometricians try to find estimators that have desirable statistical properties including unbiasedness, efficiency, and consistency. An estimator is unbiased if its expected value is the true value of the parameter; it is consistent if it converges to the true value as sample size gets larger, and it is efficient if the estimator has lower standard error than other unbiased estimators for a given sample size. Ordinary least squares (OLS) is often used for estimation since it provides the BLUE or "best linear unbiased estimator" (where "best" means most efficient, unbiased estimator) given the Gauss-Markov assumptions. When these assumptions are violated or other statistical properties are desired, other estimation techniques such as maximum likelihood estimation, generalised method of moments, or generalised least squares are used.

Estimators that incorporate prior beliefs are advocated by those who favour Bayesian statistics over traditional, classical or "frequentist" approaches

1.7 References/ Further Reading

Asteriou D and Hall S.G. (2007) App*lied Econometrics: A Modern Approach using Eviews and Microfit.* Macmillan pal Grave New York.

Berndt, Ernst R.(1991) *The practice of Econometrics: Classic and contemporary*, Addison-Wesley,

Goldberger, Arhur S. (1998) *Introductory Econometrics*, Harvard University Press.

Gujarati, D.N. (2003. *Basic Econometrics.* Tata Mc-Graw - Hill Publishing Company Ltd New-Delhi.

Koutsoyiannis, A. (1977) *Theory of Econometrics An Introductory Exposition Econometric.* Methods Macmillan

Oosterbaan, R.J. (1994), *Frequency and Regression Analysis*. In: H.P.Ritzema (ed.), Drainage Principles and Applications, Publ. 16, pp. 175-224, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. ISBN 90-70754-33-9 . Download as PDF

Oosterbaan, R.J. (2002). *Drainage research in farmers' fields*: analysis of data. Part of project “Liquid Gold” of the International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. Download as PDF

Wooldridge, J. M. (2009) *Introductory Econometrics A modern Approach*, Cengage Learning Singapore 4th Edition.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Econometric research involves using statistical methods to estimate economic models and test economic theories. Here are the basic stages of an econometric research:

1. **Formulation of a Theoretical Model**: This is the initial phase where the researcher develops a theoretical framework for the research. The framework should be able to capture the key variables and the relationships among them.
2. **Specification of the Econometric Model**: The theoretical model is then translated into an econometric model. This involves specifying the functional form of the relationship between variables and deciding which variables are endogenous (dependent) and which are exogenous (independent).

3. **Data Collection**: Once the model is specified, the next step is to collect data. This could involve the use of surveys, experiments, or secondary data sources. The type and quality of the data collected will significantly influence the reliability of the econometric analysis.
4. **Estimation of the Model**: Once the data has been collected, it is used to estimate the parameters of the model. This usually involves some form of regression analysis.
5. **Model Diagnostics**: The estimated model is tested for statistical validity. This could include tests for autocorrelation, multicollinearity, heteroscedasticity, model specification errors, and more.
6. **Hypothesis Testing**: The researcher uses statistical tests to determine whether the coefficients on the variables in the model are statistically significant and therefore provide support for the theoretical model.
7. **Model Refinement**: Based on the results of the diagnostic tests and hypothesis testing, the model may need to be refined. This could involve adding or removing variables, changing the functional form, or incorporating lagged variables.
8. **Interpretation of Results and Policy Implication**: The final step is to interpret the results and to draw conclusions about the relationships between the variables in the model. The researcher can also discuss the implications of the results for economic policy or business decisions.
9. **Reporting the Findings**: The findings are then reported in the form of a research paper or report, which includes a detailed description of the methodology used, the data collected, the results obtained, and the conclusions drawn.

Answer to Self- Assessment 2

A good econometric model possesses several important properties. Below are some of these key characteristics:

1. **Theoretical Consistency**: The model should be based on sound economic theory. It should represent logical and plausible relationships between the variables.
2. **Statistical Adequacy**: The model should satisfy the basic assumptions of the statistical techniques used for estimation. For example, in a linear regression model, assumptions include linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors.
3. **Identifiability**: The model should be identifiable, meaning the parameters of the model can be uniquely estimated based on the observed data. In other words, there should not be an infinite number of parameter values that can produce the same probability distribution for the observed data.
4. **Estimability**: Given the available data, it should be possible to estimate the parameters of the model.

5. **Simplicity**: The model should be as simple as possible, given the research question it is intended to answer. This is often referred to as the principle of parsimony or Occam's Razor.
6. **Goodness of Fit**: The model should provide a good fit to the observed data. This is often assessed using measures such as the R-squared in a linear regression model.
7. **Predictive Accuracy**: A good econometric model should be able to accurately predict out-of-sample observations.
8. **Robustness**: A robust model gives similar results with minor changes in model specification or the sample of data used.
9. **Interpretability**: The model should be easily interpretable. This means the coefficients of the model parameters should have clear meanings.
10. **Validity**: The model should pass diagnostic tests for things like autocorrelation, multicollinearity, heteroscedasticity, and model specification errors.
11. **Flexibility**: The model should be flexible enough to adapt to new data or changes in the underlying economic relationships.

UNIT TWO: TIME SERIES AND ITS COMPONENTS

CONTENTS

1.1 Introduction

In all the social sciences, and particularly economics and business, the problem of how condition changes with the passage of time is of utmost importance. For study of such problems, the appropriate kind of statistical information consist of data in the form of time series, figures which shows the magnitude of a phenomenon month after month or year after year. The proper methods for treating such data and thus summarizing the experience which they represent are indispensable part of the practicing statistician equipment.

1.2 Learning Outcomes

At the of this unit, you should be able to:

- Understand or define time series
- Understand component part of time series
- Understand methods of estimating time series
- Estimation and graphical representation of the trend

1.3 Time Series Defined

A time series consists of numerical data collected, observed or recorded at more or less regular intervals of time each hour, day, month, quarter or year. This might be the collection of observations of some economic or physical phenomenon drawn at discrete points in time, equally spaced. You might wonder why we should spend so much effort constructing series showing what has happened in the past. This is history and should we not rather be looking to the future? As you know the twentieth century is age of planning: government plans the economy for many years ahead; public corporation plan output and investment; most state plan to keep the rate of inflation down to an acceptable level.

Good planning is usually based on information and this is where the time series comes into its own. It provides information about the way in which economic and social variable have been behaving in the recent past, and provides an analysis of that behaviour that planner cannot ignore. Naturally, if we are looking into the future, there is certain assumption we have to make, the most important of which is that the behavioural pattern that we have found in the past could continue into the future. In looking to the future there are certain pattern that we assume will continue and it is to help in the determination of these pattern that we undertake the analysis of the time series.

Time series is usually ordered in time or space. Time series is denoted by sequence (Yt) where Yt is the observed value at time t.

Essentially, time series is usually applied to economic and business problems whose purpose of analyses data is to permit a forecast to the future both in the long term and short term. It may be used as an essential aid to planning. Example of time series data are volume of sales, the character and magnitude of its cost of production etc. population figure, price level, demand of a commodity.

Self-Assessment Exercise 1

The essence of time series is forecast. (true/false)

1.4 Components of Time Series

The nature or variation or type of changes in times series can be categorise into:

- Secular trend or long-term movement

- Seasonal variation

- Cyclical variation

- Irregular or residual variation

1.4.1 Secular Trend

This refers to the general direction in which the graph of time series appears to be going over a long period of time. This explains the growth or decline of a time series over a long period. Time series is said to contain a trend if the mean or average of series changes

systematically with time. The trend could be upward or downward, this could take any of the shape below.

Graphical Representation of Secular Trends

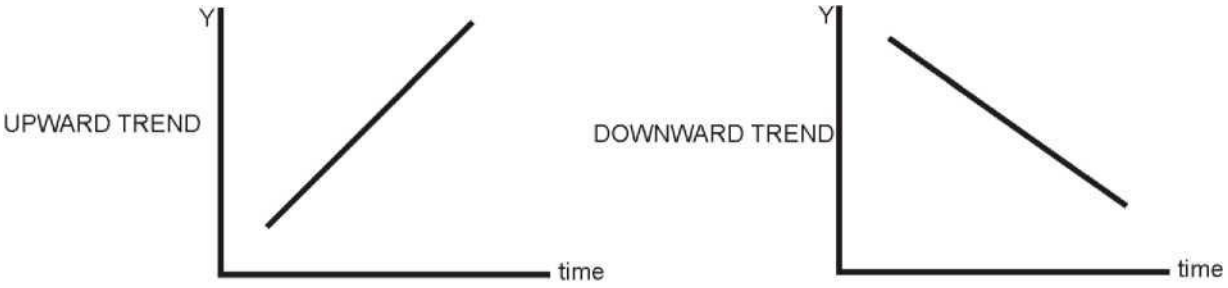


Fig. M1.2.1

1.4.2 Seasonal Variation

This refers to short term fluctuation or changes that occur at regular intervals less than a year. It is usually brought about by climatic and social factor(s), it is usually because of an event occurring at a particular period of the year. Examples of these are sale of card during valentine period, sale of chicken during xmas, new year or any festive period(s).

Graphical Representation Of Seasonal Variation

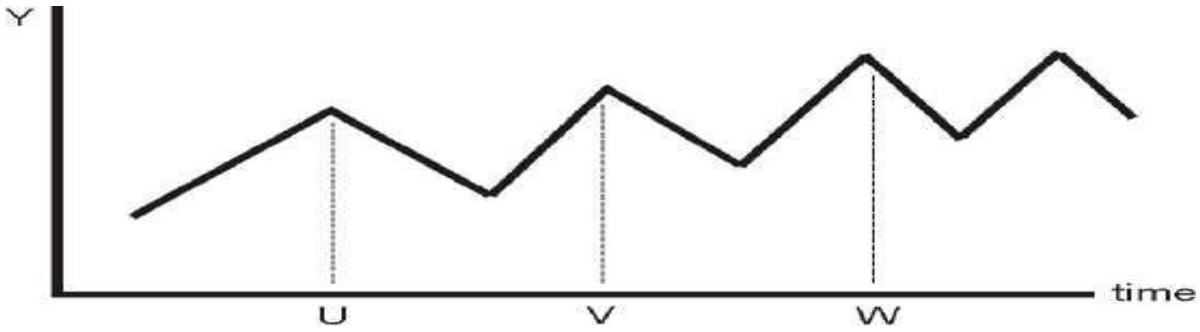


Fig. M1.2.2

1.4.3 Cyclical Variation

This refers to long term variations about the trend usually caused by disruption in services or socio-economic activities, cyclical variations are commonly associated with economic cycles, successive boom and slumps in the economy. A good example of this is business cycle.

Graphical Representation Of Cyclical Variation

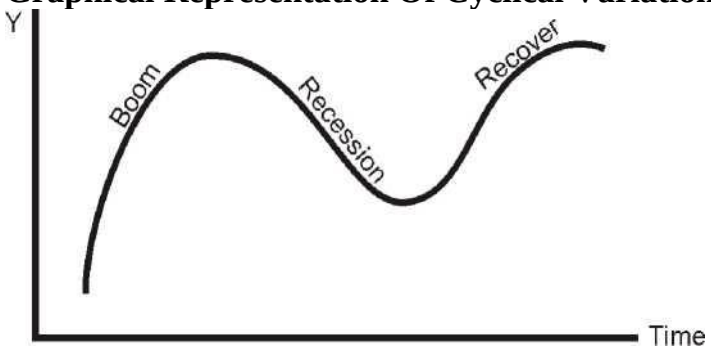


Fig. M1.2.3

1.4.4 Irregular Variation

This refers to time series movement that are not definite this is usually caused by unusual or unexpected and unpredictable events such as strike, war, flood, disasters. Here, there's no definite behavioural pattern.

Graphical Representation Of Irregular Variation

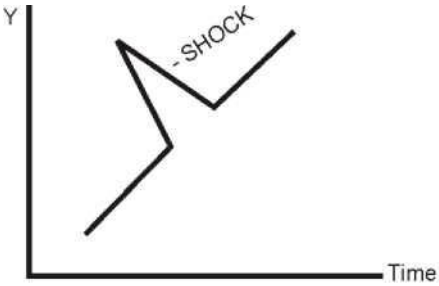


Fig. M1.2.4

Self-Assessment Exercise 2

The trend of secular trend can either be upward or downward. (true/false)

1.5 Measurement of Trend

rends in econometric data represent the underlying patterns or directions that the data seems to follow over a period of time. Detecting these trends is essential as they provide insights into the future, help formulate economic policies, and offer crucial information for investment and business decisions.

There are multiple ways to measure trends in econometrics, each varying according to the nature and requirements of the dataset and the underlying economic questions. Basically, trend values of a time series can be estimated by any of the following methods:

- Free hand

- Regression

- Moving average

1.5.1 Free Hand Method

This method involves the drawing a scattered diagram of the values with time as the independent variable on the x-axis and then drawing the trend line by eye. This method is condemned because it is subjective and inaccurate method of obtaining a Trend line.

Graphical Representation of Free Hand Method

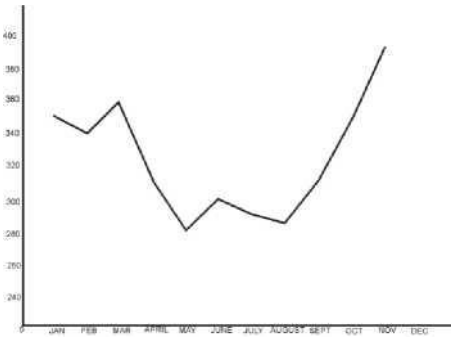


Fig. M1.2.5

1.5.2 Regression

Regression analysis is a powerful tool used extensively in econometrics. It's a statistical process that estimates the relationship among variables. The main use of regression analysis in econometrics is to understand how the dependent variable changes when one or more independent variables vary. When used in trend analysis, regression analysis allows us to quantify an underlying trend in a data set over a period of time.

Linear regression is the most commonly employed method of regression analysis in trend measurement. In the context of time series data (where observations are recorded at regular time intervals), the independent variable is time, and the dependent variable is the phenomenon being observed. For instance, the variable of interest could be GDP, stock prices, or unemployment rate, among others.

The basic form of a linear regression model for trend analysis can be represented as

$Y_t = \alpha + \beta t + \varepsilon_t$, where:

- Y_t is the dependent variable at time t
- α is the y-intercept or the value of Y_t when time t is zero (a constant term)
- β is the slope coefficient, representing the rate of change in Y_t for a one-unit change in time
- t is the time period
- ε_t is the error term, capturing all other factors affecting Y_t that are not included in the model

The slope coefficient β is of special interest in trend analysis. If β is positive, it suggests an upward trend in the dependent variable over time. Conversely, if β is negative, it indicates a downward trend. The magnitude of β tells us about the rate of change in the dependent variable per unit of time.

Using statistical software, we can estimate the parameters α and β from our data. These estimated parameters allow us to model the trend line, understand the direction and rate of trend, and make future predictions.

However, it is important to note that a linear regression model assumes a linear trend, which might not always be the case in real-world data. Non-linear trends might be better fitted with polynomial or logarithmic regression models. Moreover, regression analysis

assumes a constant trend over time, which may not hold true in the presence of structural breaks or changes in trend patterns.

1.5.3 Moving Average

Moving averages is another widely used method for trend measurement in econometric analysis. This technique is essentially a method of smoothing out data to identify the underlying trend, and it is particularly beneficial when dealing with time-series data.

In a moving average, for a given point in time, you calculate the average of the data points around it. The "window" of data points you use for this calculation can vary depending on the nature of the data and the trend you're trying to capture. For instance, if you are using a 12-month moving average on monthly data, you'd calculate the average of the current month and the 11 preceding months. This process is then repeated for each month in the series, hence the term "moving" average.

The moving average technique helps to smooth out short-term fluctuations and seasonality to show an underlying trend in the data. It can help to reveal if there's an overall upward or downward pattern in the data over time.

One advantage of moving averages is their simplicity and easy interpretability. They can provide a quick and intuitive understanding of the trend in a dataset. They are also useful for forecasting, as the moving average at the end of a time series can be used as a forecast for the next period.

However, moving averages also have some limitations. First, they can lag behind trends, because they're based on past data. This lag becomes more pronounced with larger window sizes. Secondly, they assume that recent historical data is a good predictor of the future, which may not always be true. Lastly, moving averages may not be effective at identifying sudden shifts or turning points in trends since they are designed to smooth out these fluctuations.

Despite these limitations, moving averages remain a valuable tool in econometrics for trend measurement, especially in conjunction with other methods. By understanding the strengths and weaknesses of moving averages, you can use them effectively to draw insights from economic data.

1.6 Summary

In the course of our discussion on time series analysis you have learnt about: Time series Time series data, Component of time series, and trend measurement. The measurement of trends is a vital component of econometric analysis. By identifying patterns in economic data, economists can forecast future trends, shape policy decisions, and contribute to the understanding of economic phenomena. While methods such as graphical analysis, regression analysis, and moving averages each have their unique advantages and limitations, their combined usage can often provide a comprehensive understanding of trends in the data. Ultimately, the appropriate method of trend measurement depends on the specific nature of the data and the economic questions at hand.

1.7 References/Further Readings

Adedayo, O.A. (2006): Understanding Statistics. JAS Publishers, Akoka, Lagos.
Dawodu, A.F. (2008): Modern business Statistics 1. NICHO Printing Works, Agbor, Delta State.
Esan, E.O. and Okafor, R.O. (2010): Basis Statistical Method. Tony Chriisto Concept, Lagos.
Olufolabo, O.O. & Talabi, C.O. (2002): Principles and Practice of Statistics HAS-FEM (NIG) ENTERPRISES Somolu Lagos.
Owen, F. and Jones, R. (1978): Statistics. Polytech Publishers Ltd. Stockport.
Oyesiku, O.K. and Omitogun, O.(1999): Statistics for social and Management Sciences. Higher Education Books Publisher, Lagos.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

False. While forecasting is a significant aspect of time series analysis, it's not the sole essence of it.

Time series analysis involves the study of data points collected over time to identify patterns, trends, cycles, and other characteristics. It has many applications beyond forecasting.

Answer to Self- Assessment 2

True. A secular trend, which refers to a long-term trend in time series data, can indeed be either upward or downward.

An upward secular trend indicates a consistent increase in the variable over time. This is commonly seen in time series data such as population growth, long-term economic growth, or the increasing trend in global temperatures due to climate change.

A downward secular trend, on the other hand, indicates a consistent decrease in the variable over time. This might be observed in data such as the rate of smoking over the past several decades, certain commodity prices subject to technological advancements, or the fertility rates in many developed countries.

Therefore, the direction of a secular trend (upward or downward) is determined by the nature of the variable and the time period being studied.

UNIT 3: SIMPLE LINEAR REGRESSION MODEL

1.1 Introduction

In the preceding unit you learnt the meaning of econometrics research approach. That laid the foundation for the present unit in which you will learn linear regression model. Linear model shows the relationship between two variables. In this relationship, one variable is depending on the other variable. The model consists of the independent variables and the constant term, with their respective coefficient and we need to estimate the parameters of

the model in order to know the magnitude of their relationship. Consider a familiar supply function of the form:

$Y = bo + b1x$ (1)

This function shows the positive linear relationship between quantity supply, Y and price of the commodity, X. The dependent variable in this model is the quantity supply, denoted by Y, while the independent variable (explanatory variable) is the price, X. This is a two-variable case with two parameters representing the intercept and the slope of the function. This supply-price relationship, $Y = f(x)$ is a one-way causation between the variables Y and X: price is the cause of changes in the quantity supply, but not the other way round. From the above equation (1), the parameters are bo and b1, and we need to obtain numerical value of these parameters. The left-hand variable Y is variously referred to as the endogenous variables, the regressand, the dependent variable or the explained variable. Similarly, the right-hand variable X is variously described as exogenous variables, the regressor, the independent variable or the explanatory variable.

1.2 Learning Outcomes

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

- Explain linear regression approach
- Discuss parameter estimation procedures.
- Evaluate the assumptions of the stochastic variable.
- Analyze the assumptions of the explanatory variables.
- Attempt algebraic and software estimation of simple regression.

1.3 Linear Regression Approach

Linear regression is a special case of regression analysis, which tries to explain the relationship between a dependent variable and one or more explanatory variables. Mathematical functions are used to predict or estimate the value of the dependent variables. In linear regression, these functions are linear. Linear regression was the first type of regression analysis to be studied rigorously. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters. What is more, the statistical properties of the resulting estimators are easier to determine. Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- Linear regression can be used to fit a predictive model to a set of observed values (data). This is useful, if the goal is prediction, or forecasting, or reduction. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
- Given a variable y and a number of variables Xi, ..., Xp that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the Xj, to assess which Xj has no relationship with y at all, and to identify which subsets of the Xj contain redundant information about y.

Linear regression models are often fitted using the least squares approach. Other ways of fitting exist; they include minimizing the "lack of fit" in some other norm (as with least

absolute deviations regression), or minimizing a penalized version of the least squares loss function as in ridge regression. The least squares approach can also be used to fit models that are not linear. As outlined above, the terms "least squares" and "linear model" are closely linked, but they are not synonymous.

1.4 Parameter Estimation Strategies

The parameters of this model are to be estimated using ordinary least square (OLS) method. We shall employ this method 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 are simple to understand.

iii. This method always produces satisfactory results.

iv. The parameter estimates using the O.L.S. method are best, Linear and unbiased. This makes the estimates to be more accurate compared with the estimates obtained using other methods.

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

Note that the model Y = bo + b1x 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, the pairs of observation (Y and X) would all lie on a straight line. 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.

1.4.1 Sources of Deviations in Parameters and Models

There are deviations of observations from the line. These deviations are attributable to the following factors:

- Omission of variable(s) from the function on ground that some of these variables may not be known to be relevant
- Random behaviour of human beings. Human reactions at times are unpredictable and may cause deviation from the normal behavioral pattern depicted by the line.
- Imperfect specification of the mathematical form of the model. A linear model, for instance, may mistakenly be formulated as a non-linear model. It is also possible that some equations might have been left out in the model
- Error of aggregation - usually, in model specification, we use aggregate data in which we add magnitudes relating to individuals whose behavior differs. The additions and approximations could lead to the existence of errors in econometric models
- Error of measurement - this error arises in the course of data collection, especially in the methods used in the collection of data. Data on the same subject collected from central bank of Nigeria and National Bureau of statistics could vary in magnitude and units of measurements. Therefore when you use different sources you could get different results.

Self-Assessment Exercise 1

Outline sources of deviations in parameters estimation.

1.4.2 The uses of Random Variable in Models

The inclusion of a random variable usually denoted by U, into the econometric function help in overcoming the above stated sources of errors. The U's is variously termed the error term, the random disturbance term, or the stochastic term.

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

That is $Y = bo + b1x + ui$ (ii)

Where Y = variation in Y; bo + b1x = systematic variation, Ui = random variation. Simply put, variation in Y = explained variation plus unexplained variation. Thus,

Y = bo + b1x + Ui is the true relationship that connects the variable Y and X and this is our regression model 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 variables, thus, the following assumptions hold:

1.5 Assumptions of Linear Stochastic Regression Model

1.5.1 Assumptions with respect to the random variable

In respect to the random variable U, the following assumptions apply to any given econometric model that is used in prediction of any economic phenomenon:

(i) Ui is a random variable, this means that the value which Ui takes in any one period depends on chance. Such values may be positive, negative or Zero. For this

assumption to hold, the omitted variables should be numerous and should change in different directions.

(ii) The mean value of "U" in any particular period is zero. That is, E (ui) denoted by U is zero. By this assumption, we may express our regression in equation (ii) above as Yi = bo + b1x.

(iii) The variance of ui is constant in each period. That is, Var (ui) = E $(ui)^2$ = $\delta(ui)2$ = $\delta 2u$ which is constant. This implies that for all values of x, the U's will show the same dispersion about their mean. Violation of this assumption makes the Us heteroscedastic.

(iv) U has a normal distribution. That is, a bell shaped symmetrical distribution about their zero mean. Thus, U = N(0, 1).

(v) The covariance of ui and uj = 0. $i \neq j$. This assumes the absence of autocorrelation among the ui. In this respect, the value of u in one period is not related to its value in another period.

Self - Assessment Exercise 2

Discuss the assumptions with respect to the random variable

1.5.2 Assumption of error term with respect to the explanatory Variable

The following assumptions also hold when you conduct a regression analysis in terms of the relationship between the explanatory variable and the stochastic variable:

i. U and X do not covary. This means that there is no correlation between the

disturbance term and the explanatory variable. Therefore, cov. Xu = 0.

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

1.5.3 Assumption in Relation to the Explanatory Variable.

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 that cov. Xi Xj = 0. $i \neq j$.

(This assumption applies to multiple regression model).

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

(iii) The coefficients of the relationships to be estimated are assumed to have

a unique mathematical form. That is, the variables are easily identified.

(iv) The relationships to be estimated are correctly specified.

1.5.4 Numerical Estimation of Parameters

The following procedures are used in finding numerical values of the parameters bo and bi.

From the true relationship Yi = bo + bi x + u, (i)

and the estimated relationship Y = bo + b1 x + ei,

the residual $e_i = Y_i - \hat{Y}$ …….(ii)

Squaring the residual and summing over n, gives:

$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (Y - Y)^2 = \sum_{i=1}^{n} (Y - \hat{\beta}_0 - \hat{\beta}_1 X)^2 \quad \text{.....(iii)}$$

The expression in (iii) is to be minimized with respect to bo and b1 respectively.

Partial derivative with respect to

$$\frac{\partial \sum e^2}{\partial \hat{\beta}_0} = \frac{\partial \sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X)^2}{\partial \hat{\beta}_0} = 0$$

$$2\sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X) \cdot (-1) = 0$$

$$\sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X) = 0 \quad \text{.....(iv)}$$

Partial derivative with respect to

$$\frac{\partial \sum e^2}{\partial \hat{\beta}_1} = \frac{\partial \sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X)^2}{\partial \hat{\beta}_1} = 0$$

$$2\sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X) \cdot (-X) = 0$$

$$\sum (XY - \hat{\beta}_0 X - \hat{\beta}_1 X^2) = 0 \text{.....(v)}$$

Combining both (iv) and (v), we have

$$\sum Y - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 X = 0$$

$$\sum XY - \sum \hat{\beta}_0 X - \sum \hat{\beta}_1 X^2 = 0$$

Rewriting, we have:

$$\sum Y = \hat{\beta}_0 n + \hat{\beta}_1 \sum X \text{.....(vi)}$$

$$\sum XY = \hat{\beta}_0 \sum X + \hat{\beta}_1 \sum X^2 \text{.....(vii)}$$

The two equations (vi) and (vii) are the normal equation of the regression model.

Using Crammer's rule, the values of the parameter's bo and b1 are respectively:

$$\hat{\beta}_0 = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

$$\hat{\beta}_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

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

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

$$\hat{\beta}_1 = \frac{\sum xy}{\sum x^2}$$

Where:

$\overline{Y} = \frac{\sum Y}{n}$, $\overline{X} = \frac{\sum X}{n}$, $x = X - \overline{X}$ and $y = Y - \overline{Y}$

A study is conducted involving 10 observations to investigate the relationship and effects of price on quantity demanded

Table M1.3.1: Relationship between Price and Quantity demanded

Observation	Price	Quantity demanded
	X	Y
1	35	112
2	40	128
3	38	130
4	44	138
5	67	158
6	64	162
7	59	140
8	69	175
9	25	125
10	50	142
Total	**491**	**1410**

Table M1.3.2: Calculating the Linear Regression between Price and Quantity demanded

Observation	Price	Quantity demanded		
	X	Y	XY	X^2
1	35	112	3920	1225
2	40	128	5120	1600
3	38	130	4940	1444
4	44	138	6072	1936
5	67	158	10586	4489
6	64	162	10368	4096
7	59	140	8260	3481
8	69	175	12075	4761
9	25	125	3125	625
10	50	142	7100	2500
Total	**491**	**1410**	**71566**	**26157**

$$\bar{X} = \frac{\Sigma X}{n} = \frac{491}{10} = 49.1; \bar{Y} = \frac{\Sigma Y}{n} = \frac{1410}{10} = 141$$

$$\beta_1 = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$\beta_1 = \frac{(10 * 71566) - (491 * 1410)}{10 * 26157 - (491)^2}$$

$$\beta_1 = \frac{715660 - 692310}{261570 - 241081}$$

$$\beta_1 = \frac{23350}{20489} = 1.140$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\beta_0 = 141 - (1.140 * 49.1)$$

$\beta_0 = 141 - 55.974$
$\beta_0 = 85$

1.5.5 EViews Software Applications

The linear regression can be solved using the software package using the following steps. Create an excel worksheet type the data for quantity (Y)) and price (X). Open your Eviews software, go to file, create worksheet, copy the data on excel worksheet paste it on the Eviews worksheet already created go to file, import data. Go to 'Quick' on the tool bar scroll to estimate equation and click on it a dialogue box opens, type the respective quantity and price click ok, the output is as follows:

Table M1.3.3: Regression Result

Dependent Variable: Y
Method: Least Squares
Date: 08/04/22 Time: 13:18
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	1.139636	0.194949	5.845812	0.0004
C	85.04388	9.970461	8.529583	0.0000

R-squared	0.810308	Mean dependent var	141.0000
Adjusted R-squared	0.786596	S.D. dependent var	19.10207
S.E. of regression	8.824328	Akaike info criterion	7.369759
Sum squared resid	622.9502	Schwarz criterion	7.430276
Log likelihood	-34.84879	Hannan-Quinn criter.	7.303371
F-statistic	34.17352	Durbin-Watson stat	1.885155
Prob(F-statistic)	0.000385		

To copy the estimated result from EViews to words format, highlight the output (result), click copy and select HTML from the drop-down menu. The software estimated model

for the data can be presented in a line form as follows: Y = 85 + 1.14X, Y represent the quantity while X represent the price. This similar to the one computed using the manual method in the regression.

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 b0 and b1, the estimated regression lines or equation is got by substituting these values into Y = bo + b1X and this gives Y = 85 + 1.14X.

Thus, given the values of x1 (1 = 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 regression equation. Y = b0 + b1xi

This is also the equation of the line with intercept b0 and slope b1. Note that

Therefore, price elasticity

Taking the mean of $\bar{X} = \frac{491}{10} = 49.1$ and $\bar{Y} = \frac{1410}{10} = 141$, we have average elasticity,

$ep = b1.\frac{\bar{X}}{\bar{Y}}$. Therefore $ep = 1.14 \times \frac{49.1}{141} = 1.14 \times 0.348 = 0.397$

1.6 Summary

In econometrics, linear regression is a linear approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple

linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

In this unit which is the second in our Module 1 you learnt the practical applications of econometrics as well as basic stages of applied econometric research, the assumptions of econometric model estimation. Of course, you need to learn these assumptions well because their violation leads to several econometric problems that we shall study in this course such as simultaneous equation bias. The next unit showcases how to run time series data in Eviews software.

1.7 References/Further Reading

Asteriou D and Hall S.G. (2007) Applied Econometrics: A Modern Approach using Eviews and Microfit. Macmillan pal Grave New York.

Bemdt, Ernst R. (1991) The practice of Econometrics: Classic and contemporary, Addison Wesley,

Goldberger, Arhur S. (1998) Introductory Econometrics, Harvard University Press.

Gujarati, D.N. (2003. Basic Econometrics. Tata Mc-Graw - Hill Publishing Company Ltd New-Delhi.

Koutsoyiannis, A. (1977) Theory of Econometrics An Introductory Exposition Econometric Methods Macmillan

Oosterbaan, R.J. (1994), Frequency and Regression Analysis. In: H.P.Ritzema (ed.), Drainage Principles and Applications, Publ. 16, pp. 175-224, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. ISBN 90-70754-33-9. Download as PDF

Oosterbaan, R.J. (2002). Drainage research in farmers' fields: analysis of data. Part of project “Liquid Gold” of the International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. Download as PDF

Wooldridge, J. M. (2009) Introductory Econometrics A modern Approach, Cengage Learning Singapore 4th Edition.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Estimating parameters in econometric models can be subject to various sources of deviation. These are primarily due to issues in model specification, data quality, and the statistical techniques used. Here are some key sources of deviations in parameter estimation:

1. **Sampling Error**: This occurs because the model parameters are estimated based on a sample rather than the entire population. Estimates can therefore vary from sample to sample.
2. **Measurement Error**: If the data used to estimate the parameters are measured inaccurately, the estimates of the parameters may be biased. This is particularly problematic when the error in the measurement is correlated with the true value of the variable.
3. **Model Misspecification**: If the true relationship between the variables is not correctly specified in the model, this can lead to biased parameter estimates. Examples of misspecification include omitting important variables, including irrelevant variables, or specifying the wrong functional form.
4. **Multicollinearity**: This occurs when two or more independent variables in a regression model are highly correlated. It can lead to unstable parameter estimates and large standard errors.
5. **Heteroscedasticity**: This refers to the variability of the error term being unequal across different levels of the independent variables. When present, it can lead to inefficient parameter estimates and incorrect inference based on standard errors.
6. **Autocorrelation**: This occurs when the error terms are correlated over time. It can lead to inefficient parameter estimates and incorrect standard errors.
7. **Endogeneity**: This occurs when an explanatory variable is correlated with the error term, often due to omitted variables, measurement error, or simultaneity. It can cause parameter estimates to be biased and inconsistent.
8. **Non-Normality of Errors**: Most statistical techniques assume that the errors are normally distributed. If this assumption is violated, the estimates may still be unbiased and consistent, but inference based on standard errors may be incorrect.
9. **Data Limitations**: This can include issues like small sample size, data collected over insufficient time period, or non-random sampling, all of which can impact the accuracy and reliability of parameter estimates.
10. **Outliers**: Outliers can significantly affect parameter estimates, particularly in small samples or if the outliers represent extreme values.

Addressing these issues often requires a combination of robust statistical techniques, improved data collection, and careful model specification.

Answer to Self- Assessment 2

The assumptions with respect to a random variable depend on the statistical model and method being used. However, there are several common assumptions that are often made:

1. **Independence**: This assumption suggests that the observations of a random variable are independent of each other. In other words, the outcome of any individual observation does not influence the outcomes of other observations. This is a key assumption in many statistical models, including regression models and most types of probability models.
2. **Identically Distributed**: Also known as the "IID assumption", this suggests that every observation comes from the same distribution and that distribution does not change over time. This is a common assumption when working with simple random samples.
3. **Normal Distribution**: In many statistical analyses, it's assumed that the random variable follows a normal distribution. This is a key assumption in parametric statistical tests and methods, such as t-tests, ANOVA, and linear regression.
4. **Homoscedasticity**: This assumption, critical in regression analysis, means that the variability of the random variable is constant across all levels of an independent variable. If this assumption is violated (leading to heteroscedasticity), it can result in inefficient estimates and incorrect standard errors.
5. **No Autocorrelation**: In time series analysis, it's often assumed that the errors associated with a random variable are not autocorrelated. Violation of this assumption can lead to inefficient estimates and misleading statistical tests.
6. **Linearity**: In the context of regression analysis, this assumption means that the relationship between the dependent and independent variables is linear. Non-linearity can often be addressed through transformation of the variables or by using non-linear models.
7. **Expectation and Variance**: It is often assumed that a random variable has a finite expected value (mean) and a finite variance.

These assumptions are critical for ensuring the validity of statistical inferences. If these assumptions are violated, the results may be biased, inconsistent, or inefficient. Therefore, diagnostic tests and checks are often used to test whether these assumptions hold in a given dataset.

UNIT 4: TIME SERIES DATA ANALYSIS

1.1 Introduction

In the preceding unit, you learnt simple linear regression. The stage is now set for you to learn the processes of running time series data in Eviews software. The history of methodological developments in econometrics appears to be broadly classified into Traditional Econometrics and Modern Econometrics in literature. The former often refers to the use of economic theory and the study of contemporaneous relationships to explain relationships among dependent variables. It is concerned with building structural models, understanding of the structure of an economy and making statistical inference. In contrast, Modern Econometrics is based on exploiting the information that can be gotten from a variable that is available through the variable itself. It is concerned with building

efficient models which forecasts the time path of a variable very well. Essentially, the term “modem econometrics” refers to Time Series Analysis.

1.2 Learning Outcomes

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

- Discuss history of Time Series Data Analysis
- Explain the Stochastic Process
- Evaluate Stationary and Nonstationary Variables
- Determine weakly Stationarity and Strict Stationarity
- Demonstrate how to run Time Series Data in Eviews 10.0 or 12.0

1.3 History of Time Series Data Analysis

The analysis of time-series is of particular interest to many groups, such as;

(a) Macroeconomists: studying the behaviour of national and international economies.

(b) Finance economists: analysing the stock market.

(c) Agricultural economists: predicting supplies and demands of agricultural products.

Regression models with time series data often exhibit some special characteristics designed to capture their dynamic nature. For instance, including lagged values of the dependent variable or explanatory variables as regressors, or considering lags in the errors, can be used to model dynamic relationship. Regression of the current value of

series on its past values can be used in forecasting. An important assumption for using time series data in regression analysis is that the series have a property called Stationarity. However, many economic variables are nonstationary and the consequences of nonstationary variables for regression modeling are profound. Therefore, the aim of this section is to examine the various data generating processes, the concept of stationarity and how to graphically examine the properties of a time series.

1.4 Stochastic Process

A random or stochastic process is a collection of random variables ordered in time. A Stochastic random variable could be continuous in time or discrete in time. Most economic data are collected at discrete points in time, for instance, GDP.

A stochastic process is a mathematical concept used to define a sequence of random variables evolving over time. It provides a mathematical framework for dealing with systems that evolve in a way that contains a random element.

Notation, let y denote the random variable at time t.

Here's a brief outline of the stochastic process:

1. **Definition**: A stochastic process is a collection of random variables ordered in time. Each random variable in the collection represents the state of the system at a certain point in time.
2. **Types of Stochastic Processes**:

- **Discrete-Time vs. Continuous-Time**: In a discrete-time stochastic process, the state of the system changes at specific time intervals (e.g., every minute, hour, day). In a continuous-time stochastic process, the state of the system can change at any time.
- **Discrete-State vs. Continuous-State**: In a discrete-state stochastic process, the state of the system can only take on a discrete set of values (e.g., integers). In a continuous-state stochastic process, the state can take on any value within a specified range.
- **Markov Process**: A special type of stochastic process where the future state of the system depends only on the current state, not on how the system arrived at its current state.

3. **Probability Distribution**: Each random variable within the stochastic process has a probability distribution, which can change over time. In addition, there is a joint distribution for any subset of the random variables.
4. **Expectation and Variance**: Just like with individual random variables, you can calculate the expected value (mean) and variance of each random variable within the stochastic process.
5. **Stationarity**: A stationary stochastic process is one in which the statistical properties (like mean and variance) do not change over time.
6. **Applications**: Stochastic processes have many applications in fields like finance (e.g., modeling stock prices), physics (e.g., quantum mechanics), biology (e.g., population dynamics), and engineering (e.g., signal processing).

7. **Prediction and Control**: A key goal in studying stochastic processes is to be able to predict future states of the system and to control the system in some way (e.g., reduce variability).

Understanding stochastic processes is a key part of fields such as statistics, econometrics, and machine learning.

Example

If we let y represents GDP, then denotes the third observation on GDP. The economic variable observed over time is random because we cannot perfectly predict it. The econometric model generating is called a Stochastic or random process. A sample of observed values is called a particular realization of the stochastic process. It is one of many possible paths that the stochastic process could have taken.

1.4.1 Stationary and Nonstationary Variables

A time series is stationary if its mean and variance are constant over time and if the covariance between two values from the series depends only on the length of time separating the two values, and not on the actual times at which the variables are observed. This can be summarized as follows:

(i) $E(Y_t) = U$…….(ia

The mean is constant over time. It exhibits ***mean reversion*** in that it fluctuates around a constant long-run mean. Shocks are temporary; over time, the effect of the shocks will dissipate and the series will revert to its long-run mean.

(ii) $V(Y_t) = q^2$……..(ib)

The variance is constant over time. It has a finite variance that is time-invariant.

(iii) $Cov(Y_t, Y_{t+s}) = Cov(Y_t, Y_{t-s})$……(ic)

The covariances are not constant over time. The covariances depend on the lag length, not time.

1.4.2 Weakly Stationarity and Strict Stationarity

A process is defined to be weakly stationary (or covariance stationary) if for all t, equations (1a), (1b), and (1c) holds. That is, the mean, variance and autocovariances are independent of time. Unlike weakly stationarity, strict stationarity is stronger because it requires that the whole distribution is unaffected by a change in time horizon, not just first and second order moments. Under joint normality assumption, the distribution is ompletely characterized by first and second order moments, and strict stationarity and weak stationarity are equivalent.

Self - Assessment Exercise 1

Outline stochastic process

1.5 Illustrative Example of How to Run Time Series Data in Eviews 12

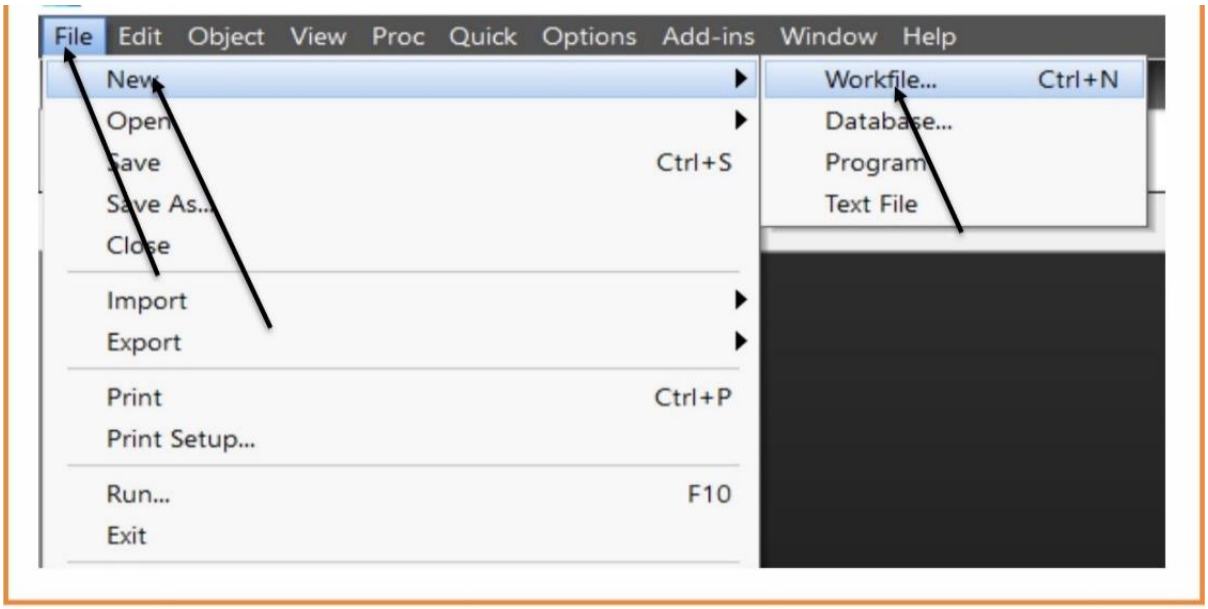
The following table contains time series data for;

Loading Data into EViews

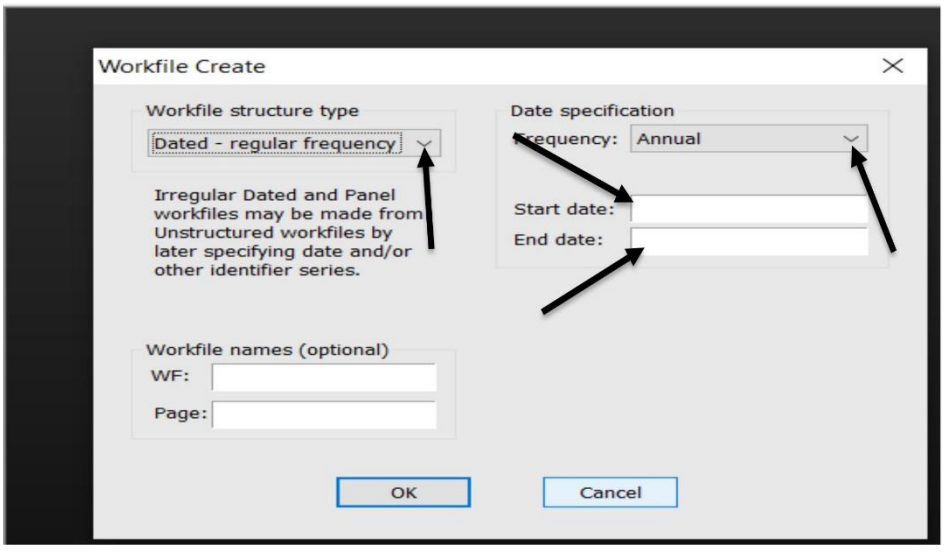
Before loading data into EViews, the first thing to do is to create a workfile.

How do one create a Workfile?

- To create a workfile, click **File / New / Workfile** on the ***Toolbar*** of the ***Main Window*** as shown below:

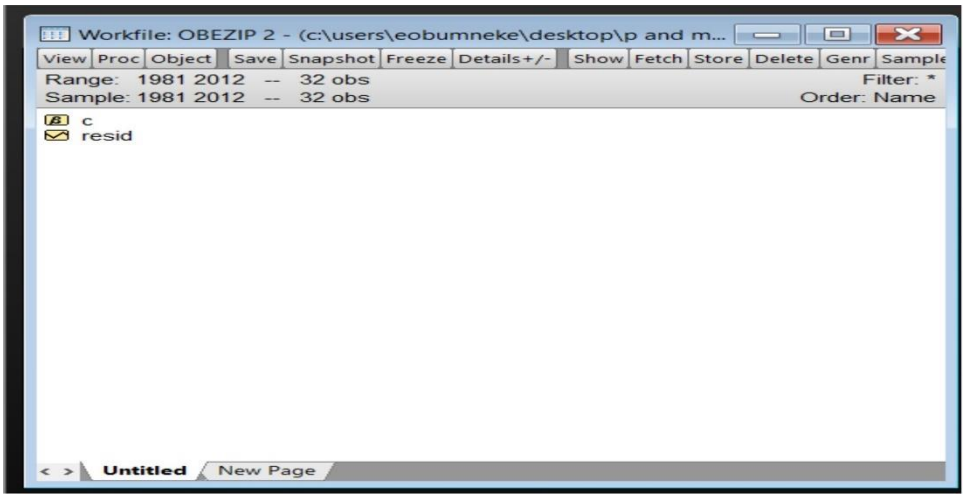


A dialogue box appears as below:



- In the dialogue box, choose the appropriate **Workfile structure type, Frequency, start date**, **End date** and (if so desired) the **Workfile name** (say obezip 2).

- Then click OK, a workfile named **OBEZIP 2** is created with 2 objects **C** and **RESID** appearing by default.



- Having created a workfile, one can then load data into EViews.

- **How do one load data into EViews?**

First, ensure that, the data is in conformity with one of the EViews supported foreign file formats such as Microsoft Excel or CSV (Comma delimited). This book presents the data in Excel format and consider three possible ways of loading data into EViews, these are:

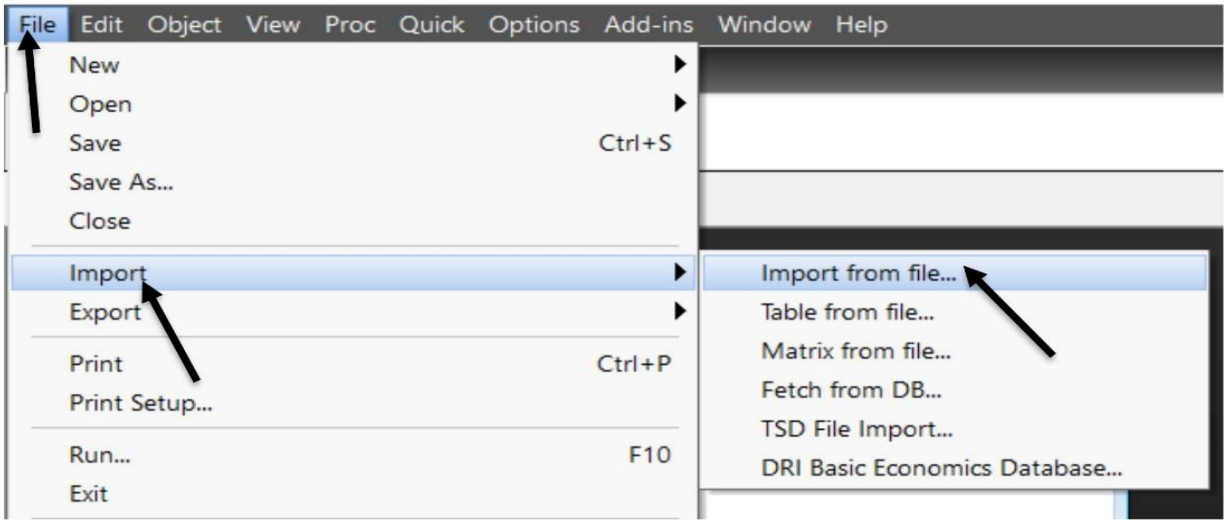
i. Import Method

ii. Drag and Drop Method

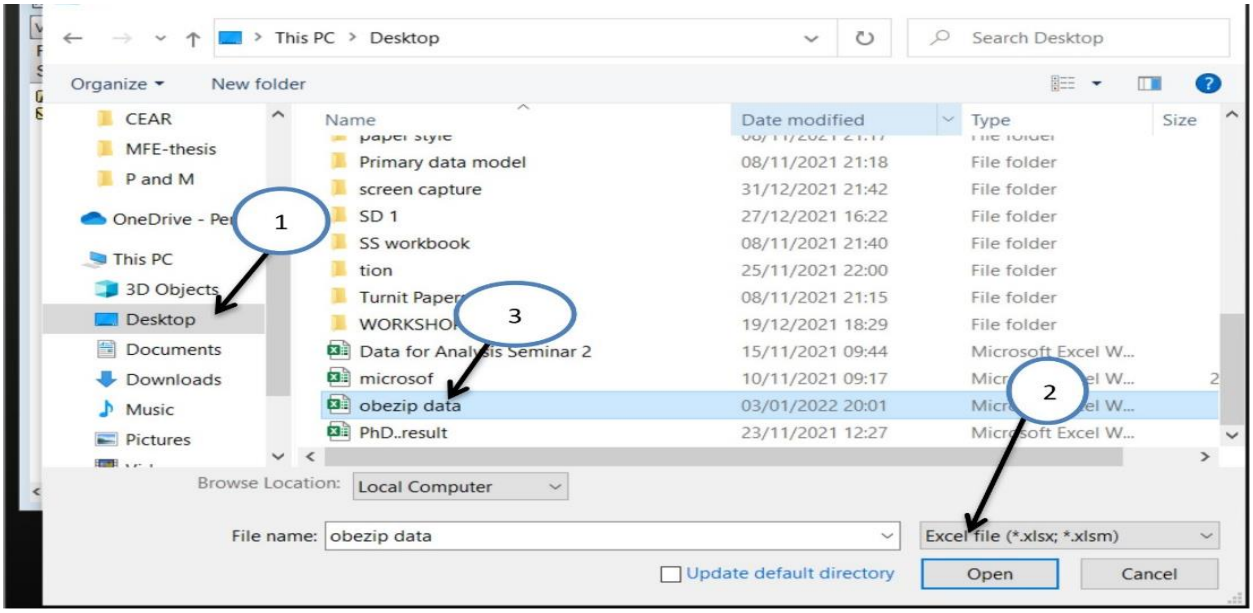
iii. Copy and Paste Method

i. Import Method

To import data into EViews workfile, click **Proc/Import/Import from File…**as below:

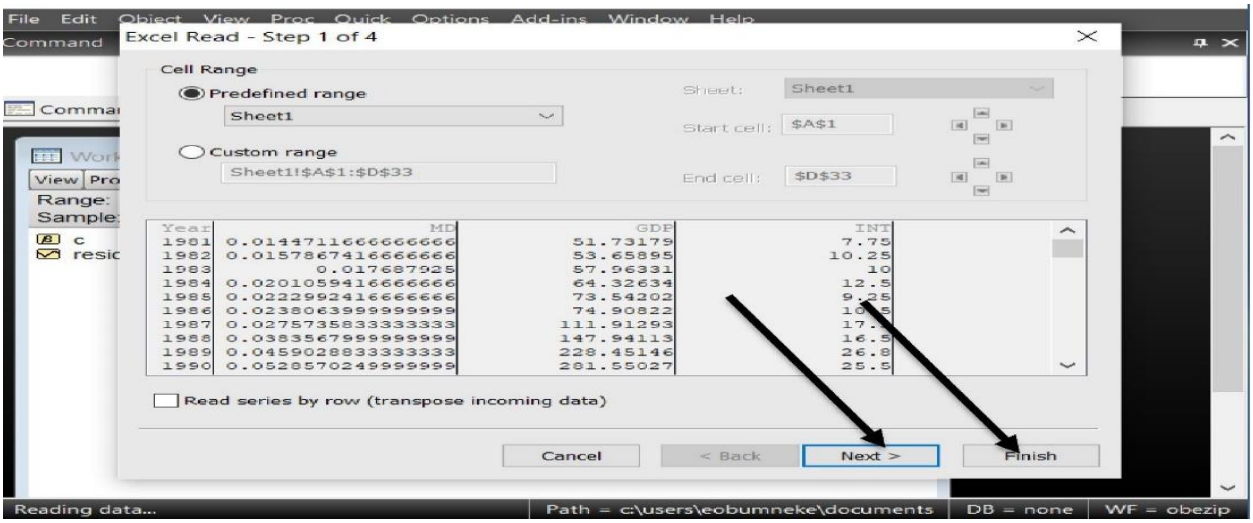


This takes one to a destination page, where one is requested to locate the data, it intends to load. To import data through this process, **the data must be in Microsoft Excel format (.xls or .xlsx).**

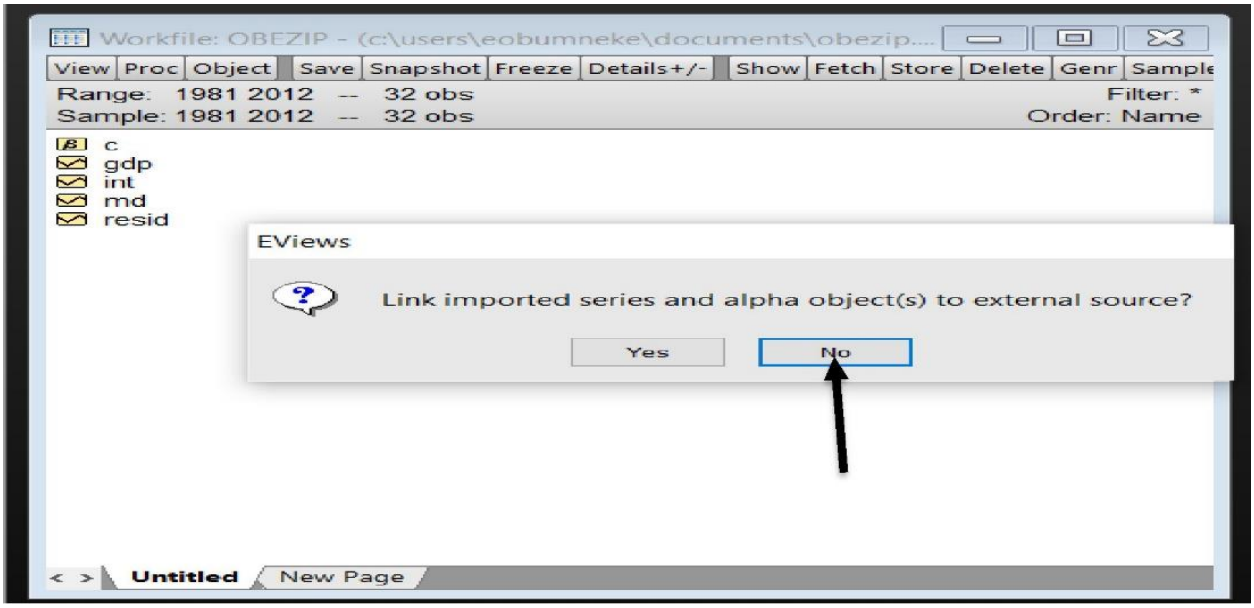


- Now, locate the destination of your file, select the appropriate file format (*.xls), then, locate your choice file. For example, **Desktop/Excel (*.xls)/obezip data.xls,** then click **Open**, an **Excel Spreadsheet Import** page will appear.

- In the Excel Spreadsheet Import, **Upper –left Cell Range** requires you to specify the sheet and cell where your actual data starts from. By default, **Sheet1** is a predefined range selected.
- By clicking on next pops up another window**. Column header** requires you to specify the header lines and header type. **Column info** describes the column for editing, the description and the data type. Usually, column series is expected to be displayed on the workfile, thus, **series in column** is selected by default. **Click on next**



- The next window describes the structure of data to be Imported, the import method, and the identifier series. **Click on finish**
- A pop-up window comes up asking whether to link imported series and alpha object(s) to external source (where the data are imported from). Click on No.

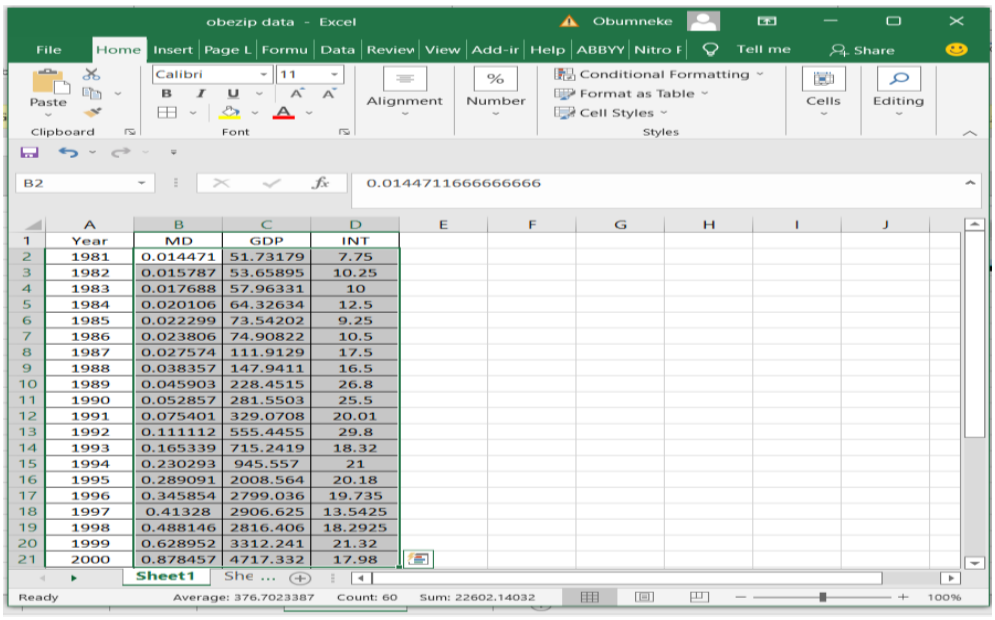


- You have now successfully loaded data into EViews workfile.

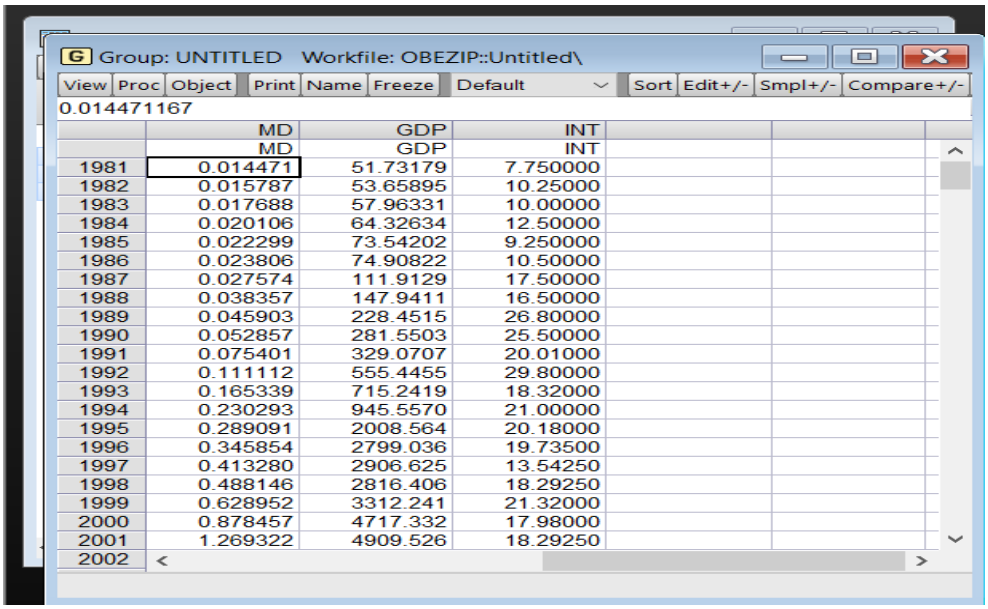
ii. Copy and Paste Method

To use the Copy and Paste method, it is more convenient to type *data md gdp int* in the command window, and then press the enter button (i.e. type *data* followed by the *series labels*). This allows you to create series named **md gdp** and **int** each containing N/A values.

Thereafter, select **Edit +/-** (as shown above) from the Group Menu to enable you paste data into the opened worksheet. Then, copy your data from the Excel sheet and paste into the EViews Group window. Make sure the ordering is in line.



	A	B	C	D
1	Year	MD	GDP	INT
2	1981	0.014471	51.73179	7.75
3	1982	0.015787	53.65895	10.25
4	1983	0.017688	57.96331	10
5	1984	0.020106	64.32634	12.5
6	1985	0.022299	73.54202	9.25
7	1986	0.023806	74.90822	10.5
8	1987	0.027574	111.9129	17.5
9	1988	0.038357	147.9411	16.5
10	1989	0.045903	228.4515	26.8
11	1990	0.052857	281.5503	25.5
12	1991	0.075401	329.0708	20.01
13	1992	0.111112	555.4455	29.8
14	1993	0.165339	715.2419	18.32
15	1994	0.230293	945.557	21
16	1995	0.289091	2008.564	20.18
17	1996	0.345854	2799.036	19.735
18	1997	0.41328	2906.625	13.5425
19	1998	0.488146	2816.406	18.2925
20	1999	0.628952	3312.241	21.32
21	2000	0.878457	4717.332	17.98



	MD	GDP	INT
	MD	GDP	INT
1981	0.014471	51.73179	7.750000
1982	0.015787	53.65895	10.25000
1983	0.017688	57.96331	10.00000
1984	0.020106	64.32634	12.50000
1985	0.022299	73.54202	9.250000
1986	0.023806	74.90822	10.50000
1987	0.027574	111.9129	17.50000
1988	0.038357	147.9411	16.50000
1989	0.045903	228.4515	26.80000
1990	0.052857	281.5503	25.50000
1991	0.075401	329.0707	20.01000
1992	0.111112	555.4455	29.80000
1993	0.165339	715.2419	18.32000
1994	0.230293	945.5570	21.00000
1995	0.289091	2008.564	20.18000
1996	0.345854	2799.036	19.73500
1997	0.413280	2906.625	13.54250
1998	0.488146	2816.406	18.29250
1999	0.628952	3312.241	21.32000
2000	0.878457	4717.332	17.98000
2001	1.269322	4909.526	18.29250
2002			

Then close the Group workfile. A dialogue box will appear to confirm your action, click **Yes.**

1.6 Summary

While reading this course, you were made to understand the definition a random or stochastic process, which entails a collection of random variables ordered in time. A Stochastic random variable could be continuous in time or discrete in time. Most

economic data are collected at discrete points in time, for instance, GDP

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Ilorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Ezie, O. (2022). A Practical Guide on Data Analysis Using EViews. Kabod Limited Publisher, Kaduna.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

A stochastic process is a sequence of random variables observed over time. It serves as a mathematical representation of systems evolving randomly. These processes can be categorized based on time (discrete-time or continuous-time) and state (discrete-state or continuous-state). A key type of stochastic process is the Markov process, where the future state depends only on the present state.

Each variable within the process has a probability distribution, which might change over time. Each variable also has an expectation (mean) and variance. If the process's statistical properties, such as mean and variance, remain constant over time, the process is considered stationary.

Stochastic processes find applications across many fields, including finance, physics, biology, and engineering. One of the main objectives of studying these processes is to predict future states of the system and, where possible, control the system.

UNIT 5: MULTICOLLINEARITY

1.1 Introduction

Multicollinearity is a frequently encountered phenomenon in econometric analyses that can significantly impact the results and interpretations of a regression model. This unit aims to elucidate the meaning, implications, and potential remedies of multicollinearity, enhancing our understanding of this crucial concept in econometrics.

1.2 Learning Outcomes

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

- Define the meaning of multicollinearity
- Explain the types of multicollinearity
- Discuss the consequences of multicollinearity
- Highlight the solution to multicollinearity problems

1.3 The meaning of multicollinearity

In econometrics and statistics, multicollinearity refers to a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy. In other words, it's a condition where the independent variables (or the predictors) are highly correlated with each other.

While multicollinearity does not undermine the predictive power of the model as a whole, it can significantly impact the individual predictive power of a variable and the model's interpretability. The reason is that multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your model. As a result, you might not be able to trust the p-values to identify independent variables that are statistically significant.

More specifically, with multicollinearity, the standard errors associated with the coefficients of the correlated variables tend to be large. This implies that coefficient estimates are less precise and can change dramatically with small variations in the model or the data. Therefore, the estimates of the coefficients can be unreliable and unstable.

It's also worth noting that multicollinearity can be an issue of degrees - ranging from perfect multicollinearity (where one independent variable is a perfect linear function of another independent variable) to no multicollinearity (where independent variables are not linearly related). The degree of multicollinearity can be measured using tools such as the Variance Inflation Factor (VIF), tolerance, or condition index.

1.4 Types of multicollinearities

This section delves into the types of multicollinearity—perfect multicollinearity and imperfect multicollinearity.

1. **Perfect Multicollinearity:** This occurs when one independent variable in a regression model can be expressed as an exact linear function of one or more other independent variables. Perfect multicollinearity may be due to several reasons. It can occur as a result of dummy variables, derived variables, or simply because of data errors. For example, if in a model, one includes both temperature in Fahrenheit and Celsius, they are perfectly collinear because they can be converted from one to the other using a linear function.

   Perfect multicollinearity is a severe issue because it makes it impossible to estimate the regression model using standard regression techniques like Ordinary Least Squares (OLS). The consequence of this is that the OLS estimator will not be unique. In essence, there will be an infinite number of solutions, making the estimates indeterminate. As a result, most statistical software will fail to produce output in the presence of perfect multicollinearity.

2. **Imperfect Multicollinearity:** This type of multicollinearity arises when two or more independent variables in a regression model are highly correlated but not perfectly so. Unlike perfect multicollinearity, imperfect multicollinearity doesn't completely hinder the model estimation process; however, it may lead to

inefficient and unstable estimates of the regression coefficients, which can in turn lead to unreliable hypothesis tests.

With imperfect multicollinearity, it becomes difficult to ascertain the individual influence of the correlated variables on the dependent variable, even though the overall model may still be valuable in predicting the dependent variable. This is because the estimates will have high standard errors, making them sensitive to minor changes in the model specification or the data.

1.5 Causes and consequences of multicollinearity

This section delves into the various causes of multicollinearity, thus enhancing our understanding of its origins and how it can be addressed.

1.5.1 Causes of multicollinearity

1. **Inclusion of Identical or Similar Variables:** Multicollinearity frequently occurs when a model includes variables that measure the same or similar constructs. For instance, including both a person's weight in pounds and weight in kilograms in the same model would induce perfect multicollinearity, as these two variables convey identical information.
2. **Creation of Derived Variables:** Creating variables based on other variables in the model can also lead to multicollinearity. For example, if a model includes a variable for income and another variable for income squared (to account for non-linear effects), these two variables will likely be highly correlated.

3. **Use of Dummy Variables:** The use of dummy variables can lead to multicollinearity, especially when the categories are not mutually exclusive or when a dummy variable is created for every category of a categorical variable (also known as the "dummy variable trap").
4. **Small Sample Sizes:** Multicollinearity can be more likely in studies with small sample sizes, where there may not be enough variation in the data to discern distinct patterns of influence for the independent variables.
5. **Over-Specification of the Model:** Over-specification of the model by including too many variables, especially when they are not all necessary or relevant, can lead to multicollinearity. Overly complex models may capture noise rather than signal, leading to high correlations between variables.
6. **Data Collection Issues:** In some cases, multicollinearity can arise due to the way data is collected. For example, if a survey asks similar questions in slightly different ways and responses to these questions are used as independent variables, this can create multicollinearity.

Self- Assessment 1

What are the causes of multicollinearity.

1.5.2 Consequences of multicollinearity

This section explores these implications.

1. **Inflated Standard Errors:** One of the most notable consequences of multicollinearity is the inflation of standard errors of the regression coefficients. When independent variables are highly correlated, it becomes challenging to isolate the effect of each variable on the dependent variable, which leads to higher standard errors. The larger the standard errors, the less precise the estimated coefficients and the less reliable the hypothesis tests based on these estimates.
2. **Unstable Coefficients:** Multicollinearity can make the estimated regression coefficients unstable. This means that with slight changes in the data or model specification, the estimated coefficients can fluctuate significantly. This lack of robustness can be problematic, particularly when the model is used for forecasting or policy evaluation.
3. **Problematic Interpretation:** The high correlation between variables can make it difficult to interpret the coefficients. In a multicollinear model, the estimated coefficient of a variable is contingent on the other variables in the model. As a result, the estimated coefficient of a particular variable may not reflect its true effect on the dependent variable, complicating the interpretation of the model's results.
4. **Reduced Statistical Power:** Multicollinearity can reduce the statistical power of the model, which is the ability of a test to detect an effect if there is one. This is because multicollinearity inflates the standard errors, making it harder to reject the null hypothesis. Consequently, there is an increased risk of committing a Type II error—failing to reject a false null hypothesis.

1.5.3 Solution to multicollinearity

1. **Removing Variables:** One straightforward way to deal with multicollinearity is by removing one or more of the correlated variables from the model. This decision should ideally be informed by domain knowledge, theoretical considerations, and the importance of the variables for the analysis.
2. **Combining Variables:** Instead of dropping one of the correlated variables, another approach is to combine them into a single predictor. For example, if two variables are measured in different units but represent the same underlying construct, they could be standardized and averaged to create a single composite measure.
3. **Principal Component Analysis (PCA) or Factor Analysis:** These statistical techniques can help reduce the dimensionality of the data and transform the original variables into a new set of uncorrelated variables, which can then be used in the regression analysis.
4. **Increasing Sample Size:** If possible, increasing the sample size can help mitigate the effects of multicollinearity. Larger samples provide more information and can help disentangle the effects of the correlated independent variables.
5. **Ridge Regression:** This is a type of regression analysis that introduces a small bias into the regression estimates to obtain substantial reductions in the variance and more stable estimates. It can be particularly useful when dealing with multicollinearity.

6. **Variance Inflation Factor (VIF):** VIF can be used to detect the presence and severity of multicollinearity. If a variable has a high VIF, it might be a good idea to remove it or combine it with other variables.

Self- Assessment 2

Highlight the consequences of multicollinearity

1.6 Summary

In this unit, you have been able to learn the meaning of multicollinearity, types, causes and possible solutions. While multicollinearity does not bias the estimates in a linear regression model, it can make them inefficient and difficult to interpret. The two types—perfect and imperfect multicollinearity—represent different degrees of correlation between independent variables. Understanding these types is essential for identifying, diagnosing, and remedying multicollinearity in econometric analyses.

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Ilorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Ezie, O. (2022). A Practical Guide on Data Analysis Using EViews. Kabod Limited Publisher, Kaduna.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated. This makes it difficult to determine the separate impact of each variable on the dependent variable. Here are some common causes of multicollinearity:

1. **Inclusion of Identical Variables**: Multicollinearity can occur when the same variable is inadvertently included multiple times in a regression model under different names.
2. **Inclusion of Linear Combinations**: If one independent variable is a linear combination of one or more other independent variables, this can lead to perfect multicollinearity. For example, if you include both "total sales" and "total sales per store" in a regression model, these variables will be highly correlated.
3. **Inclusion of Dummy Variables**: If a categorical variable with n categories is included in the model by creating n dummy variables, this will result in multicollinearity. This is because the dummy variables will be perfectly correlated with the intercept term. To avoid this, we typically create n-1 dummy variables for a categorical variable with n categories.
4. **Data Collection Method**: Sometimes the way data is collected can result in multicollinearity. For example, in a survey, if two questions are very similar, the responses to these questions will likely be highly correlated.
5. **Over-Specification of the Model**: Including too many variables in a model, especially in small sample sizes, can lead to multicollinearity. Some variables may be unnecessary and simply introduce noise and correlation with other variables.
6. **Limited Variation in the Data**: If there is limited variation in the data, small correlations between variables can become inflated, leading to multicollinearity.
7. **Temporal Data**: In time series data, variables can become highly correlated simply due to trends over time. For example, both income levels and education levels might increase over time, but this doesn't necessarily mean they are causally related.

Answer to Self- Assessment 2

Multicollinearity can cause several issues in a regression analysis. Here are some key consequences:

1. **Unreliable Parameter Estimates**: Multicollinearity can result in unstable and unreliable estimates of the regression coefficients. A small change in the data can

lead to a large change in the coefficient estimates. This makes it difficult to interpret the individual coefficients.

2. **Inflated Standard Errors**: The standard errors of the coefficient estimates can be inflated when multicollinearity is present. This can lead to a failure to reject the null hypothesis of no effect (type II error) because the confidence intervals around the coefficients are wider.
3. **Significance Tests May Be Misleading**: When multicollinearity is present, significance tests on individual coefficients may be misleading. A coefficient might not be statistically significant, even though it is theoretically expected to have an impact on the dependent variable.
4. **Poor Model Generalizability**: Models with multicollinearity may perform well on the training data but poorly on new, unseen data. This is because the unstable coefficients are tailored to the idiosyncrasies of the training data.
5. **Incorrect Model Interpretation**: Due to multicollinearity, the signs of the coefficients could be opposite to what is expected based on theoretical reasoning. This can lead to incorrect interpretations of the model.
6. **Difficulty in Identifying the Most Important Predictors**: Multicollinearity makes it challenging to identify the most important predictors due to the shared variance among predictors.

MODULE 2: STATIONARITY AND AUTOREGRESSIVE PROCESS

Unit 1. Autoregressive Process

Unit 2: Concept of Stationarity

Unit 3: Cointegration

Unit 4: Autoregressive Distributed Lag (ARDL) Model

Unit 5: ARDL Post Estimation Tests

UNIT 1: AUTOREGRESSIVE (AR) PROCESS

1.1 Introduction

In the preceding unit, you learnt time series. In the present unit, we will discuss autoregressive models. In regression analysis involving time series data, if the regression model includes one or more lagged values of the dependent variable among its explanatory variables, it is called autoregressive model.

1.2 Learning Outcomes

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

- Discuss the meaning of the term Autoregressive (AR)
- Explain estimation of an Autoregressive Model (AR)
- State autocorrelation or Serial Correlation
- Analyze consequences of Serial Correlation
- State steps for carrying out the LM Test

1.3 Meaning of the Term Autoregressive (AR)

The term autoregressive (AR) describes a random or stochastic process used in econometrics through which future values are estimated based upon a weighted sum of previous or past values. The "auto" signals autoregressive models are regression of variable in question against itself.

If yt is univariate for instance $yt-1, yt-2, \cdots yt-p$ the model is called autoregressive AR(P) on the other hand if yt is multivariate, i.e $yt=[y_{1t}, y_{2t}, \cdots y_{Nt}]$, the model is called vector autoregressive model VAR (P)

The AR model has proven to be useful for describing the dynamic behaviour of economic and financial time series and forecasting. It often provides superior forecast and are quite flexible because they can be made conditional on the potential future paths of specified variables in the model.

The AR is also used for structural inference and policy analysis. In structural analysis, certain assumptions about the causal structure of the data under investigation are imposed, and the resulting causal impacts of unexpected shocks or innovations to specified variables on the model are summarized Wooldridge (2013)

1.4 Estimation of an Autoregressive Model (AR)

Autoregressive (AR) models constitute an essential category within time series analysis in econometrics. As a cornerstone of understanding temporal dependencies in sequential data, AR models play a pivotal role in diverse fields, including economics, finance, engineering, and physical sciences.

Autoregressive models are used when a value from a time series is regressed on previous values from the same time series. The basic structure of an autoregressive model of order p, or AR(p), is expressed as:

$$Y_t = c + \varphi 1 Y_{(t-1)} + \varphi 2 Y_{(t-2)} + ... + \varphi p Y_{(t-p)} + \varepsilon_t$$

where Y_t represents the current value of the time series, Y_(t-1), Y_(t-2), ..., Y_(t-p) represent past values, c is a constant, φ1, φ2, ..., φp are the parameters of the model, and ε_t is a white noise error term.

Estimation of an AR model typically involves the following steps:

1. Identification of Model Order: The first step is to determine the order of the AR model (p), which indicates the number of previous observations (lags) to be included in

the model. This can be done using various statistical tools like Partial Autocorrelation Function (PACF) plots or using information criteria like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), which trade off model fit with model complexity.

2. Parameter Estimation: After deciding on the order of the AR model, the next step is to estimate the parameters (φ1, φ2, ..., φp). The most common method for estimating these parameters is the method of Ordinary Least Squares (OLS). This method minimizes the sum of the squared residuals, yielding the best linear unbiased estimates.

3. Model Checking: Once the model parameters have been estimated, it is essential to check the model's adequacy. This is often achieved by examining the residuals of the fitted model. If the model provides an adequate description of the data, the residuals should behave like white noise. Tools such as autocorrelation and partial autocorrelation functions of the residuals, or the Ljung-Box test, are commonly used.

4. Forecasting: With the AR model estimated and checked, it can be used for forecasting future values of the time series. Since AR models use past values, the accuracy of the forecasts typically diminishes as the forecast horizon increases.

1.5 Autocorrelation or Serial Correlation

It is now a common practice to treat the terms autocorrelation and serial correlation synonymously although there may be technical differences. But for this course, we will use both concepts interchangeably.

Autocorrelation, also known as serial correlation, is a crucial concept in time series analysis within econometrics. It measures the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

Autocorrelation is based on the principle of correlation, which measures the statistical association between two variables. However, in autocorrelation, we are essentially comparing a variable to itself at different points in time. This is particularly useful in time series data where observations are often related across time. The autocorrelation function (ACF) is a tool that provides correlation coefficients of a time series with its own lagged values.

Mathematically, the autocorrelation of a time series Y at lag k can be expressed as:

$$\rho_k = Cov(Y_t, Y_{(t-k)}) / \sqrt{[Var(Y_t) * Var(Y_{(t-k)})]}$$

where Y_t is the current time period, $Y_{(t-k)}$ is the time period k steps in the past, Cov is the covariance, and Var is the variance.

The implications of autocorrelation in econometric analysis are significant:

1. **Model Specification:** Autocorrelation can help determine the order of an autoregressive (AR) model. A plot of autocorrelation against various lags can assist in identifying the number of past values (lags) that significantly affect the current value.

2. **Violation of Classical Regression Assumptions:** In regression models, the assumption of no autocorrelation is often required. Autocorrelation in the error terms violates the Gauss-Markov assumptions, which state that error terms should be uncorrelated for the ordinary least squares (OLS) estimator to be the best linear unbiased estimator (BLUE). Violation of this assumption leads to inefficient parameter estimates, although they remain unbiased.
3. **Diagnostic Tool:** Autocorrelation can serve as a diagnostic tool in checking the adequacy of a fitted model. If a time series model is a good fit, then the residuals should exhibit no autocorrelation.
4. **Modeling Dependence:** Autocorrelation is a fundamental concept underlying many types of time series models, including AR models, moving average (MA) models, and ARIMA models. These models explicitly incorporate autocorrelation to capture the dependency structure in time series data.

1.5.1 Consequences of Serial Correlation or Autocorrelation

When the observations of a dataset are correlated with each other, several problems may arise, especially in the context of econometric models that rely on the assumption of independence between observations:

1. **Inefficiency of Ordinary Least Squares (OLS) Estimates:** One of the principal assumptions of OLS estimation in linear regression is that the error terms are uncorrelated (no autocorrelation). When autocorrelation is present, this assumption

is violated. While the OLS estimators remain unbiased, they are no longer efficient. This means that there are other estimators with a smaller variance, implying that OLS does not provide the best linear unbiased estimates (BLUE).

2. **Misleading Test Statistics:** The presence of autocorrelation leads to incorrect standard errors of the estimates, which subsequently leads to misleading t-statistics and incorrect inference about the coefficients. Confidence intervals and hypothesis tests can be invalid, and p-values can be misleading. In particular, if positive autocorrelation is present, the confidence intervals for the coefficient estimates are typically narrower than they should be, leading to overconfidence in estimates and increased Type I error rates.
3. **Model Mis-specification:** The presence of autocorrelation often indicates model mis-specification. It could mean that important variables have been omitted or that the functional form of the model needs to be rethought. For example, a lagged dependent variable might need to be included to correct autocorrelation in an autoregressive model.
4. **Impaired Forecasting:** Models suffering from autocorrelation are less accurate in their forecasts. Since these models do not fully account for the relationship between time periods, the forecast error may be higher.
5. **Durbin Watson Statistic:** The presence of autocorrelation also affects the value of the Durbin Watson statistic, which is a test statistic used to detect the presence of autocorrelation. If the Durbin Watson statistic deviates significantly from 2, it signals the presence of autocorrelation.

If the error term is known to exhibit serial correlation, then the consequences for the OLS estimates can be summarized as follows:

i. the OLS estimators of the $\hat{\beta}s$ are still unbiased and consistent. This is because, both unbiasedness and consistency in this case were violated;
ii. the OLS estimators will be inefficient and therefore no longer BLUE;
iii. the estimated variances of the regression coefficients will be biased and inconsistent, and therefore hypothesis testing is no longer valid. In most of the cases, R^2 will be overestimated (indicating a better fit than the one that truly exists) and the t-statistics will tend to be higher (indicating a higher significance of our estimates than for the correct one).

Causes Of Autocorrelation

The presence of autocorrelation in a dataset can stem from several sources. Here are some of the main causes:

1. **Inherent Nature of the Data:** Autocorrelation often arises naturally in time series data due to the inherent temporal ordering of observations. Certain data sets, such as economic and financial time series, often exhibit autocorrelation because the value of the variable in the current period is influenced by its value in the previous period. For example, today's stock price is usually strongly related to yesterday's stock price.

2. **Omitted Variables:** Autocorrelation can also occur when an important variable that influences the dependent variable is omitted from the model. If this omitted variable follows a trend over time or is autocorrelated, then this can induce autocorrelation in the residuals of the model.
3. **Incorrect Model Specification:** Autocorrelation can be the result of incorrect model specification. For example, if a linear model is used when the true relationship is nonlinear, the residuals from the model may show patterns over time, indicating autocorrelation.
4. **Lagged Dependent Variables:** The inclusion of lagged dependent variables as regressors can also induce autocorrelation. In such models, by construction, the error term in one period is related to the error term in previous periods.
5. **Measurement Error:** Sometimes, autocorrelation may be a consequence of the measurement errors or data collection process. For example, if the same error is made when measuring the variable in successive time periods, this can result in autocorrelation.
6. **Simultaneity:** Autocorrelation can also arise from simultaneity, where the dependent variable and one or more independent variables are determined simultaneously and affect each other.
7. Specification Bias. Excluded Variables Case and Incorrect Functional Form - Suppose that Y_t is connected to X_t with a quadratic relationship, but one wrongly, assume and estimate a straight-line. The error term obtained from the straight-line

specification will depend on X_t of quadratic relationship. If X_t is increasing or decreasing over time, u_t will be doing the same, indicating auto-correlation.

8. Manipulation of Data. In empirical analysis, the raw data are often "manipulated". For example, in time series regressions involving quarterly data, such data are usually derived from the monthly data by simply adding three monthly observations and dividing the sum by 3. The averaging introduces smoothness into the data by dampening the fluctuations in the monthly data. Another source of manipulation is interpolation or extrapolation of data.

Self-Assessment Exercise 1

Discuss the consequences of serial correlation

1.5.2 Testing for serial correlation

Usually, we estimate a linear model by ordinary least squares (OLS) assuming that the classical assumptions holds, and then attempt to test whether those assumptions appear to be satisfied for the estimated model.

We therefore need to be able to test whether the assumption of zero covariance of the disturbances appears to hold, based on the estimated residuals.

One method commonly used in applied econometric research for detecting autocorrelation is to plot the regression residuals, *e*, against time. If the residuals in

successive periods show a regular time pattern (for example, a saw tooth pattern, or a cyclical pattern) we conclude that there is autocorrelation.

Until the 1990s, the most commonly used test for serial correlation was the Durbin-Watson test for first order autocorrelation. Nowadays, the Lagrange Multiplier (LM) test is more popular because it can be applied in a wider set of circumstances and can test for higher-order serial correlation such as AR(1), AR(2), AR(3), etc.

The Durbin–Watson Test

The most frequently used statistical test for the presence of serial correlation is the Durbin–Watson (DW) test (see Durbin and Watson, 1950), which is valid when the following assumptions are met:

a) the regression model includes a constant;
b) serial correlation is assumed to be of first-order only, hence, cannot detect higher order serial correlation; and
c) the equation does not include a lagged dependent variable as an explanatory variable or among the regressors.

Therefore, alternative tests are often sought when the above assumptions are not met. An important one is the Breusch-Godfrey Lagrange multiplier test.

DW test statistic given by:

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

A Rule of Thumb for the DW Test

From the estimated residuals an estimate of ρ can be obtained as:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{n} \hat{u}_t}$$

DW statistic is approximately equal to $d = 2(1-\hat{\rho})$. Because ρ by definition ranges from -1 to 1, the range for d will be from 0 to 4.

According to Ezie and Ezie (2021), one can have three different cases:

a) $\rho = 0; d = 2$: therefore, a value of d close to 2 indicates that there is no evidence of serial correlation.

b) $\rho \approx 1; d \approx 0$: a strong positive auto-correlation means that ρ will be close to $+1$, and thus d will have very low values (close to zero) for positive auto-correlation.

c) $\rho \approx -1; d \approx 4$: similarly, when ρ is close to -1, then d will be close to 4, indicating a strong negative serial correlation.

The Breusch–Godfrey LM Test for Serial Correlation

The DW test has several drawbacks that make its use inappropriate in various cases. For example: (a) it may give inconclusive results; (b) it is not applicable when a lagged dependent variable is used; and (c) it can't take into account higher orders of serial correlation. For these reasons, Breusch (1978) and Godfrey (1978) developed an LM test that can accommodate all the above cases.

Example

Table M2.1.1: Breusch-Godfrey Serial Correlation LM Test Result
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	1.277441	Prob. F(2,27)	0.2951
Obs*R-squared	2.852697	Prob. Chi-Square(2)	0.2402

From the result in Table M2.1.1, one accepts the null hypothesis that there is no serial correlation among the error terms used in a model. This is premised on the probabilities of "F-statistic" and "Obs*R-squared" statistic (which are both statistics for the LM test) are greater than 0.05 (that is, 0.2951 and 0.2402 are both greater than 0.05). Thus, one accepts the null hypothesis of no serial correlation. It implies that there is no presence of serial correlation in the estimated result.

1.6 Summary

Autoregressive (AR) models describe a random or stochastic process used in econometrics through which future values are estimated based upon a weighted sum of previous or past values. The “auto” signals autoregressive models are regression of variable in question against itself. In this unit you have learn the meaning of the term Autoregressive (AR) and how to estimate of an Autoregressive Model (AR). You also learn autocorrelation or Serial Correlation and the consequences of Serial Correlation as well as Testing for serial correlation and steps for carrying out the LM Test.

Tutor Marked Assignment

Interpret the serial correlation result below:

Table M2.1.2: Breusch-Godfrey Serial Correlation LM Test

F-statistic	9.215102	Prob. F(2,414)	0.0001
Obs*R-squared	17.81519	Prob. Chi-Square(2)	0.0001

Source: EViews-12

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Ilorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Ezie, O. (2022). A Practical Guide on Data Analysis Using EViews. Kabod Limited Publisher, Kaduna.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The presence of serial correlation can lead to several problems:

1. **Inefficient Parameter Estimates**: While the OLS estimates remain unbiased in the presence of serial correlation, they are no longer efficient. This means that there are other estimators that can provide a smaller variance than the OLS estimates, hence more accurate predictions.
2. **Standard Errors Are Misestimated**: The presence of serial correlation leads to an underestimation of the standard errors of the regression coefficients. This can

result in overconfidence in the significance of the predictors (too many false positives) and can lead researchers to erroneously conclude that a variable is significant when it is not.

3. **Inferential Statistics Are Invalid**: Serial correlation invalidates typical inferential statistics, such as t-tests and F-tests, that are based on the assumption of independent errors. This can lead to incorrect conclusions about the relationship between the variables.
4. **Model Misspecification**: Serial correlation often indicates that the model is misspecified. It may suggest that an important variable has been omitted, that the functional form is incorrect, or that there is an issue with the measurement of the dependent variable.
5. **Reduced Forecasting Accuracy**: If the model's error terms are serially correlated, the model might not produce the most accurate forecasts.

To rectify these issues, one can apply techniques such as adding lags of the dependent and/or independent variables, differencing the data, or using time-series specific methods like autoregressive integrated moving average (ARIMA) models. Other methods like the Durbin-Watson test, Breusch-Godfrey test, or visual inspection of the residuals can be used to detect the presence of serial correlation.

UNIT 2: CONCEPT OF STATIONARITY

1.1 Introduction

In econometric analysis the use of stationarity is an inevitable way of ensuring that that the data used for regression is reliable and can be used for prediction of future performance. It help to ensure that the data that will be used for regression is mean reverting and stable.

1.2 Learning Outcomes

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

- Discuss stationarity and non-stationarity
- Explain the meaning of unit root and its importance
- Explain how to stationarise non-stationary series
- Conduct unit roots test in the AR(1) Model
- Estimate stationarity of variables in EViews software

1.3 What is Stationarity

A time series is stationary, if its mean, variance, and autocovariance (at various lags)

remain the same no matter at what point we measure them; that is, they are time invariant. Such a time series will tend to return to its mean (called mean reversion) and fluctuations around this mean (measured by its variance) will have a broadly constant amplitude (Gujarati, 2004).

It is assumed under multiple regression analysis, that, all the series are stationary at level (that is, the order of integration of each of the series is zero, $I(0)$. However, in reality, this might not be the case; and estimating a model with non-stationary series could produce spurious results. The unit root test is used to determine the stationarity or non-stationarity of a given time series. It is a test of stationarity or non-stationarity of series data used in the model.

1.4 Unit Root Test/ Stationarity Test

In the field of econometrics, understanding the concepts of unit roots and stationarity is paramount to the accurate modeling and forecasting of time series data. These concepts are crucial as they help to ensure the validity and reliability of statistical results, thereby aiding sound decision-making in economic policy, investment strategies, and various other applications.

A unit root is a feature of some stochastic or random processes that can cause issues in statistical data analysis. A time series with a unit root is non-stationary, meaning that its statistical properties such as mean, variance, and autocorrelation change over time. The

presence of a unit root can lead to 'spurious regression', where it appears that two unrelated series are significantly related due to their shared trends over time.

Formally, a simple autoregressive model AR(1) is defined as $Y_t = \alpha + \rho Y_{(t-1)} + \varepsilon_t$, where Y_t is the variable of interest at time t, α is a constant, ρ is the autoregressive parameter, and ε_t is a white noise error term. If the absolute value of ρ is equal to 1, the process has a unit root and is non-stationary.

In contrast, stationarity refers to a statistical concept in which a time series' key properties do not change over time. A stationary process has a constant mean, variance, and autocorrelation structure. Most statistical modeling techniques require data to be stationary because they are based on the assumption that the underlying data's properties remain constant over time.

When working with time series data, it's crucial to test for stationarity and the presence of unit roots before performing further analysis. There are several statistical tests to examine these, including the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron test, and the KPSS test. These tests formulate a null hypothesis that a unit root is present (i.e., the series is non-stationary). If the test statistic is less than the critical value, then the null hypothesis is rejected, indicating the series is stationary.

If a time series is found to be non-stationary, there are several techniques that can be used to transform it into a stationary series. The most common is differencing, where instead of analyzing the series itself, we analyze the difference between consecutive

observations. This often helps stabilize the mean of a time series by removing changes in the level over time and thereby eliminating trends and seasonality.

Why it is important to Test for Unit Root

We usually consider unit root for the following reasons:

i. to evaluate the behaviour of series over time.
ii. to determine how series respond to shocks
iii. to test for market efficiency

Several tests have been developed in the literature to test for unit root. Prominent among these tests are ADF, PP, DFGLS and KPSS.

1.5 How Non-stationary series be stationarised

A non-stationary time series, one where the statistical properties such as mean and variance change over time, poses challenges in econometric analysis. Stationarity is a desirable property for many statistical models because they assume the underlying data's properties remain constant over time. When faced with non-stationary data, certain transformations can be applied to "stationarize" the series. Here are some commonly used methods:

1. **Differencing:** The most common way to stationarize a series is through differencing. A differenced series is the change between consecutive observations. For example, the first differenced series is Y'*t* = *Y_t* - *Y*(t-1), where Y represents

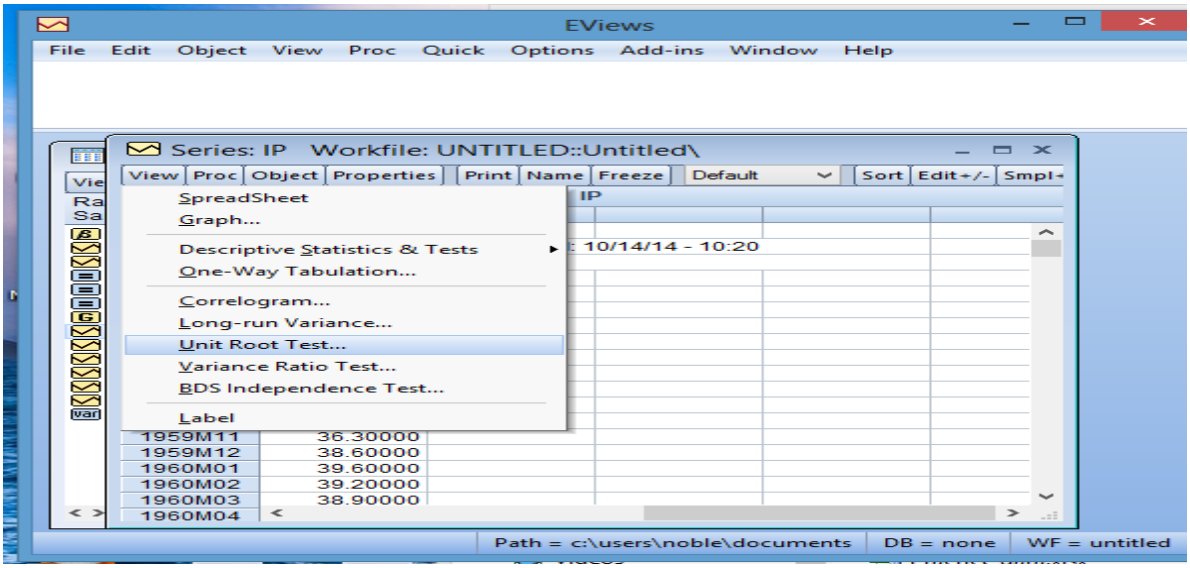
the non-stationary series. Sometimes, more than one round of differencing may be needed to achieve stationarity.

2. **Log Transformation:** Taking the log of a series can help to stabilize an exponential growth trend and reduce the variance of the series. This transformation is commonly used when dealing with economic time series data that exhibit exponential growth, such as GDP or stock prices.
3. **Seasonal Differencing:** In the presence of a seasonal trend, a seasonal differencing may be applied. For instance, in monthly data with yearly seasonality, the difference from the same month of the previous year can be taken.
4. **Detrending:** If a non-stationary series exhibits a deterministic trend, you can fit a trend line (using, say, least squares regression) and then analyze the detrended series (the residuals from the trend fit).
5. **Deflating:** If the series represents economic data and is influenced by price changes over time, you can adjust it for the effects of inflation. The series can be deflated by dividing it by a price index.
6. **Applying Mathematical Transformations:** At times, mathematical transformations like square roots or cube roots can help in achieving stationarity.

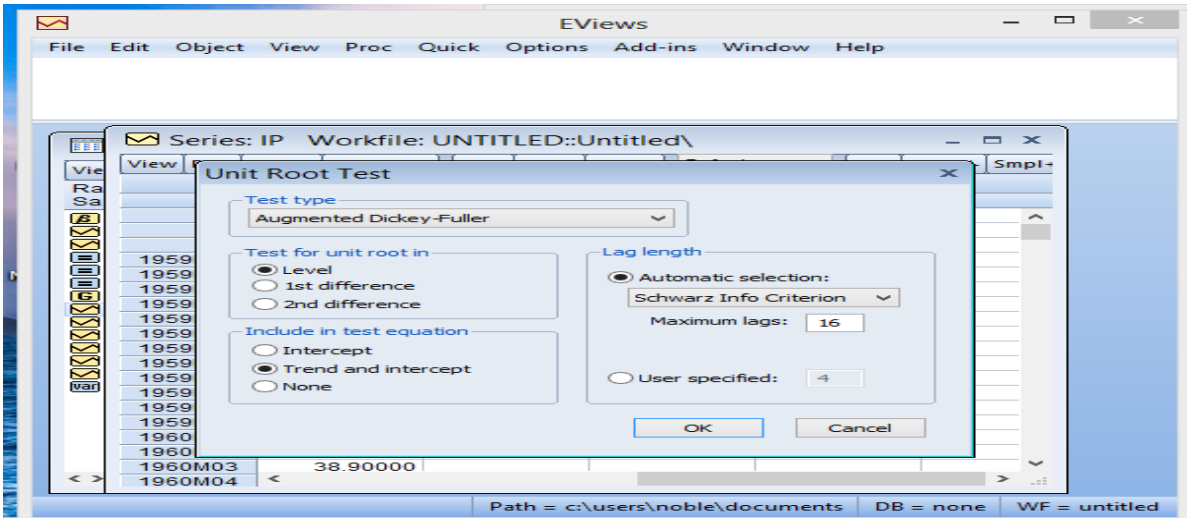
It's crucial to note that before and after applying each of these transformations, tests for stationarity should be carried out to confirm the success of the operation. The Augmented Dickey-Fuller (ADF) test is commonly used to test for stationarity.

1.5.1 Unit root/Stationarity Test on EViews

Double click on the series name to open the series window, and choose **View/Unit Root Test…**



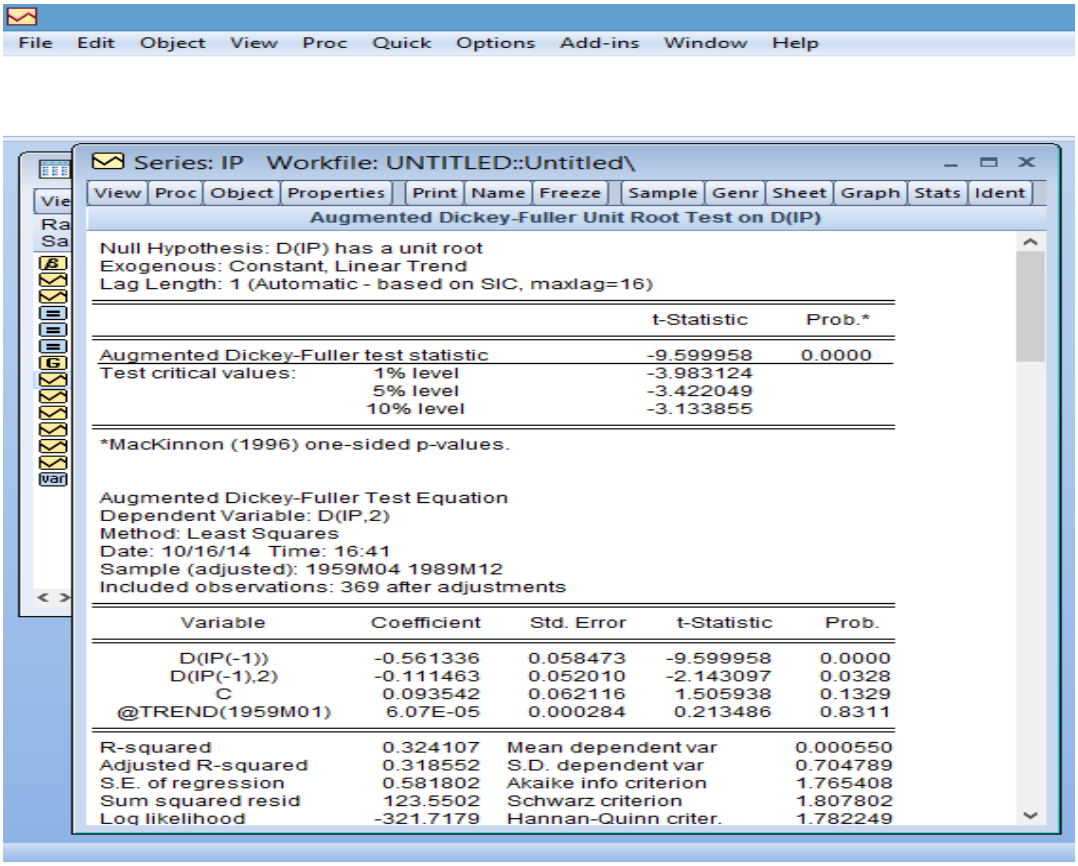
- Use the topmost combo box to select the type of unit root test that you wish to perform.
- Next, specify whether you wish to test for a unit root in the level, first difference, or second difference of the series.
- Lastly, choose your exogenous regressors. You can choose to include a constant, a constant and linear trend, or neither (there are limitations on these choices for some of the tests).
- You can click on **OK** to compute the test using the specified settings, or you can customize your test using the advanced settings portion of the dialog.



Note:

The first part of the unit root output provides information about the form of the test (the type of test, the exogenous variables, and lag length used), and contains the test output, associated critical values, and in this case, the p-value.

The second part of the output shows the intermediate test equation that EViews used to calculate the ADF/PP statistic.



Testing for the Order of Integration

A test for the order of integration is a test for the number of unit roots. In order to carry out this test, the following steps are necessary.

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 ΔY_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 Y_t as $\Delta^2 Y_t = \Delta Y_t - Y_{t-1}$, and test $\Delta^2 Y_t$ 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 instance, if $\Delta^3 Y_t \sim I(0),$ then $\Delta^2 Y_t \sim I(1),$ and $\Delta Y_t \sim I(2),$

and finally $Y_t \sim I(3)$; which means that Y_t needs to be differentiated (or differenced) three times to become stationary.

Result at Level (GDP)

Table M2.2.1: Augmented Dickey-Fuller Test-GDP (at Level)

Null Hypothesis: GDP has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=3)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.113029	0.2422
Test critical values:	1% level	-3.886751	
	5% level	-3.052169	
	10% level	-2.666593	

Source: EViews-12

Interpretation

i. The value of t-statistics must be greater than a specified significant level pre-selected, either at 1%, 5% or 10% levels respectively (in absolute terms). One must pick and stick to one.

ii. Probability value must be significant, that is, very close to zero.

From the result in Table above, it could be observed that the t-statistics is not greater than any of the levels of significance, and, at the same time, not significant, given the probability value of 0.2422. It can conclude that, the variable (GDP) is not stationary at level, and therefore, one must test at 1st difference. Testing GDP at first difference see Table below:

Result at 1st Difference (GDP)

Table M2.2.2: Augmented Dickey-Fuller Test-GDP (at 1st Difference)

Null Hypothesis: D(GDP) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=3)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.679821	0.0159
Test critical values:	1% level	-3.920350	
	5% level	-3.065585	
	10% level	-2.673459	

Source: EViews-12

Interpretation

The result in Table M2.2.2 above shows that t-statistic is greater than 5 and 10 percent levels of significance and as such, significant at both levels. This was further confirmed by the p-value of 0.0159. It can then conclude that the variable GDP is stationary at 1st difference.

Result at Level (INF)

Table M2.2.3: Augmented Dickey-Fuller Test-INF (at Level)

Null Hypothesis: INF has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=3)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.173944	0.0398
Test critical values:	1% level	-3.886751	
	5% level	-3.052169	
	10% level	-2.666593	

Source: EViews-12

Interpretation

The result in Table M2.2.3 above showed that t-statistic is also greater than 5 and 10 percent level of significance, and as such, significant at both levels. This was further confirmed by the p-value of 0.0398. One can then conclude that the variable INF is stationary at levels.

Using 5% levels of significant, one can summarize Tables M2.2.1 to M2.2.3 as follows:

Table M2.2.4: Summary of Unit Root Test Results

	Levels			**First difference**			**Order of Integration**
Variables	ADF	Critical Value (5%)	P-value	ADF	Critical Value (5%)	P-value	
GDP	-2.113029	-3.052169	0.2422	-3.679821**	-3.065585	0.0159	*I*(1)
INF	-3.173944**	-3.052169	0.0398				*I*(0)

Note: The tests include intercept with trend; ** implies significant at 5%.
Source: EViews-12

Self- Assessment 1

What are ways Non-stationary series be stationarised

1.6 Summary

In summary, understanding the concepts of unit root is crucial in econometric analysis. Stationarity ensures that the time series under study has constant statistical properties,

which is a prerequisite for many classical statistical models. Therefore, an appropriate application of these concepts aids in generating reliable forecasts and valuable insights into the dynamics of economic variables.

In this unit you learnt the answer to the question "What is Stationarity?". You also learnt how to conduct unit roots test in the AR(1) Models and stationarity of the variables using Eviews software.

Tutor Marked Assignment

Interpret the table for the Augmented Dickey Fuller statistics

Null Hypothesis: GDP has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=1)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-2.196300	0.4378
Test critical values:	1% level	-5.521860	
	5% level	-4.107833	
	10% level	-3.515047	

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Ilorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Ezie, O. (2022). A Practical Guide on Data Analysis Using EViews. Kabod Limited Publisher, Kaduna.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Green, W.H. (2012). Econometric Analysis. (7th edition). Pearson Education Limited. England

Gujarati, D.N. (2005). Basic Econometrics. (4th edition). Tata McGraw-Hill Publishing Company Limited. New Delhi

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

There are several common techniques for transforming non-stationary time series data to make it stationary:

1. **Differencing**: This involves computing the difference between consecutive observations. In some cases, more than one difference may be necessary. This is commonly used for series that exhibit a trend.
2. **Seasonal Differencing**: For a series with a seasonal pattern, you might need to compute the difference between an observation and a previous observation from the same season. For example, in monthly data with a yearly pattern, you might take the difference between each observation and the observation from 12 months earlier.
3. **Transformation**: Mathematical transformations, like taking the logarithm or square root of the series, can help stabilize a changing variance. This is commonly used for series where the variance increases or decreases with the level of the series.
4. **Detrending**: If the series has a deterministic trend (i.e., the trend does not change over time), you might remove this trend from the data. This could involve fitting a trend model (like a linear regression on time) and then subtracting the fitted trend from the original series.
5. **Decomposition**: Some series can be decomposed into trend, seasonal, and residual components. The residual component might be stationary and can be analyzed separately from the trend and seasonal components.
6. **Moving Average**: If there are short-term fluctuations that make a series appear non-stationary, a moving average can be used to smooth out these fluctuations and expose longer-term patterns or trends.

UNIT 3: COINTEGRATION ANALYSIS

1.1 Introduction

In econometric analysis, especially with time series data, the concepts of stationarity and cointegration are critical for modeling and forecasting. While stationarity is a property of a single time series, cointegration involves a long-run relationship between two or more non-stationary time series.

The concept of stationarity suggests that the statistical properties of a process generating a time series do not change over time. It implies that the mean, variance, and autocorrelation structure are constant throughout time. Most of the classical linear statistical models are based on the assumption of stationarity. Non-stationary time series,

on the other hand, have statistical properties that change over time, which can lead to unreliable and spurious statistical results.

Economically speaking, two variables will be cointegrated if they have a long-term, or equilibrium relationship between them. After conducting the stationarity test (or unit root test) on the times series, it is imperative to ascertain if the variables have long-run relationship within them.

1.2 Learning Outcomes

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

- Discuss the meaning of cointegration
- Learn how to conduct cointegration Test
- Discuss the types of cointegration tests
- Learn how to conduct cointegration test using EViews

1.3 What is Cointegration?

Economically speaking, two variables will be cointegrated if they have a long-term, or equilibrium relationship between them. After conducting the stationarity test (or unit root test) on the times series, it is imperative to ascertain if the variables have long-run relationship within them. For an equilibrium or long-run relationship to exist, what is required is a linear combination of Y_t and X_t that is a stationary variable (an $I(0)$ variable); that is, 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.

1.4 Cointegration Tests

While non-stationarity poses challenges, it is a common feature in many time series datasets, especially in economics where variables often exhibit trends. This is where the concept of cointegration becomes crucial.

Cointegration is a statistical property of two or more time series that suggests a long-term relationship among them. In simple terms, if two or more series are individually non-stationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

To illustrate, consider two economic time series, say, the GDP and consumption expenditure, both typically non-stationary because they tend to grow over time. However, if there exists a stable, long-run relationship between the two, then even if they stray apart in the short run due to various shocks, they will converge back to equilibrium in the long run. This equilibrium relationship between GDP and consumption makes them cointegrated.

Testing for cointegration is an essential part of econometric analysis with time series data. The most commonly used test is the Engle-Granger two-step method. Another widely-used procedure is the Johansen test, which allows for more than one cointegrating relationship.

Cointegration has significant implications in modeling and forecasting. It forms the basis for Error Correction Models (ECM) that combine the short-run dynamics of the variables with their long-run equilibrium relationship. ECM allows for better model specification, improved forecasting, and meaningful economic interpretation.

After unit root testing, what's next?

The outcome of unit root testing matters for the empirical model to be estimated. The following cases explain the implications of unit root testing for further analysis. Ezie an Ezie (2021) highlighted three cases and they are discussed as follows;

CASE 1: Series in the model under examination are stationary.

- That is, series are stationary. Technically speaking, one meant, they are $I(0)$ series (integrated at order zero).
    - Under this scenario, cointegration test is not required, as any shock to the system in the short run quickly adjusts to the long run. Therefore, only the long run model should be estimated.
    - In econometrics, a long run model is a static model where variables are neither lagged nor differenced.
    - Thus, the estimation of short run model is not necessary if series are $I(0)$.

It has been shown from the foregoing that if a linear combination of non-stationary series is found to be stationary (cointegrated), one would be correct to go ahead and interpret the results of a static model. The Error Correction Mechanism is a means of reconciliating the short-run behaviour of an economic variable to its long-run behaviour. ECM was first used by Hendry, Pagan and Sargan (1984) and later popularized by Engle and Granger. The ECM is important and popular for many reasons. Some of the reasons are discussed as:

i. it is a convenient model for measuring the correction of the disequilibrium of the previous period, with the present which has a very good economic implication.
ii. with 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.
iii. a very important advantage of ECMs is the ease with which they can fit into the general to specific approach to econometric modelling, which lead to, a search for the most parsimonious ECM model that best fits the given data sets.

1.5 Types or Testing for Cointegration

A lot of econometricians have presented different methods of testing for cointegration; and they include Engle-Granger Cointegration, Johansen Cointegration Test and Bounds Co-integration Test

1.5.1 Engle-Granger Cointegration

This involves Testing for Cointegration for $I(1)$ series in single-equation models. This test was developed by Engle and Granger (1987) [EG thereafter], and they show that if after using either DF or Augmented DF (ADF) unit root test, the variables in the regression model are $I(1)$ and the residual component obtained from the regression is $I(0)$, then there is a linear combination (long-run relationship or equilibrium) between or among the variables in the model (Ezie and Ezie, 2021).

Empirical Illustration

Table below shows the result of Engle and Granger (1987) (or residual-based) co-integration test generated using ADF test method and its interpretation:

Table M2.3.1: Results of Engle and Granger Residual Based Cointegration Test

Variable	ADF Test Statistic	95% Critical ADF Value	Order of Integration	Remarks
Residual	-6.2116	-4.2061 (0.00396)	$I(0)$	Stationary

Source: Ezie and Ezie, 2021

In Table M2.3.1 above, the ADF test statistic value of -6.2116 is greater than the 95 percent critical ADF value of -4.2061 (in absolute values). This was further collaborated by the p-value of 0.00396 that was found to be less than 0.05 (that is, 0.00396 < 0.05). This clearly indicates that the residuals are stationary. Thus, one cannot reject the hypothesis of cointegration among the variables in the study; it implies that, a long run relationship exists amongst the variables.

1.5.2 Johansen Cointegration Test

Cointegration, a fundamental concept in econometric analysis of time series data, is a

statistical property of a collection of time series variables which have some long-term, equilibrium relationship between them, even though the individual series themselves may be non-stationary. The Johansen cointegration test, developed by Søren Johansen, is a method that determines the cointegration relationships among several univariate series.

The Johansen cointegration test is based on the vector error correction model (VECM), where changes in variables are expressed as a function of deviations from the long-term equilibrium. The primary advantage of the Johansen methodology is its capacity to test and estimate multiple cointegrating vectors, rendering it useful for systems involving more than two variables.

In the Johansen test, the null hypothesis states that the number of cointegrating vectors is at most r, where r is less than the total number of variables in the system. The test provides two statistics: the trace statistic and the maximum eigenvalue statistic. The trace test evaluates the null hypothesis of r cointegrating relations against the alternative of more than r relations, whereas the maximum eigenvalue test checks the null hypothesis of r relations against the alternative of r+1 relations.

These two test statistics have different power properties and may occasionally yield contradictory results. As such, the researcher needs to interpret the results with caution, taking into account the context and the potential consequences of Type I and Type II errors.

An important consideration in applying the Johansen cointegration test is that the data series should be integrated of the same order. Moreover, the test assumes that the underlying data-generating process does not suffer from structural breaks. If these assumptions are not satisfied, the results of the test can be misleading.

- Johansen co-integration test uses two test criteria namely; the trace statistic and the maximum eigenvalue, both generated with ML technique.
- The null hypothesis is that there is no cointegration between/among the series under examination.
- If co-integration test is carried out and the result shown with trace statistic and maximum eigenvalue suggests that the null hypothesis of no co-integration should be rejected, then, it implies that co-integration exists and as such, there is long run relationship between/among the variables in the model.
- If otherwise, there is no long run relationship in the model.

Empirical Illustration and Interpretations of Johansen Cointegration

The results of the cointegration tests are extracted into Table below:

Table M2.3.2: Johansen Cointegration Tests Result

Date: 06/17/21 Time: 19:11 Sample (adjusted): 1988 2017 Included observations: 30 after adjustments Trend assumption: Linear deterministic trend Series: GDP EXR MS PSC Lags interval (in first differences): 1 to 1
Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.535757	57.64755	47.85613	0.0046
At most 1 *	0.440426	34.62715	29.79707	0.0129
At most 2 *	0.352221	17.20977	15.49471	0.0273
At most 3 *	0.130167	4.183609	3.841466	0.0408

Trace test indicates 4 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	0.535757	23.02040	27.58434	0.1726
At most 1	0.440426	17.41738	21.13162	0.1532
At most 2	0.352221	13.02616	14.26460	0.0777
At most 3 *	0.130167	4.183609	3.841466	0.0408

Max-eigenvalue test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Source: Ezie and Ezie, 2021

The Trace test statistics in Table M2.3.2 indicate, that the hypothesis of no cointegration, H_0, among the variables can be rejected. The Trace test results revealed that four cointegrating vectors exist among the variables of interest. However, the Max-eigenvalue test indicates no cointegration at the 0.05 level. Thus, one has a case of conflicting results between the two tests. The study thus adopts the conclusion made by Trace test which shows that, four cointegrating vectors exist among the variables of interest.

1.5.3 Bounds Co-integration Test

The Bounds Test is an approach to cointegration analysis in the context of a single-equation framework. Unlike other cointegration techniques like the Johansen test, the Bounds Test does not require variables to be of the same order of integration. This means it can be applied irrespective of whether the variables are I(0), I(1), or a combination of both, making it a more flexible approach to cointegration analysis.

The Bounds Test relies on the estimation of an unrestricted error correction model (UECM), which is an augmented version of the Autoregressive Distributed Lag (ARDL) model. This model captures both the long-run and short-run dynamics of the relationship between variables.

The null hypothesis of the Bounds Test is that there is no cointegration between the variables. This is tested against the alternative hypothesis of cointegration. The decision rule involves computing an F-statistic (or a t-statistic in some versions of the test) and comparing it to two critical value bounds. If the computed statistic is above the upper bound, the null hypothesis of no cointegration is rejected. If it falls below the lower bound, no conclusion can be drawn. If it falls between the bounds, the result is inconclusive.

One major advantage of the Bounds Test is its applicability in small sample sizes. Moreover, it is robust to the inclusion of deterministic components like intercepts and trends in the model. However, the test assumes no structural breaks in the data series, and any such breaks may bias the results.

When testing for cointegration where series are of different order of Integration, the appropriate cointegration test is the one proposed by Pesaran, Shin and Smith (2001) defined as bounds cointegration test.

Interpret your result appropriately using the following decision criteria

The three options of the decision criteria according to Ezie and Ezie (2021) are:

i. If the calculated **F-statistic** is greater than the **Critical Value Bounds** for the upper bound $I(1)$, then it can be concluded that, there is cointegration, because there is long-run relationship.

ii. If the calculated F-statistic falls below the theoretical critical value for the lower bound $I(0)$, it can be concluded that there is no cointegration, hence, no long run relationship.

iii. The test is considered inconclusive if the F-statistic falls between the lower bound $I(0)$ and the upper bound $I(1)$.

Table below presents the result of ARDL bounds test for Co-integration for the model using the recommended lag by AIC.

Table M2.3.3: Bound Test-Co-integration Results

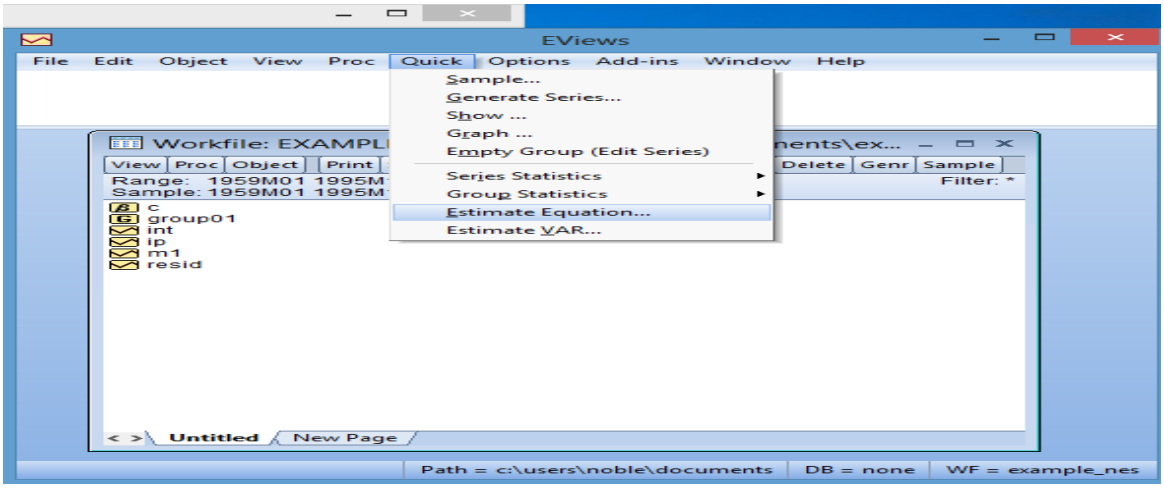
F-Bounds Test		Null Hypothesis: No levels relationship		
Test Statistic	Value	Signif.	I(0)	I(1)
F-statistic	7.288809	10%	2.37	3.2
k	3	5%	2.79	3.67
		2.5%	3.15	4.08
		1%	3.65	4.66

Source: EViews-12

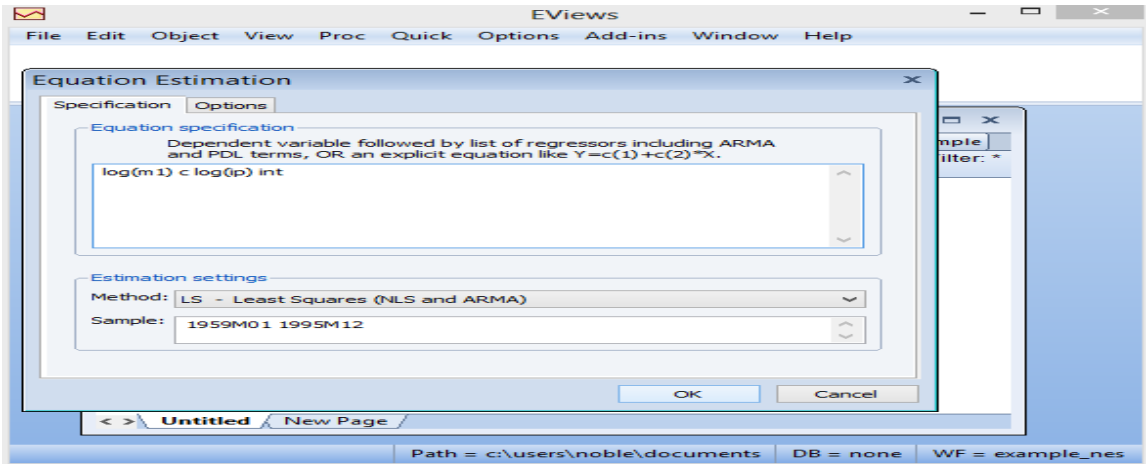
From the co-integration test captured in Table 6.3, it could be seen that F-statistic value of 7.288809 is greater than the lower ($I(0)$) and upper bound ($I(1)$) critical values of 2.79 and 3.67 respectively at the 5% significance level. It can therefore be inferred that, the variables are co-integrated, and as such, there is a long-run equilibrium relationship between dependent and the independent variables. Thus, the null hypothesis of no long-run relationship is rejected at the 5% significance level.

1.5.4 Cointegration Tests on EViews

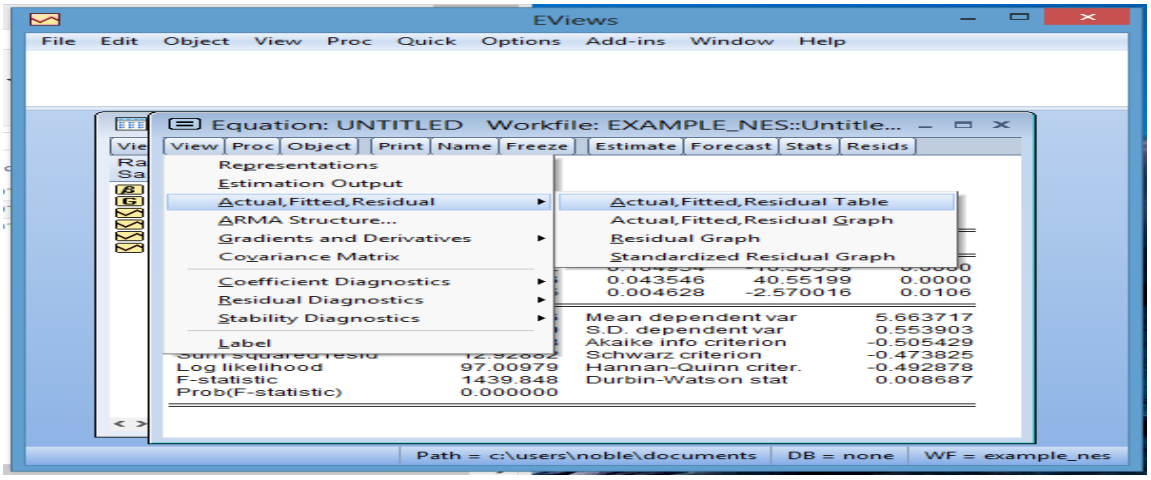
- For Engle-Granger cointegration test, choose **Quick/Estimate Equation…**



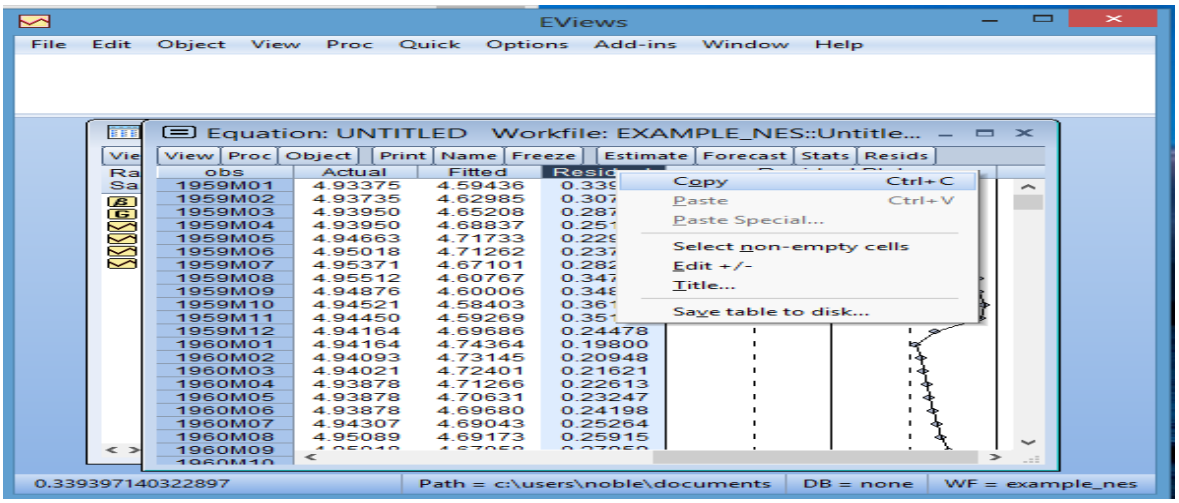
- Type the equation you want to test into the dialogue box that and click **Ok**



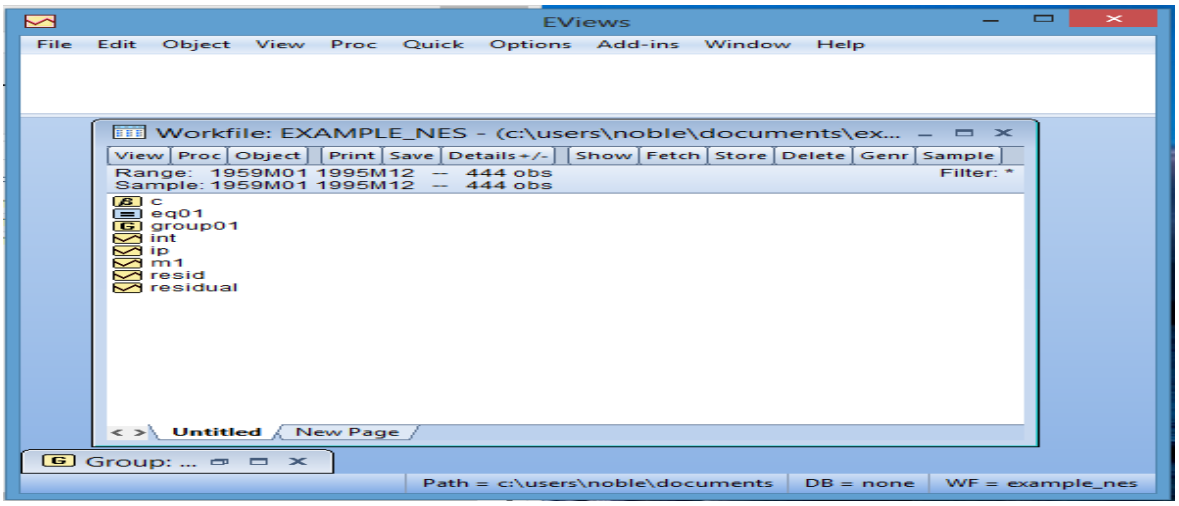
- Choose View/Actual, Fitted, Residual/Actual, Fitted, Residual Table



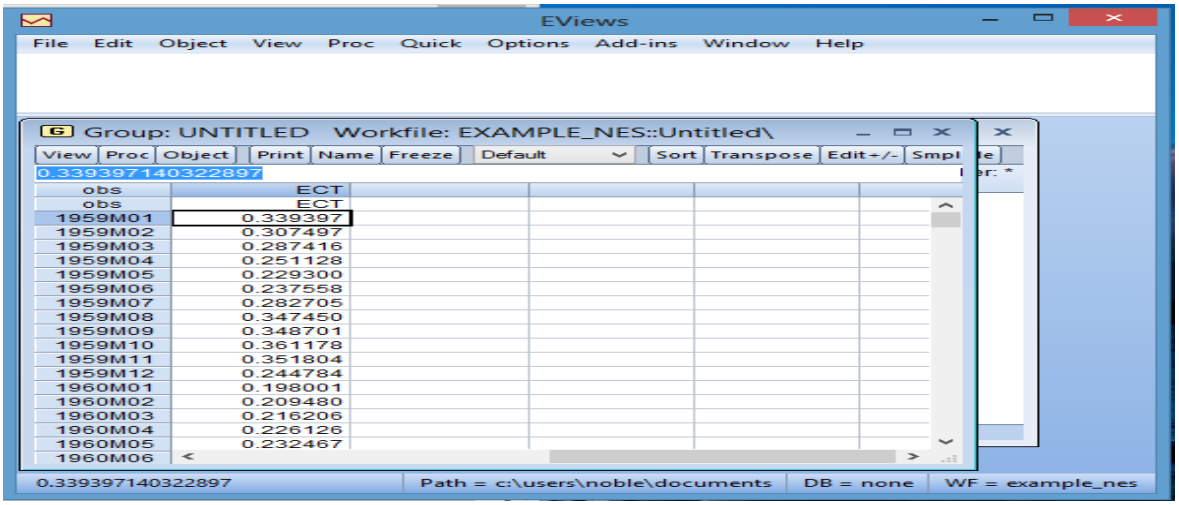
- Right click on **Residual** and select **Copy**



Choose **Quick/Empty Group (Edit Series)**

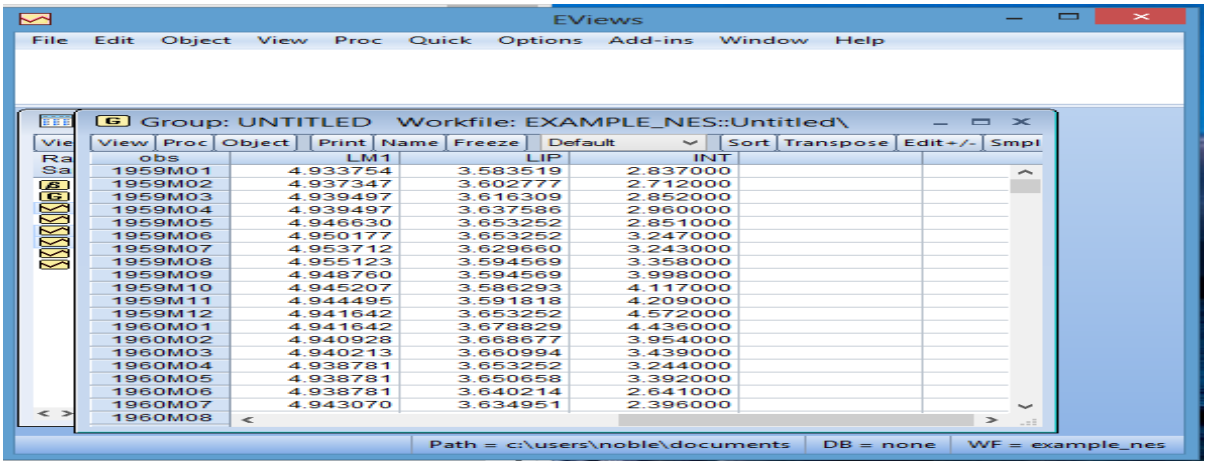


Paste the copied **Residual** into the work area that pop-up and change the series name to **"ect"**

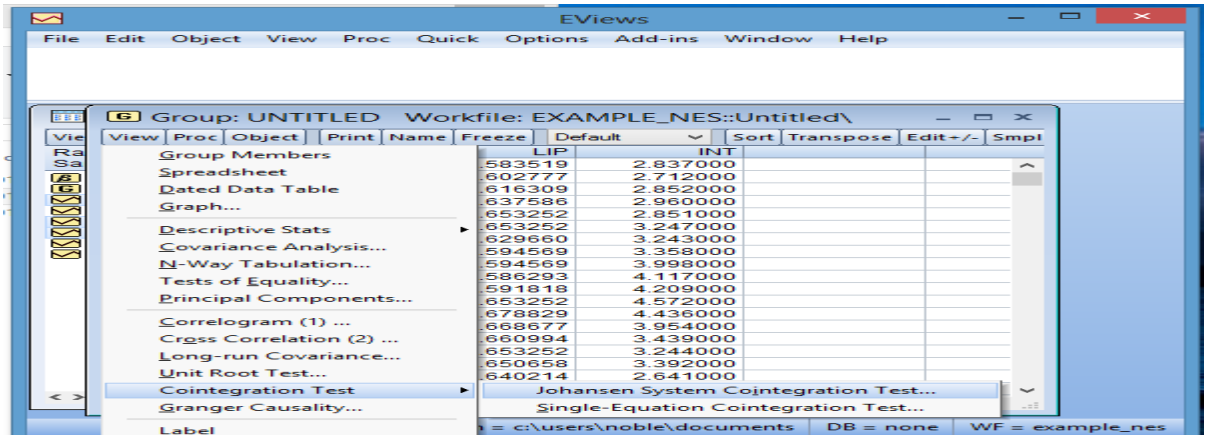


Test the **"ect"** series for unit root at **"level"** by double clicking on the series name to open the series window, and choose **View/Unit Root Test…**

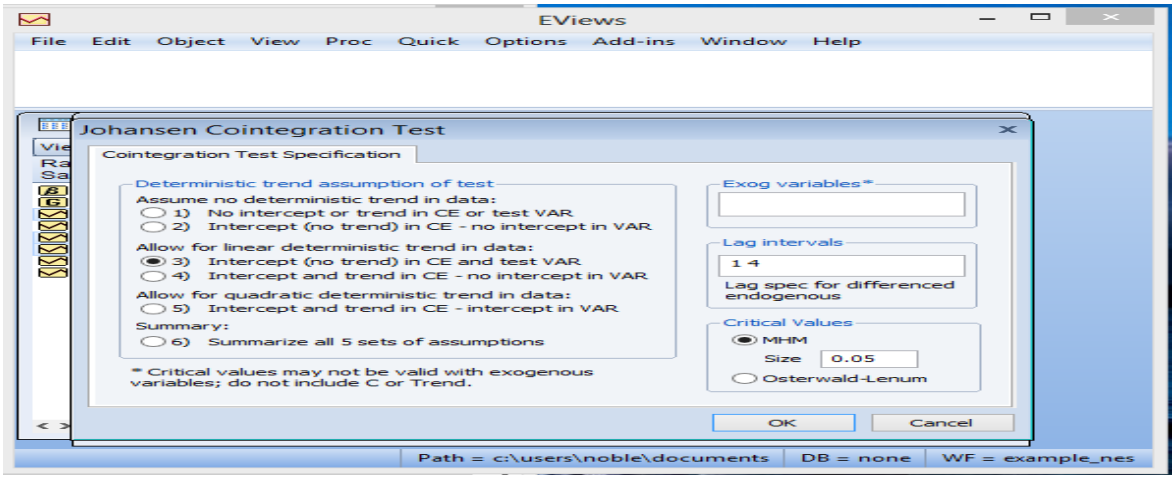
- For Johansen cointegration test, open the variables of interest as **Group**



Choose **View/Cointegration Test/Johansen System Cointegration Test…**



Click Ok



Self-Assessment Exercise 1

What is Cointegration?

1.6 Summary

Cointegration has become an overriding requirement for any economic model using non-stationary time series data. When the variables do not cointegrate, there will be problems of spurious regression and the econometric work becomes almost meaningless. On the other hand, if the stochastic trends do cancel, then it has cointegration. In this unit you learned what cointegration is and various types of cointegration test.

Tutor Marked Assignment

Using the result below, interpret if there is a cointegrating relation between X and Y at 10%:

F-Bounds Test		Null Hypothesis: No levels relationship		
Test Statistic	Value	Signif.	I(0)	I(1)
F-statistic	3.16552	10%	2.37	3.2
k	2	5%	2.79	3.67
		2.5%	3.15	4.08
		1%	3.65	4.66

1.7 References/Further Reading

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Narayan, P. K. (2005). The saving and investment nexus in China: evidence from cointegration tests. Applied Economics, 37, 1979 – 1990.

Pesaran, M., Shin, Y. & Smith, R.. (2001). Bound testing approaches to the analysis of level relationship. J. Appl. Econ. 16, 289–326.

Pesaran, H. and Shin, Y. (1999). An autoregressive distributed lag modeling approach to cointegration analysis. In: Strom, S. (Ed.), Econometrics and Economic Theory in 20th Century: The Ragnar–Frisch Centennial Symposium. Cambridge University Press: Cambridge.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

Cointegration is a statistical concept used in the field of time series analysis. It deals with the long-term relationship between two or more time series variables.

Specifically, if two or more time series are individually non-stationary (i.e., they have trends or varying variances over time), but a linear combination of them is stationary, then the series are said to be cointegrated. In other words, even though the individual series may wander over time, there is a consistent, long-run relationship between them that brings them back towards an equilibrium.

Cointegration has important implications in the modeling and prediction of time series data. It is often used in finance and economics to find pairs of assets whose prices move together in the long run, even if they might diverge in the short run. For example, if the prices of two different stocks are cointegrated, this might suggest that they are influenced by similar market or economic factors.

The concept of cointegration was developed by Clive Granger, who won the Nobel Prize in Economics for this and related work. A commonly used test for cointegration is the Engle-Granger two-step method, while the Johansen test allows for testing the cointegration of more than two series.

UNIT 4: AUTOREGRESSIVE DISTRIBUTED LAG (ARDL) MODEL

1.1 Introduction

In the present unit we shall discuss autoregressive distributed lag commonly called ARDL models. Autoregressive Distributive Lagged (ARDL)-Bounds testing approach. This test was developed by Pesaran and Shin (1999) and later extended by Pesaran, Shin and Smith (2001). The ARDL bound test has superiority over Johansen (1991) and Engle and Granger (1987) cointegration approaches because for a number of reasons The endogeneity problems and inability to test hypotheses on the limited coefficients in the long run associated with the Engle-Granger Method are avoided.

1.2 Learning Outcomes

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

- Discuss ARDL Cointegration Equations
- Explain the justification for the choice of ARDL model
- Compute ARDL in Eviews
- Design the ARDL Bound cointegration test model:
- Conduct a diagnostic check for Serial Correlation in ARDL

1.3 The Meaning of ARDL

Time series analysis plays an indispensable role in econometrics, offering tools and techniques to investigate the relationships and dynamics of variables over time. A pivotal component in this regard is the Autoregressive Distributed Lag (ARDL) model, a staple in time series econometrics known for its flexibility and versatility in handling a broad range of econometric scenarios.

The ARDL model is a type of dynamic regression model, capturing relationships between variables across time. It is called 'autoregressive' because it includes lagged values of the dependent variable as regressors, and 'distributed lag' because it involves lagged values of one or more independent variables.

ARDL Cointegration Equations

In particular, if y_t is the dependent variable and $x_1,...x_k$ are k explanatory variables, a general ARDL $(p,q_1,q_2,.....q_k)$ model is given by:

$$y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q_j} \beta_j, i_j x_j, _{t-1j} + \varepsilon_t$$

The co-integrating regression form of the (restricted) ARDL model in is specified as:

$$\Delta y_t = -\sum_{i-1}^{p-1} \gamma_i^* \Delta y_{t-1} + \sum_{j=1}^{k} \sum_{i=0}^{q_j-1} \vartheta_j \Delta x_{j,t-i'} - \partial ect_{t-1} + \varepsilon_t$$

Where; Δ is the first difference operator; ect_{t-1} is the lagged error correction term; ∂ is the co-efficient of the error correction term.

The long run equation can be written as follows:

$y_t = \alpha + \sum_{i=1}^{k} \beta_i x_i + \varepsilon_t$ The bounds test procedure thus transforms into the following representation:

$$\Delta y_t = -\sum_{i-1}^{p-1} \gamma_i^* \Delta y_{t-i} + \sum_{i=1}^{k} \sum_{i=0}^{q_j-1} \beta_j \Delta x_{j,t-i} - \rho y_{t-1} - \alpha_1 - \sum_{j=1}^{k} \delta_j x_{j,t-1} + \varepsilon_t$$

The bound test for the existence of long-run relationships is simply a test of:

$H_0 : \delta_1 = \delta_2 = ... = \delta_k = 0$

$H_1 : \delta_1 = \delta_2 = ... = \delta_k \neq 0$

The three options of the decision criteria according to Ezie and Ezie (2021) are:

i. If the calculated **F-statistic** is greater than the **Critical Value Bounds** for the upper bound $I(1)$, then it can be concluded that, there is cointegration, because there is long-run relationship.

ii. If the calculated F-statistic falls below the theoretical critical value for the lower bound $I(0)$, it can be concluded that there is no cointegration, hence, no long run relationship.

iii. The test is considered inconclusive if the F-statistic falls between the lower bound $I(0)$ and the upper bound $I(1)$.

1.4 Justification for the Choice of ARDL Model

The ARDL model is an integral part of time series econometrics due to several reasons:

i. **Dynamic Relationship:** The ARDL model enables the examination of the dynamic relationship between variables. It accounts for both the immediate and lagged effects of changes in independent variables on the dependent variable. This characteristic is particularly useful in economics, where the full impact of a change in one variable on another often unfolds over time.

ii. **Long-run and Short-run Analysis:** The ARDL model and its error correction variant (ARDL-ECM) offer the ability to estimate both short-run and long-run coefficients within a unified framework. It provides valuable insights into the speed of adjustment and how quickly the variables converge to their long-run equilibrium after a shock. This understanding is crucial in many fields, including macroeconomics and finance.

iii. **Robustness and Flexibility:** The ARDL approach is flexible in terms of the integration order of variables; it can be applied regardless of whether the variables are stationary at level I(0), first difference I(1), or mutually integrated. This makes the ARDL a more versatile model, especially when compared to other cointegration tests like the Johansen test that require all variables to be integrated of the same order.

iv. **Small Sample Properties:** The ARDL model is well-suited to situations where the available sample size is small. It has better small-sample properties compared to alternative methods, ensuring more reliable estimations when data is limited.

v. **Cointegration Analysis:** With the ARDL bounds testing approach, it is possible to test for the existence of a long-run relationship (i.e., cointegration) among variables irrespective of their order of integration. This feature broadens the range of scenarios where ARDL can be beneficially employed.

Therefore, the use of ARDL models in econometric analysis can be justified by their capability to capture dynamic relationships, their flexibility in handling variables of different integration orders, their robustness in small samples, and their utility in cointegration analysis. While the choice of an econometric model should always consider the specific research question and the characteristics of the data, the ARDL model's strengths make it a powerful tool for many scenarios.

1.5 ARDL Computation in EViews

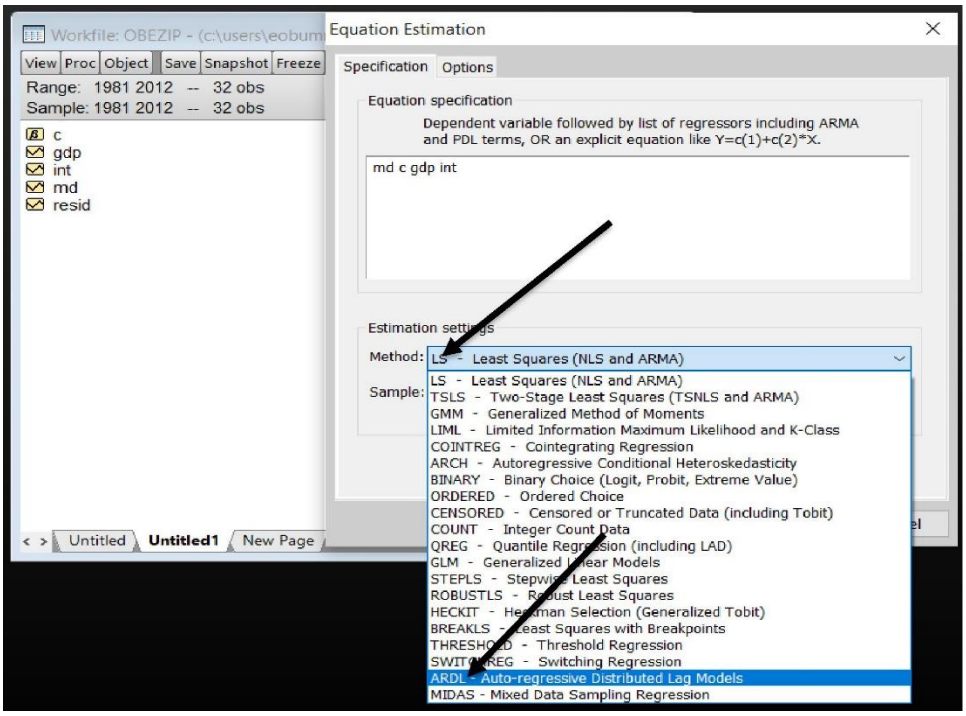
Since ARDL models are least squares regressions using lags of the dependent and independent variables as regressors, they can be estimated in EViews using an equation object with the Least Squares estimation method.

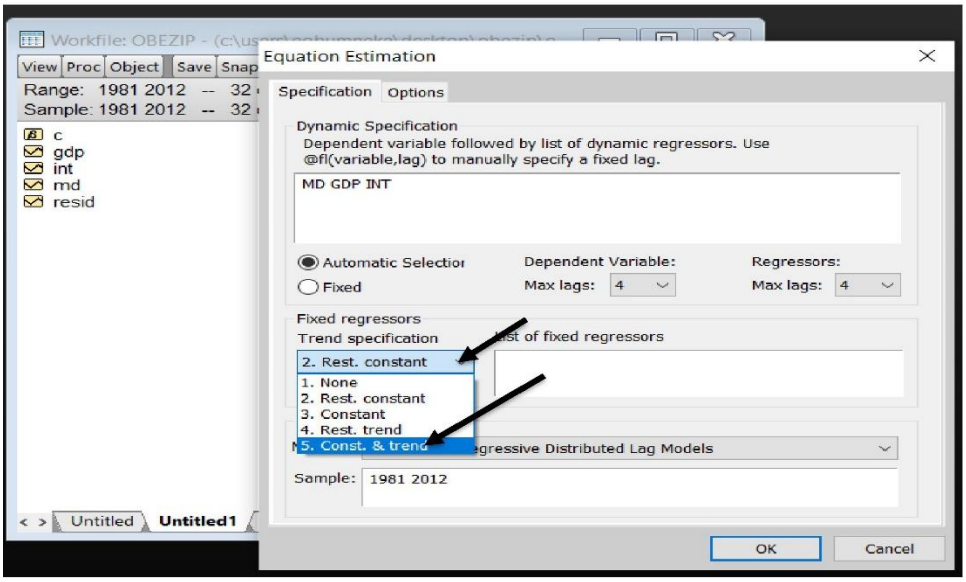
However, EViews also offers a specialized estimator for handling ARDL models. This estimator offers built-in lag-length selection methods, as well as post-estimation views. To estimate an ARDL model using the ARDL estimator, open the equation dialog by selecting Quick/Estimate Equation…, or by selecting Object/New Object…/Equation and

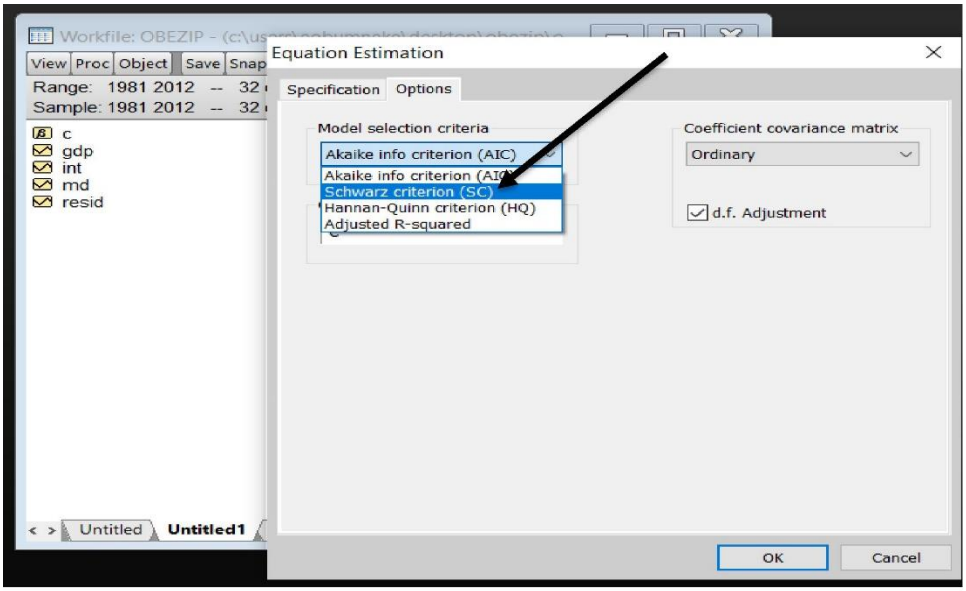
then selecting ARDL from the Method dropdown menu. EViews will then display the ARDL estimation dialog:

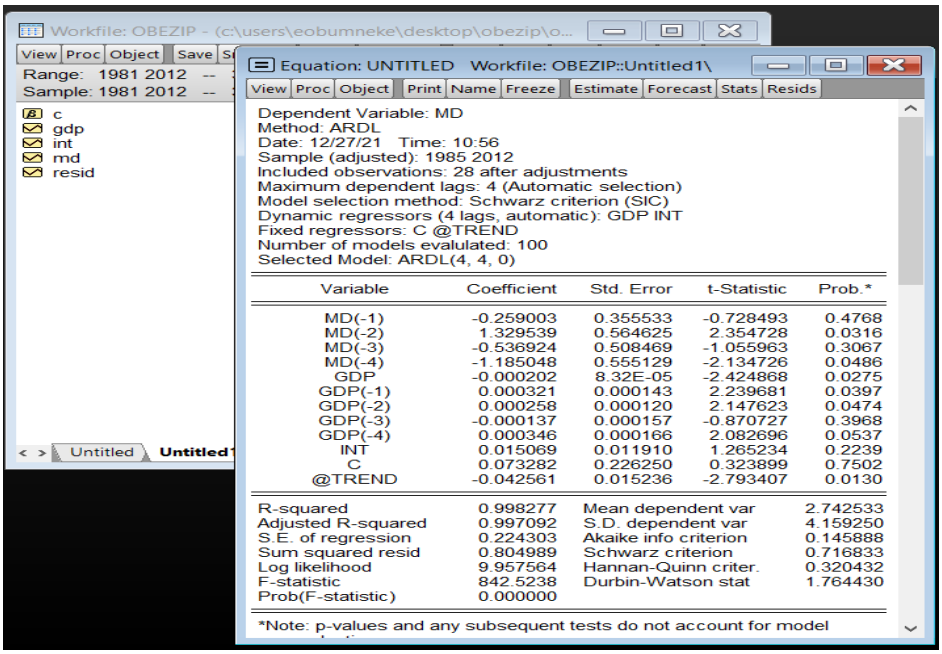
$$\Delta MD_t = \alpha_0 + \sum_{i=1}^{p} \alpha_1 \Delta MD_{t-i} + \sum_{j=0}^{q1} \alpha_2 \Delta GDP_{t-j} + \sum_{k=0}^{q2} \alpha_3 \Delta INT_{t-k} + \varphi_1 MD_{t-1} + \varphi_2 GDP_{t-1} + \varphi_3 INT_{t-1} + \varepsilon_t$$

Where MD is the Money Demand (N'Billion), GDP is the Gross Domestic Product (N'Billion) and INT is the Interest Rate (%).



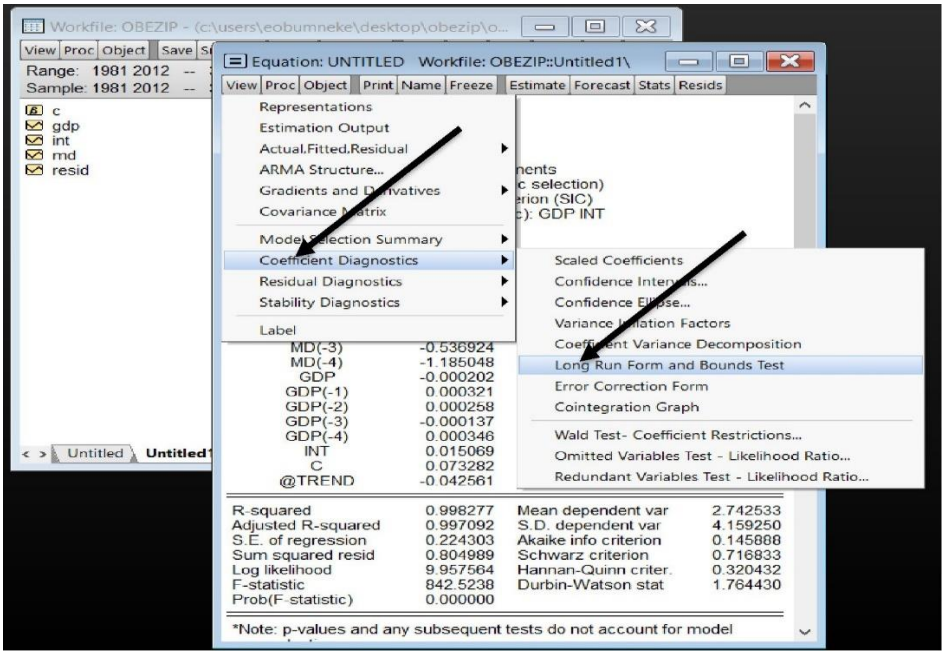

Workfile: OBEZIP
View Proc Object Save Snap
Range: 1981 2012 -- 32
Sample: 1981 2012 -- 32
c
gdp
int
md
resid
Untitled Untitled1
Equation Estimation
Specification Options
Dynamic Specification
Dependent variable followed by list of dynamic regressors. Use @fl(variable,lag) to manually specify a fixed lag.
MD GDP INT
Automatic Selectior
Fixed
Dependent Variable:
Max lags: 4
Regressors:
Max lags: 4
Fixed regressors
Trend specification
List of fixed regressors
2. Rest. constant
1. None
2. Rest. constant
3. Constant
4. Rest. trend
5. Const. & trend
gressive Distributed Lag Models
Sample: 1981 2012
OK
Cancel

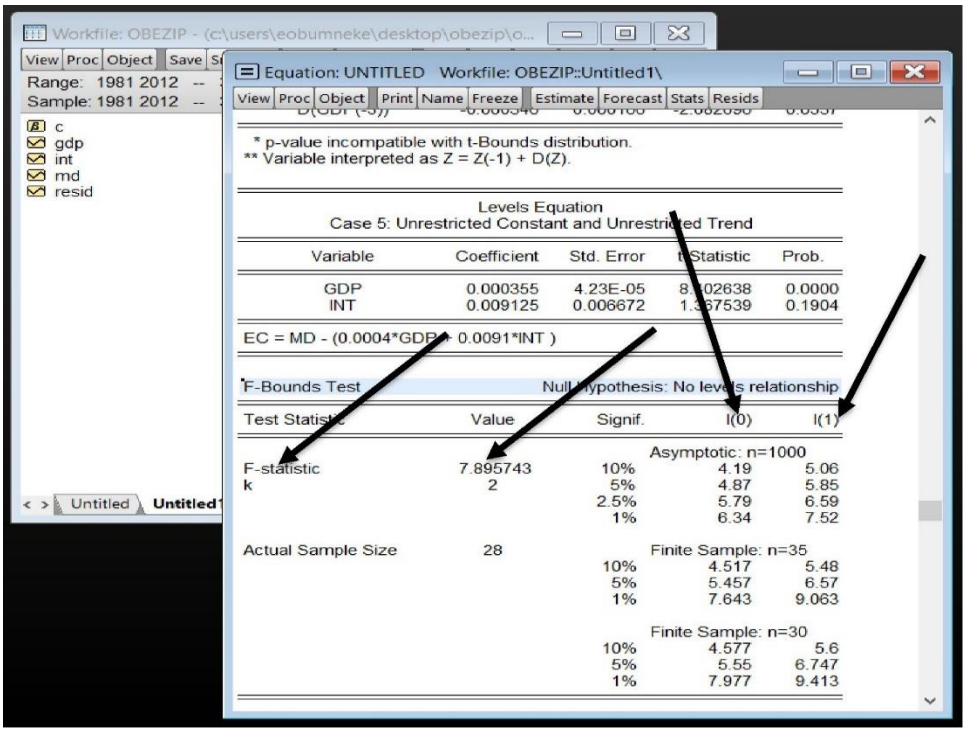

Workfile: OBEZIP
View Proc Object Save Snap
Range: 1981 2012 -- 32
Sample: 1981 2012 -- 32
c
gdp
int
md
resid
Untitled Untitled1
Equation Estimation
Specification Options
Model selection criteria
Akaike info criterion (AIC)
Akaike info criterion (AIC)
Schwarz criterion (SC)
Hannan-Quinn criterion (HQ)
Adjusted R-squared
Coefficient covariance matrix
Ordinary
d.f. Adjustment
OK
Cancel



ARDL Cointegration Bounds Test

After specifying the optimum lag model, we proceed to the ARDL Cointegration Bounds test.





Compare the F-statistic value with critical value provided by Pesaran et al. (2001). However, if the sample size is small (< 100 observations), then compare with the critical value provided by Narayan (2005).

In the case of the empirical illustration carried out, as shown in the result window, the F-statistic obtained (**7.895743**) falls above the lower bound $I(0)$ and upper bound $I(1)$. Hence, one concludes that, there is long run or equilibrium relationship amongst the variables, and one may consider both long run and short run models.

1.5.1 Diagnostic Check for Serial Correlation

Perform diagnostic check for serial correlation using the Breusch-Godfrey LM test

Select “View” – “Residual Diagnostics” – “Serial Correlation LM Test”.

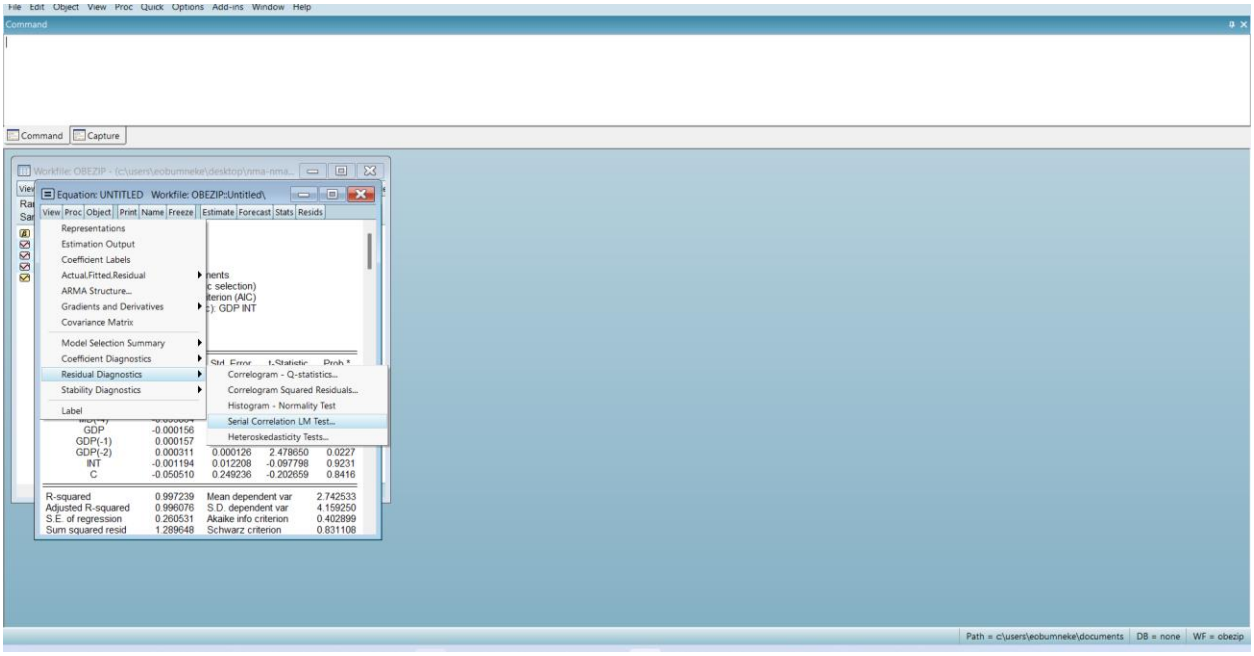


Table M.2.4.1

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	14.97137	Prob. F(2,17)	0.1782
Obs*R-squared	17.25274	Prob. Chi-Square(2)	0.1266

The null hypothesis of the test is that there is no serial correlation in the residuals up to the specified lag order. **EViews** reports a statistic labelled "F-statistic" and "Obs*R-squared" statistic. Both statistics indicate that there is no presence of serial correlation in the model.

Self-Assessment Exercise 1

What is ARDL Bound Test?

1.6 Summary

In this unit you learned ARDL cointegration equations as well as its computational procedures in Eviews. In addition, you learned a practical example on ARDL Bound cointegration test model and how to conduct diagnostic check for Serial Correlation in ARDL Cointegration Bounds test.

Tutor Marked Assignment

Use any data of your choice and conduct ARDL bound test.

1.7 References/Further Reading

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Narayan, P. K. (2005). The saving and investment nexus in China: evidence from cointegration tests. Applied Economics, 37, 1979 – 1990.

Pesaran, M., Shin, Y. & Smith, R.. (2001). Bound testing approaches to the analysis of level relationship. J. Appl. Econ. 16, 289–326.

Pesaran, H. and Shin, Y. (1999). An autoregressive distributed lag modeling approach to cointegration analysis. In: Strom, S. (Ed.), Econometrics and Economic Theory in 20th Century: The Ragnar–Frisch Centennial Symposium. Cambridge University Press: Cambridge.

1.8 Possible Answers to Self-Assessment Exercise(S) Within the Content

Answer to Self- Assessment 1

The ARDL (Autoregressive Distributed Lag) Bound Test is a methodology used to test whether a long-run level relationship (cointegration) exists between variables in a model, regardless of whether the underlying regressors are purely I(0), purely I(1), or fractionally integrated.

The approach was developed by Pesaran, Shin and Smith (2001) and is advantageous in situations where the researcher doesn't know whether the variables are stationary (integrated of order zero, or I(0)) or non-stationary (integrated of order one, or I(1)).

UNIT 5: ARDL POST ESTIMATION TESTS

1.1 Introduction

In the previous unit you learnt ARDL cointegration equations as well as its computational procedures in Eviews. In addition, you learnt a practical example on ARDL Bound cointegration test model.

The present unit which is incidentally is our last in this course is a continuation of ARDL model. It is required to verify whether the estimates from the ARDL model are reliable. These assumptions as earlier highlighted are linearity (using Ramsey Reset Test), homoscedasticity, normality, Stability test (using CUSUM test) among others. We have also demonstrated how to test for these assumptions in our previous presentations involving static regression models.

1.2 Learning Outcomes

- Should be able to know why post estimation tests are carried.
- Various types of post estimation tests required
- Interpret appropriate post estimation tests result
- How to conduct Hypotheses Testing Using Wald-Test

1.3 Post-Estimation Tests in ARDL

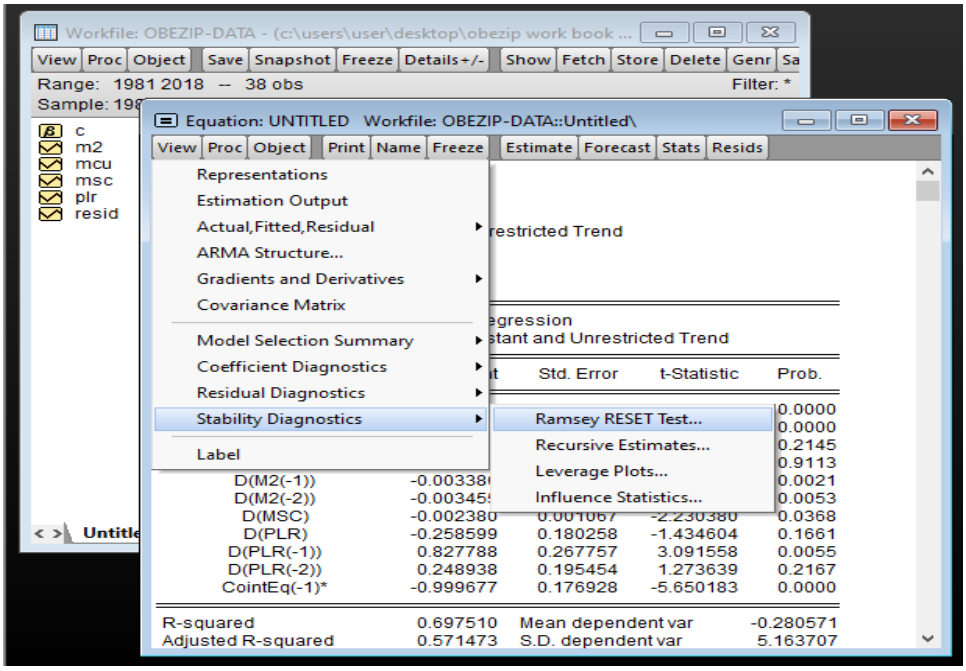
Post-estimation diagnostic tests serve to ensure the validity and reliability of the ARDL model's results. These tests primarily check the assumptions of the model.

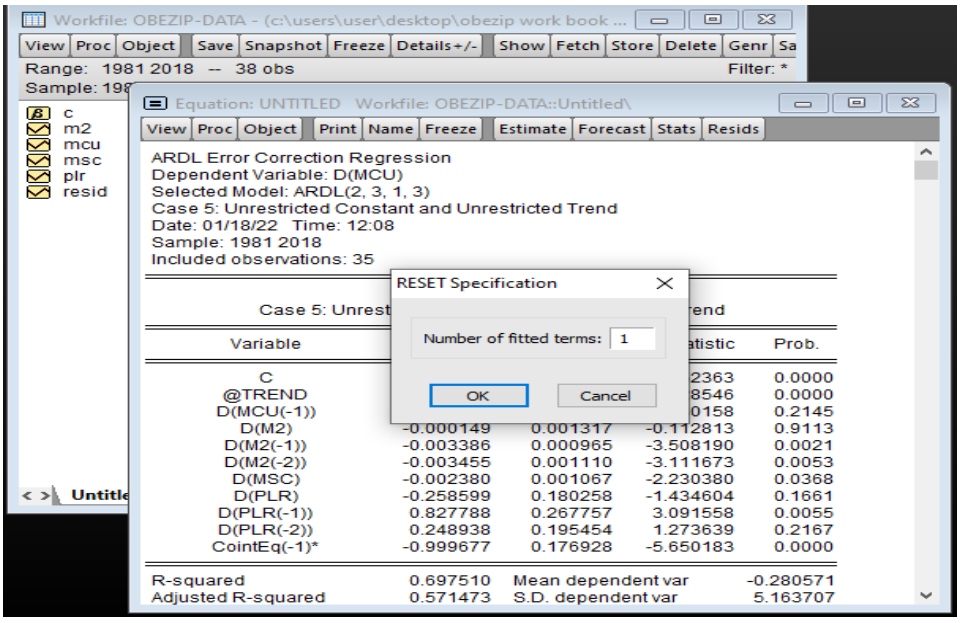
1.3.1 The Linearity Test

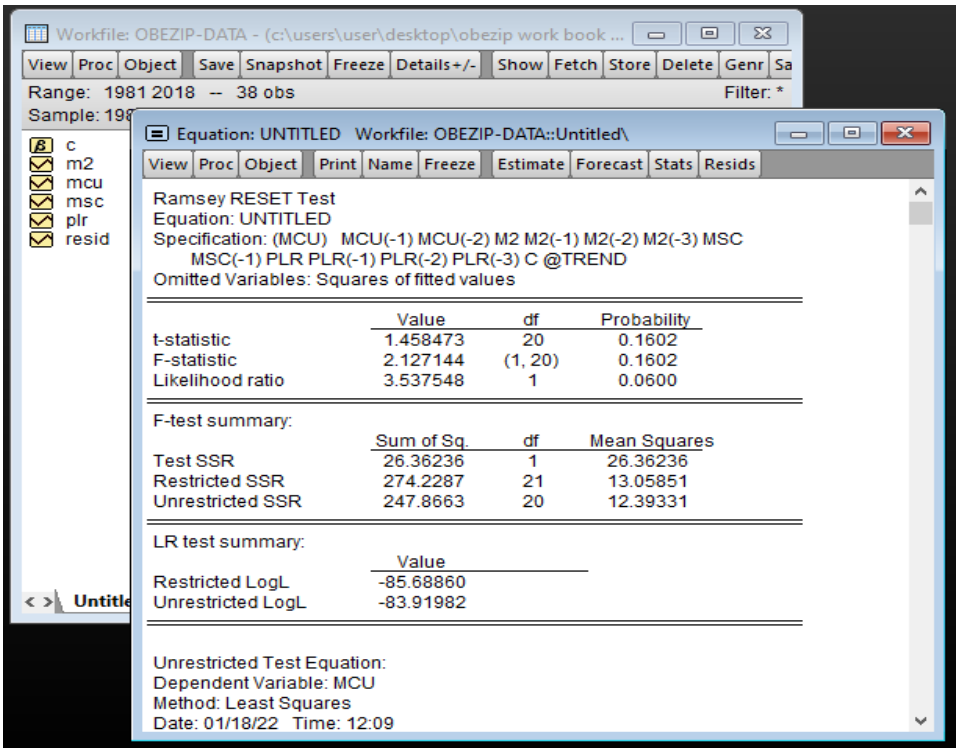
The test is meant to ascertain whether the model is linear or it is correctly specified. The essence is to find out if there is a linear relationship between the dependent variable and

the independent variables. The null hypothesis is that the model under consideration is linear or correctly specified.

To perform the test, select **Views/Stability Diagnostics** and then click **OK** on the RESET Specification. The step-by-step procedures for linearity are shown in the following figures:





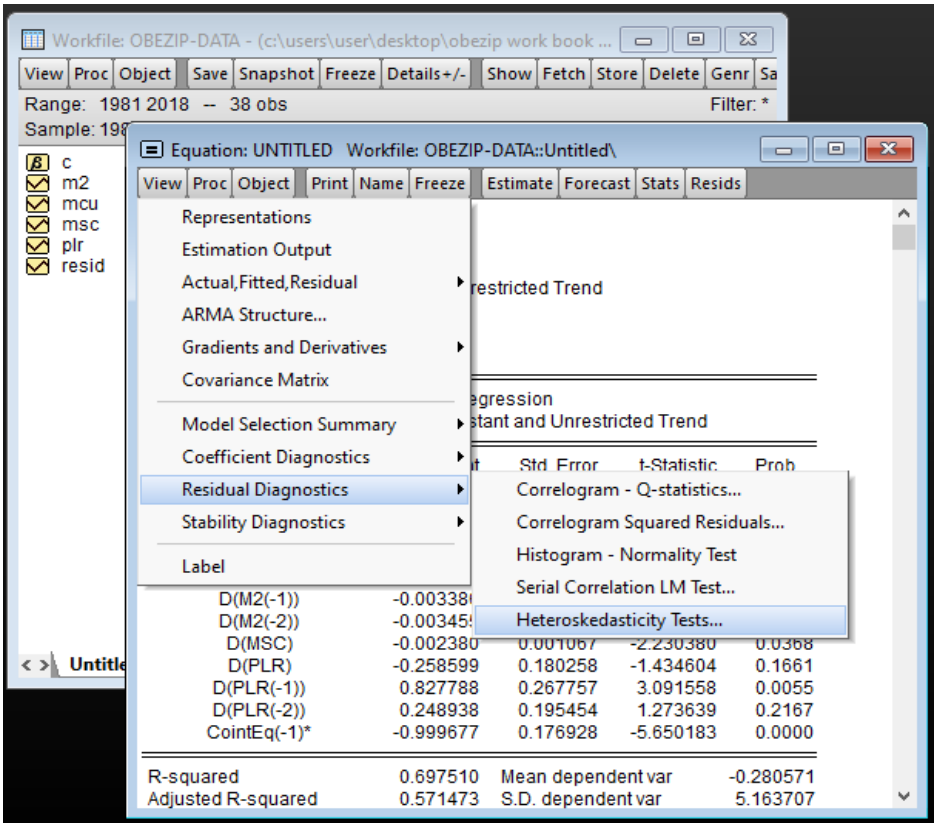


Output from the Ramsey reset test reports the test regression, the F-statistic and t-statistic for testing the hypothesis that the coefficients on the powers of fitted values from the regression are jointly zero, that is, the model is correctly specified. The null cannot be rejected since the p-value is more than 0.05.

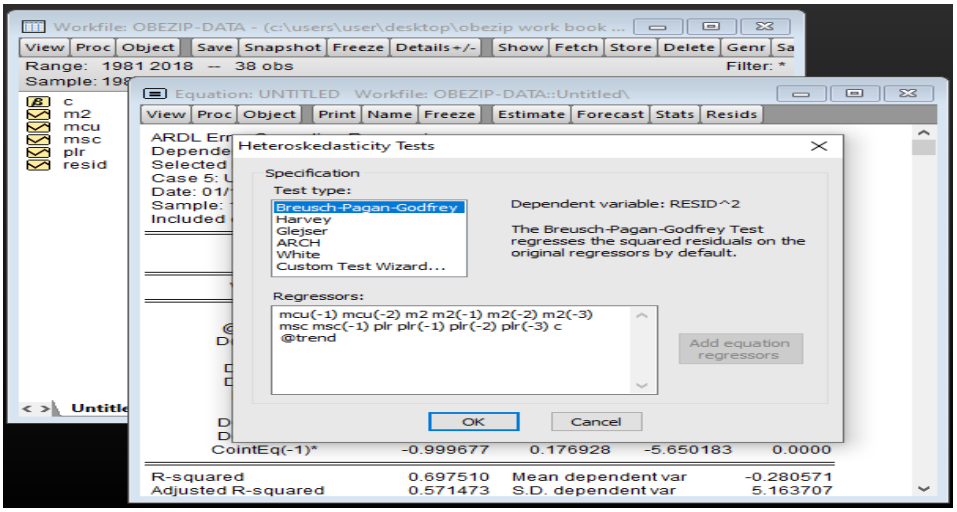
1.3.2 Heteroscedasticity Test

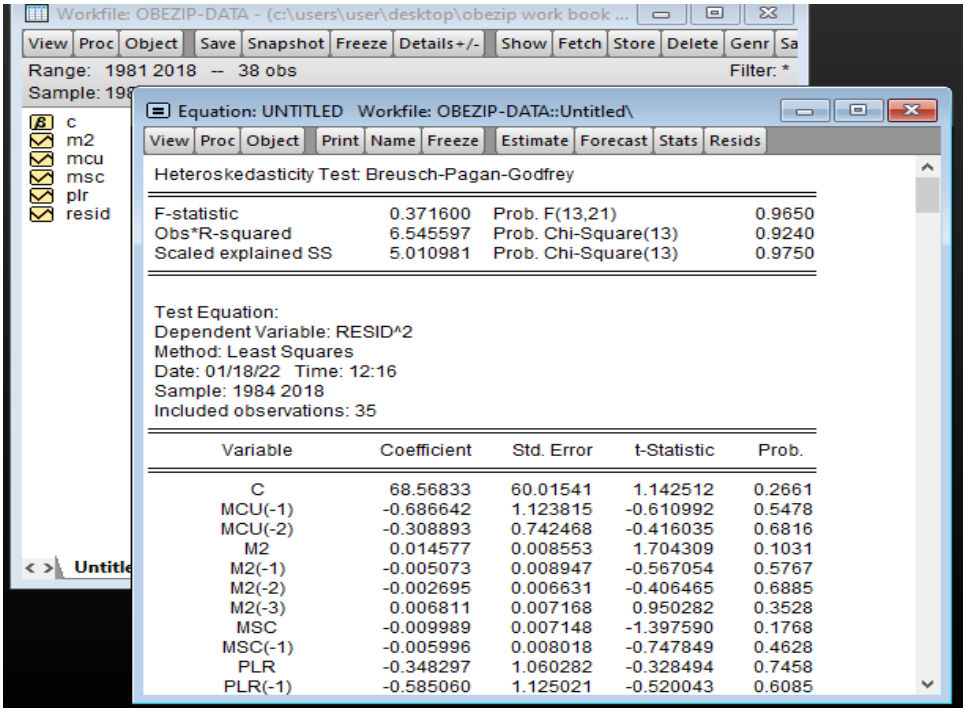
EViews offers variety of options for the Heteroscedasticity tests. For the purpose of our illustration, we consider the **Breusch-Pagan-Godfrey** test.

- Select **View/Residual Diagnostics/Heteroscedasticity Tests**.



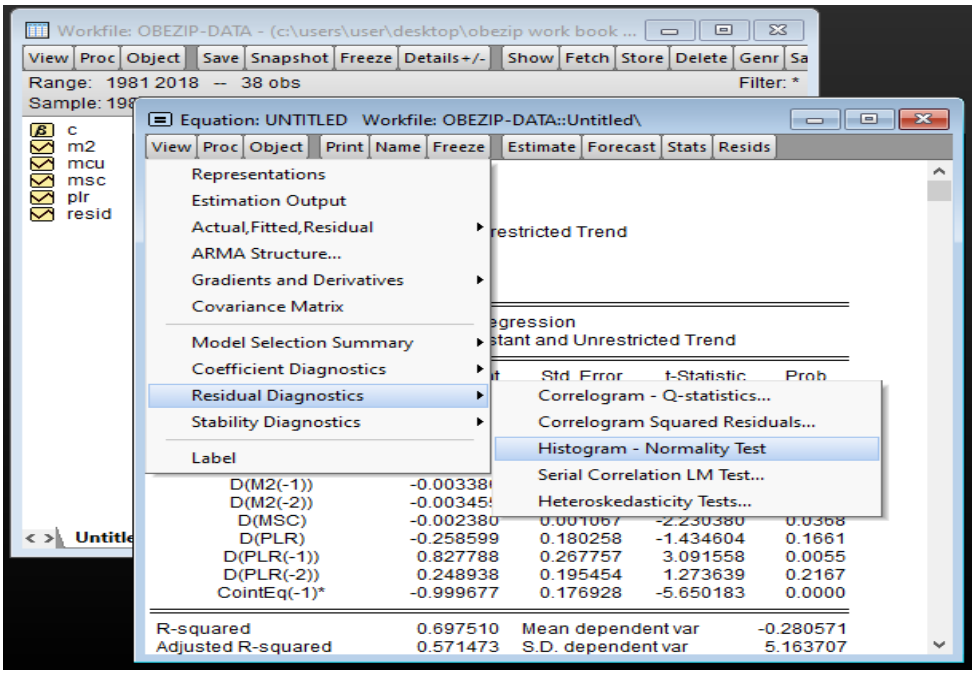
- Choose the default option by clicking on **Breusch-Pagan-Godfrey** in the **Test type box**. Click **OK** and the following results would appear:

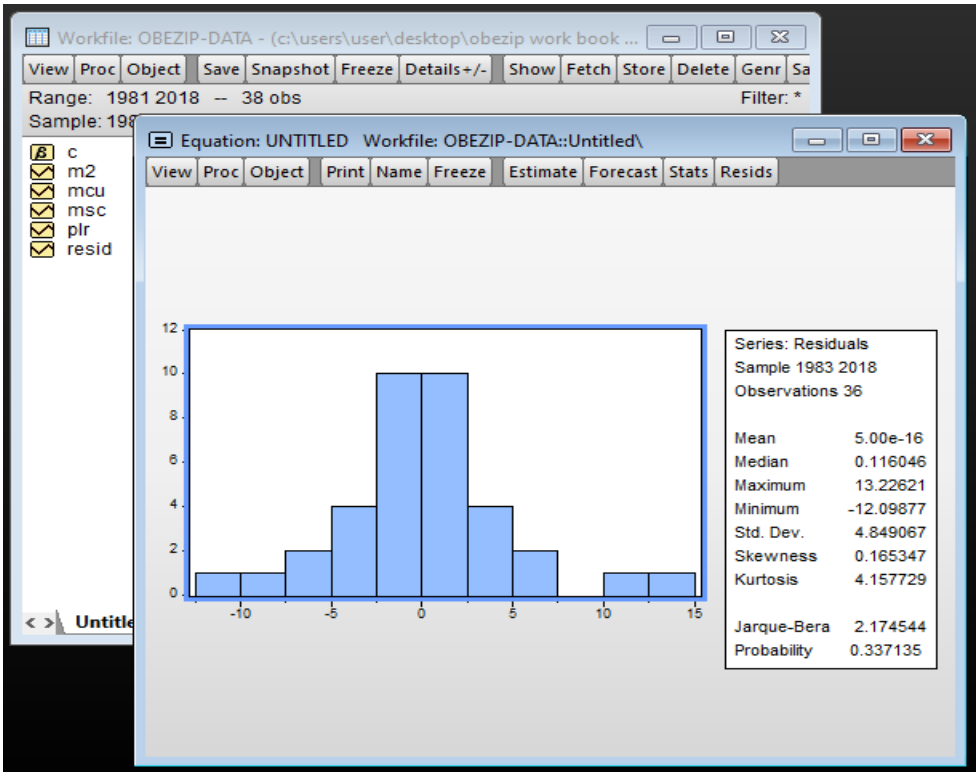




1.3.3 Normality Test

To perform the test, select ***View/Residual Diagnostic/Histogram Normality-Test***, the following result window is shown:

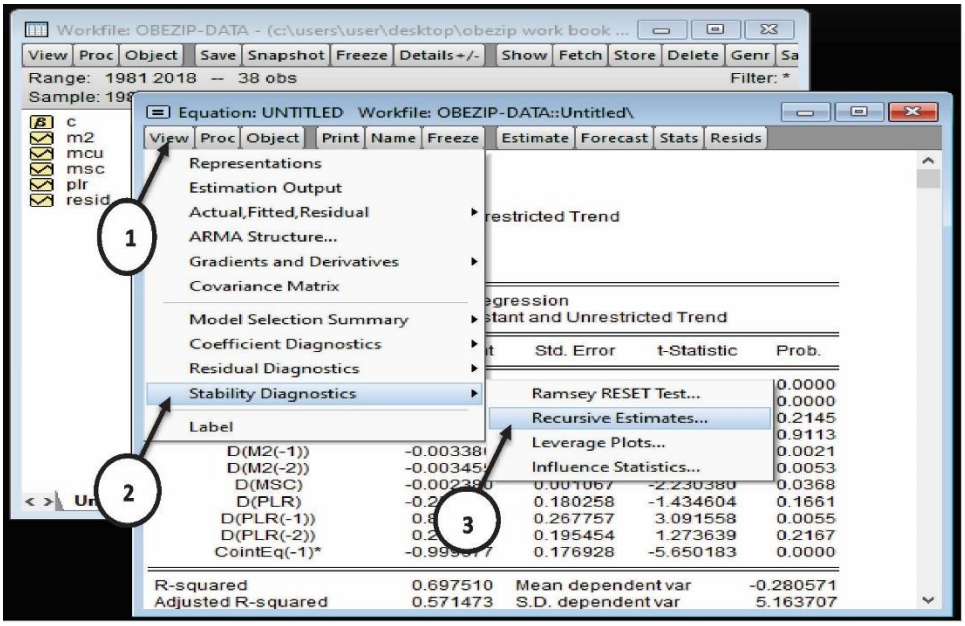


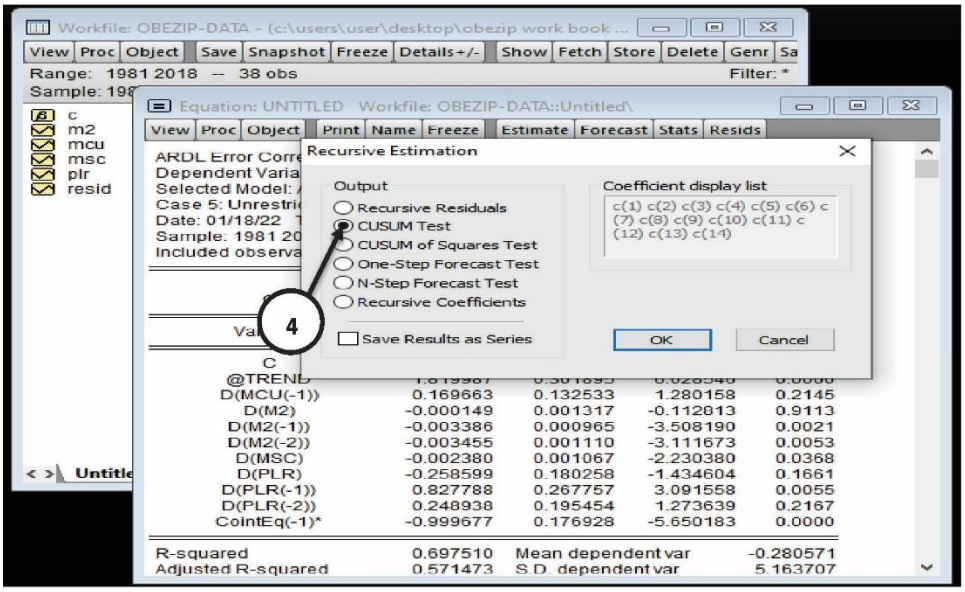


1.3.4 Stability Test (CUSUM and CUSUMQ Residual Test)

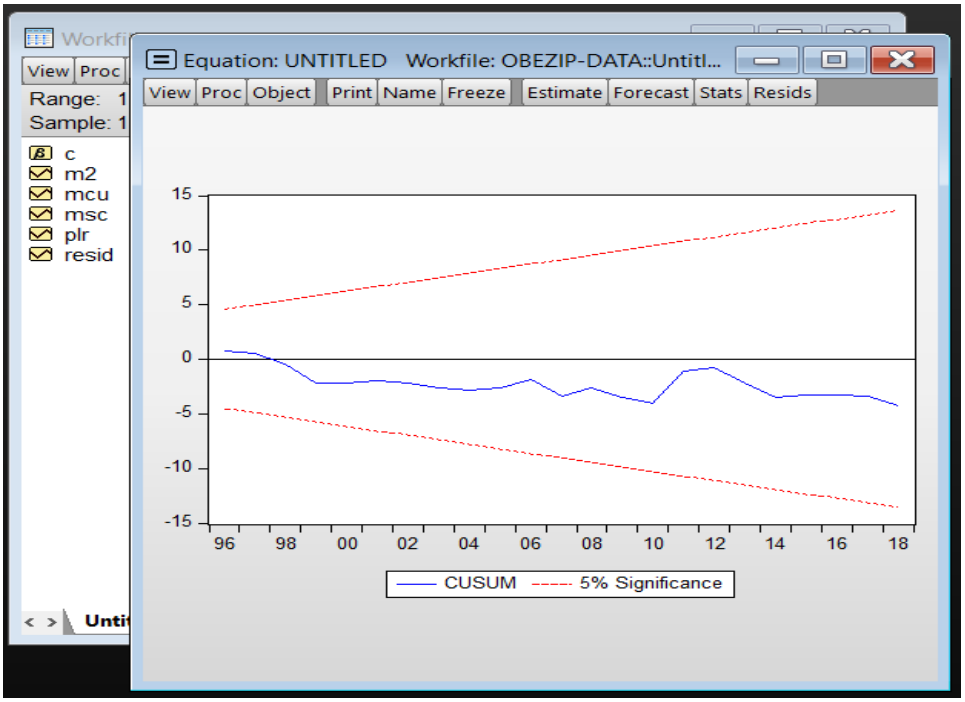
The CUSUM and CUSUM of Square test for stability is meant to determine the appropriateness and the stability of the model. Put differently, the CUSUM test is used to show whether the model is stable and is suitable for making long-run decision.

The following figures demonstrate the step-by-step procedures for performing **CUSUM test** in Views.





The fifth step is to click **OK** on the ***Recursive Estimation*** and the stability plot would appear as follows.

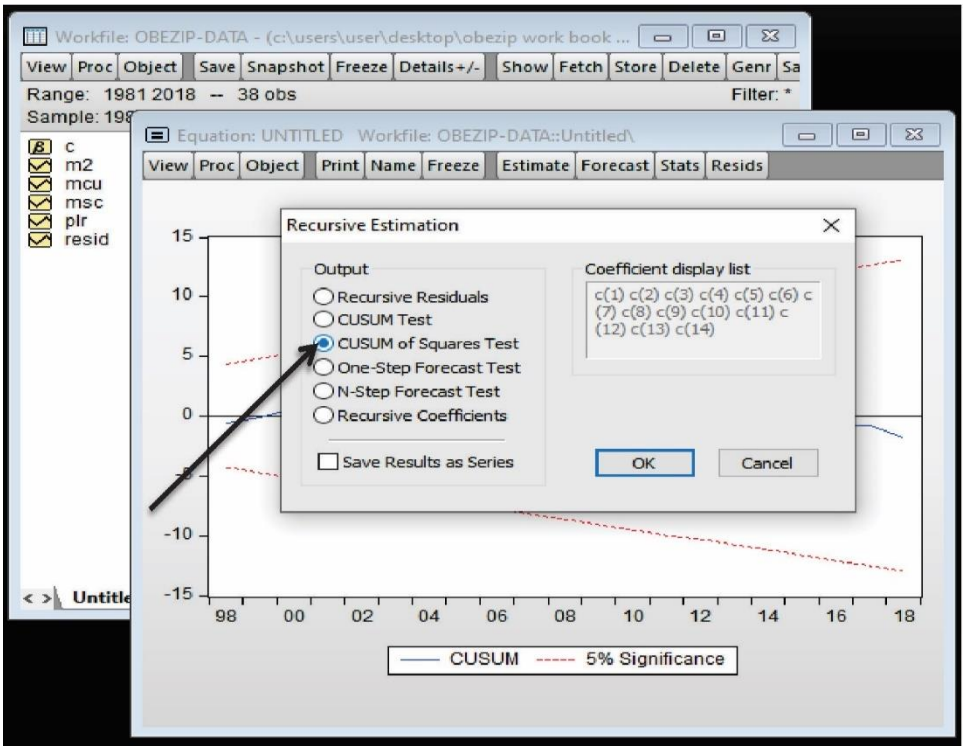


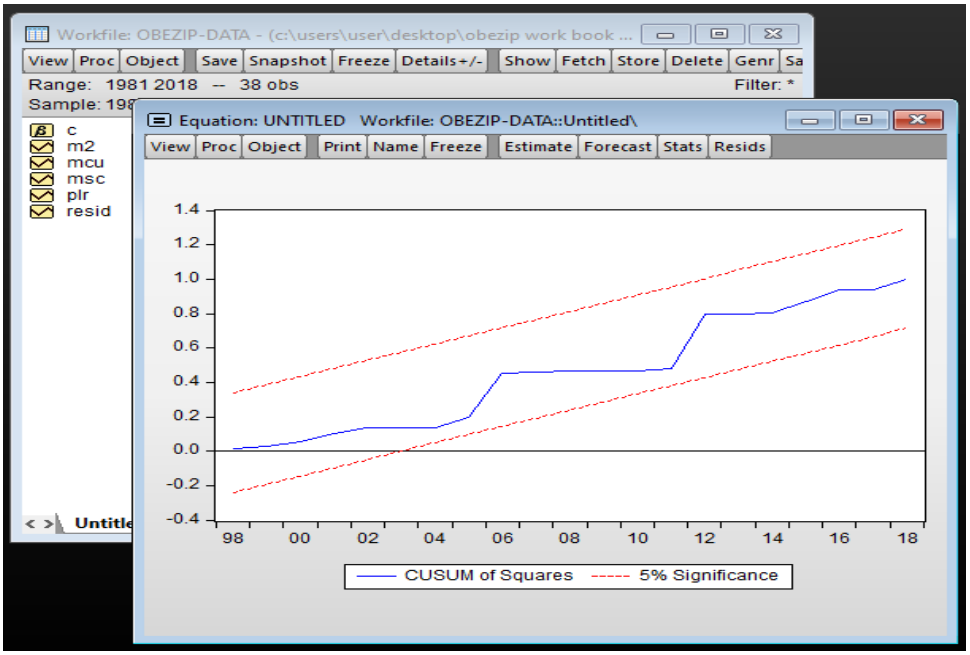
The figure above shows that, the plot of CUSUM for the model under consideration is within the five percent critical bound. This by implication suggests that, the parameters of the model do not suffer from any structural instability over the period of study. That is, all the coefficients in the error correction model are stable.

CUSUM of Square (CUSUMQ) Test Plot

The CUSUM of square statistic is a cumulative sum of squared residuals. The expectations of the CUSUM of squares statistics run from zero at the first observation until the value of one at the end of the sample period, under the null hypothesis of constant coefficients and variance.

The figures below demonstrate the step-by-step procedures for performing **CUSUMQ test** in EViews.





Likewise, the figure above shows that the plot of **CUSUMQ** for the model under consideration is within the five per cent critical bound. This also suggests that the parameters of the model do not suffer from any structural instability over the period of analysis. That is, all the coefficients in the error correction model are stable.

Answer to Self- Assessment 1

What are the Post-Estimation Tests in ARDL?

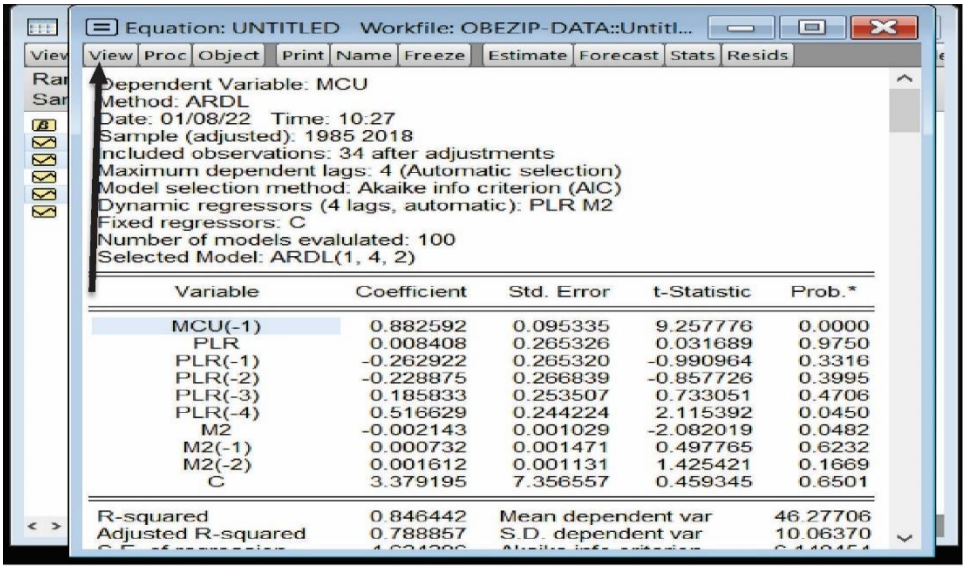
1.4 Hypotheses Testing Using Wald-Test in EViews

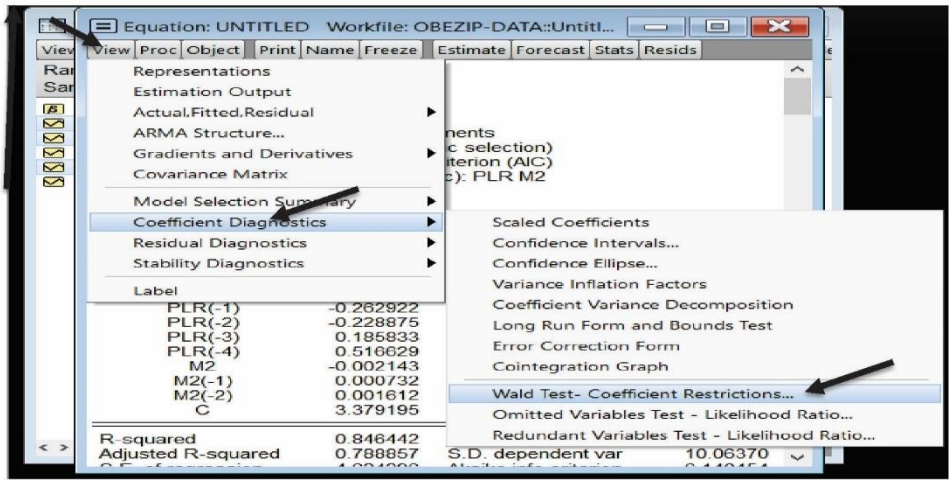
The Wald test computes a test statistic based on the unrestricted regression and tests for the joint significance of coefficients. The estimated equation is used to perform hypothesis tests on the coefficients of the model. The test is used to ascertain whether the joint impact of explanatory/exogenous/independent variables (or regressors) actually have a significant influence the dependent variable. Let's assume that the level of significance for a study is 5% (for a two-tailed test), the decision rule will be that, if the probability value (PV) is less than 5% or 0.05 (that is, PV < 0.05), it implies that, the regressors are jointly statistically significant at 5% level (that is, reject H_0); otherwise, they are not significant at that level.

The practical illustration shown on Wald Test was made using **PLR** and **M2** as **regressors** (or independent variables); while **MCU** is the **regressand (**or dependent variable)

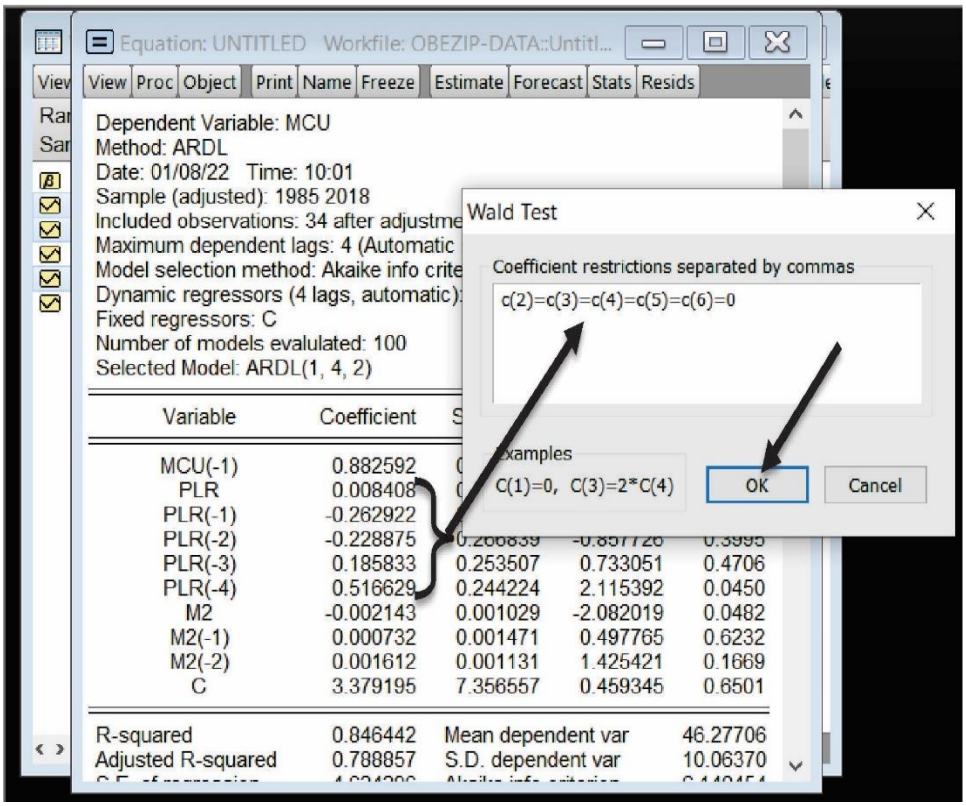
Walt Test of PLR or Joint Impact of PLR on MCU

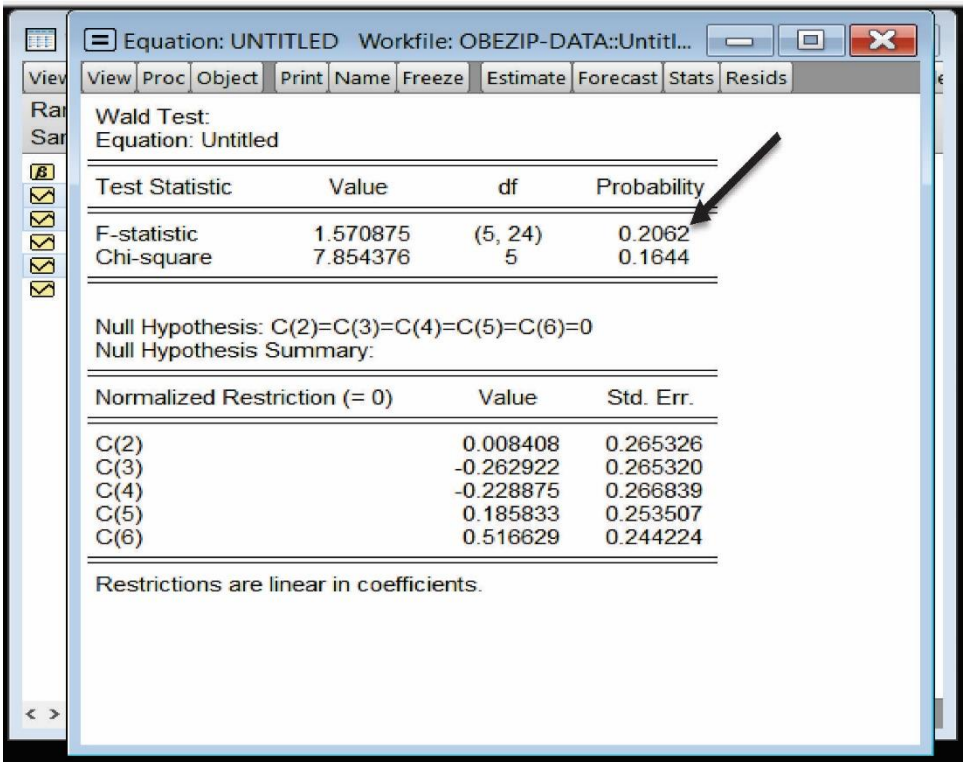
Step 1: To conduct a Wald Test, go to the generated regression result, click **View/Coefficient Diagnostics/Wald Test-Coefficient Restrictions**





Step 2: Enter or type in: c(2)=c(3)=c(4)=c(5)=c(6)=0 or c(2)=0,c(3)=0,c(4)=0,c(5)=0,c(6)=0 in the Wald Test Box. These coefficients correspond to the coefficients of PLR which are according to the order of arrangements from the estimated result as: second order coefficient PLR; third order coefficient PLR(-1), fourth order coefficient PLR(-2), fifth order coefficient PLR(-3) and sixth order coefficient PLR(-4).

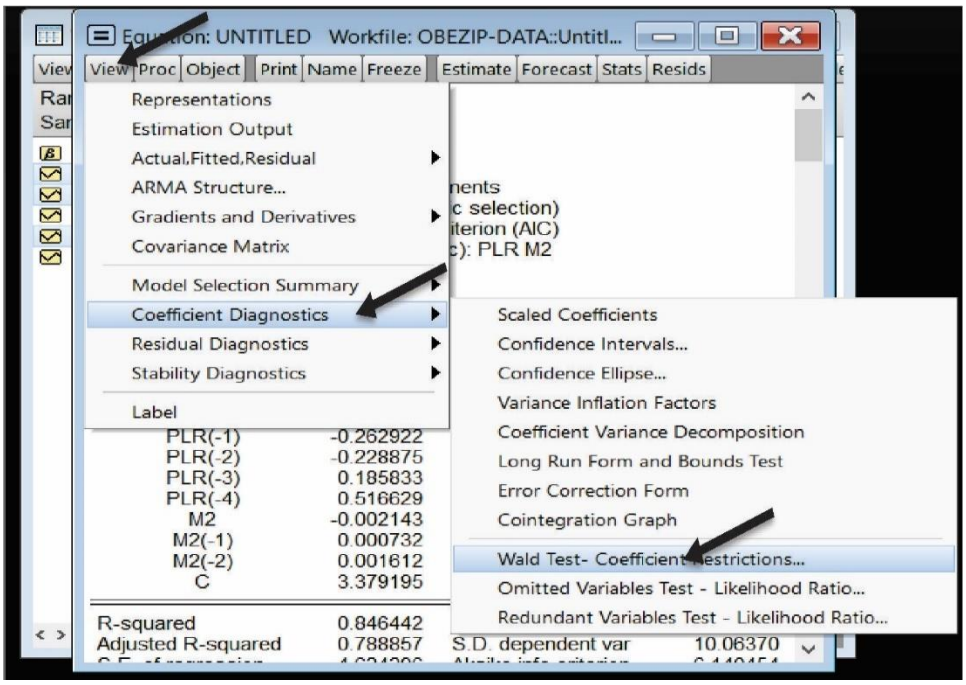

Equation: UNTITLED Workfile: OBEZIP-DATA::Untitl...
View Proc Object Print Name Freeze Estimate Forecast Stats Resids
Dependent Variable: MCU
Method: ARDL
Date: 01/08/22 Time: 10:01
Sample (adjusted): 1985 2018
Included observations: 34 after adjustme
Maximum dependent lags: 4 (Automatic
Model selection method: Akaike info crite
Dynamic regressors (4 lags, automatic):
Fixed regressors: C
Number of models evalulated: 100
Selected Model: ARDL(1, 4, 2)
Wald Test
Coefficient restrictions separated by commas
c(2)=c(3)=c(4)=c(5)=c(6)=0
Examples
C(1)=0, C(3)=2*C(4)
OK
Cancel
Variable Coefficient
MCU(-1) 0.882592
PLR 0.008408
PLR(-1) -0.262922
PLR(-2) -0.228875
PLR(-3) 0.185833 0.253507 0.733051 0.4706
PLR(-4) 0.516629 0.244224 2.115392 0.0450
M2 -0.002143 0.001029 -2.082019 0.0482
M2(-1) 0.000732 0.001471 0.497765 0.6232
M2(-2) 0.001612 0.001131 1.425421 0.1669
C 3.379195 7.356557 0.459345 0.6501
R-squared 0.846442 Mean dependent var 46.27706
Adjusted R-squared 0.788857 S.D. dependent var 10.06370


Equation: UNTITLED Workfile: OBEZIP-DATA::Untitl...
View Proc Object Print Name Freeze Estimate Forecast Stats Resids
Wald Test:
Equation: Untitled
Test Statistic Value df Probability
F-statistic 1.570875 (5, 24) 0.2062
Chi-square 7.854376 5 0.1644
Null Hypothesis: C(2)=C(3)=C(4)=C(5)=C(6)=0
Null Hypothesis Summary:
Normalized Restriction (= 0) Value Std. Err.
C(2) 0.008408 0.265326
C(3) -0.262922 0.265320
C(4) -0.228875 0.266839
C(5) 0.185833 0.253507
C(6) 0.516629 0.244224
Restrictions are linear in coefficients.

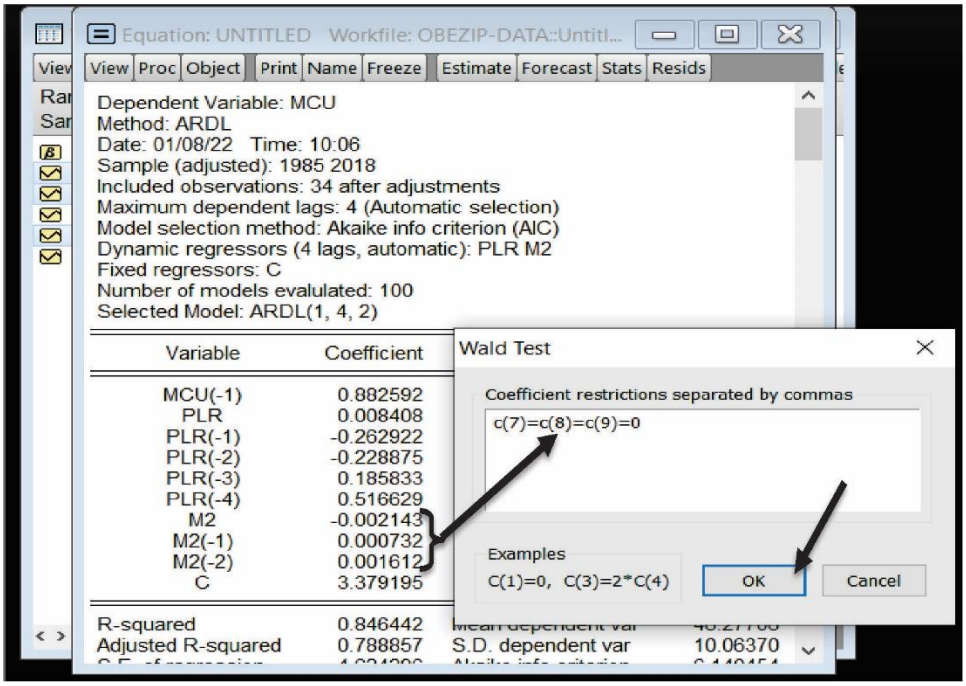
According to the Wald Test result on PLR, the Chi-square test statistic is: 1.570875. The p-value is: 0.2062; and 0.2062 > 0.05 (using 5% level of significance). One cannot reject the H_0. The test is not significant. The joint coefficient test of PLR is not statistically significant.

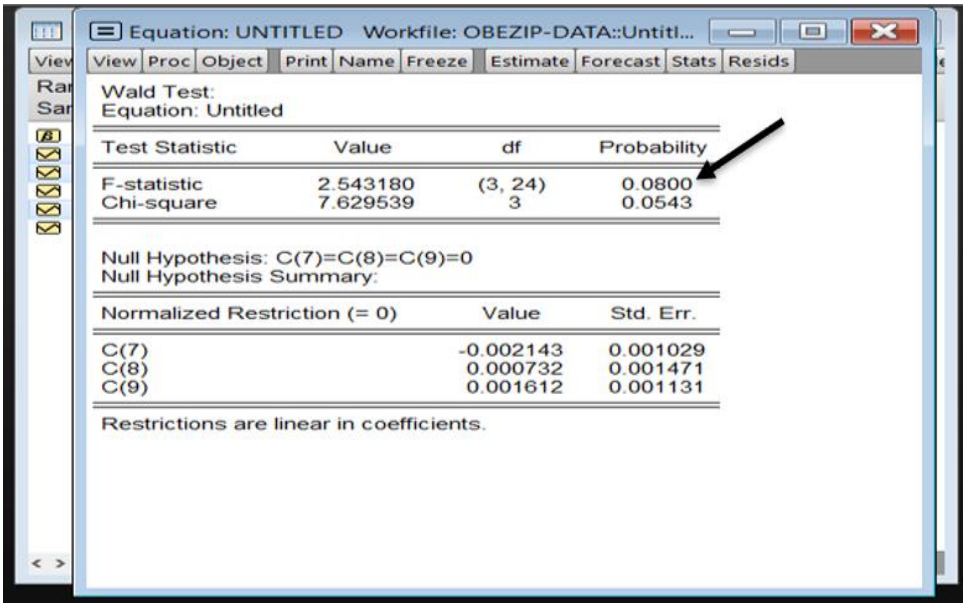
Wald Test of M2 or Joint Impact of M2 on MCU

Step 1: Go to the generated regression result, click **View/Coefficient Diagnostics/Wald Test-Coefficient Restrictions**



Step 2: Enter c(7)=c(8)=c(9)=0 or c(7)=0,c(8)=0,c(9)=0 in Wald Test box as shown below:





1.5 The Long Run Model and Error Correction Model (ECM) in EViews

If a linear combination of nonstationary (cointegrated) series is discovered to be the results of a static model like equation, it would be correct to interpret the non-stationary

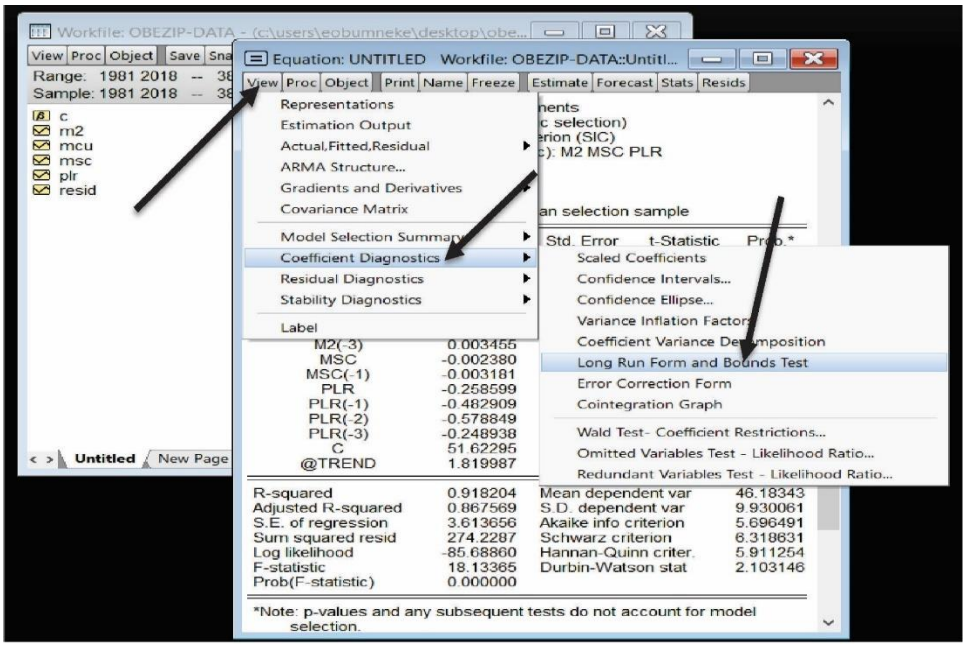
series as the results of the static model without paying attention to the time series properties and short-run dynamics of the model. Despite this, a substantial argument has been made that, because equilibrium (i.e. steady state) is rarely observed, it may be required to analyse the series' short-run evolution and adjustment dynamics.

To conduct, The Long Run Model and Error Correction Model (ECM) in EViews

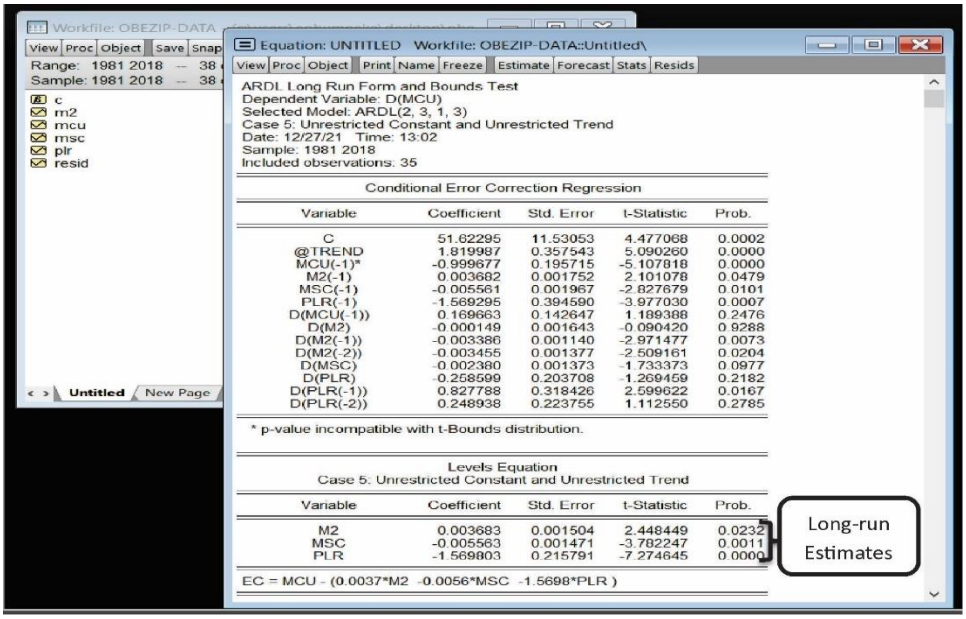
i. From the result, Click on **View** on the Menu Bar

ii. Click on **Coefficient Diagnostics**

iii. Select the **Long Run Form and Bounds Test** option



The diagram below shows the result window for the **Long Run Form**

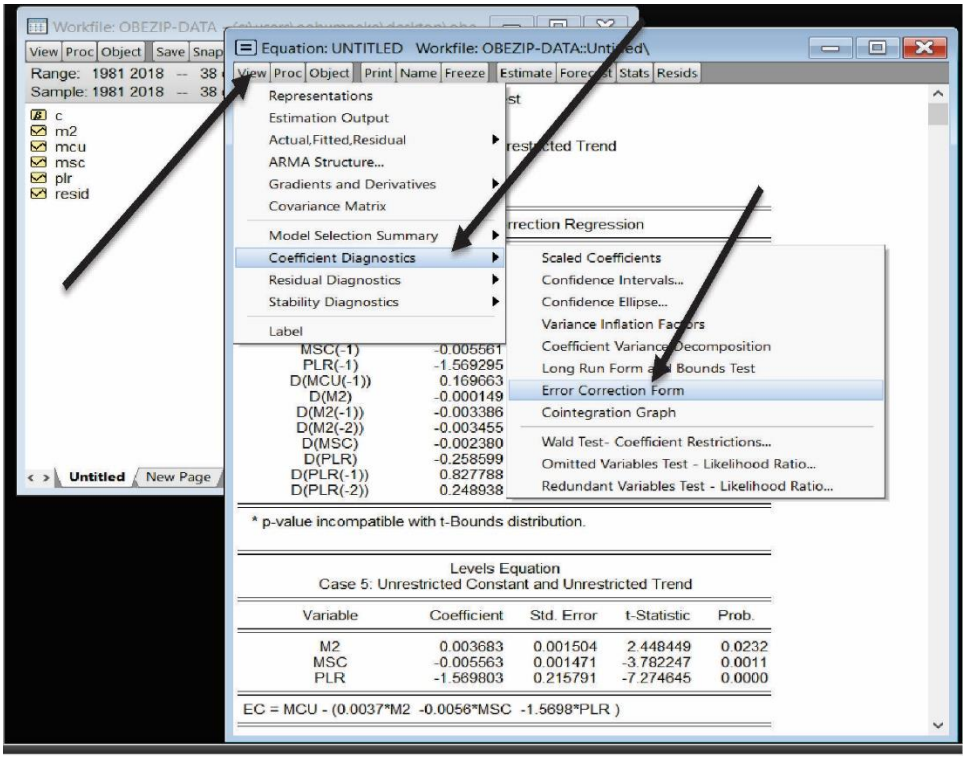


Estimate the Error Correction Model

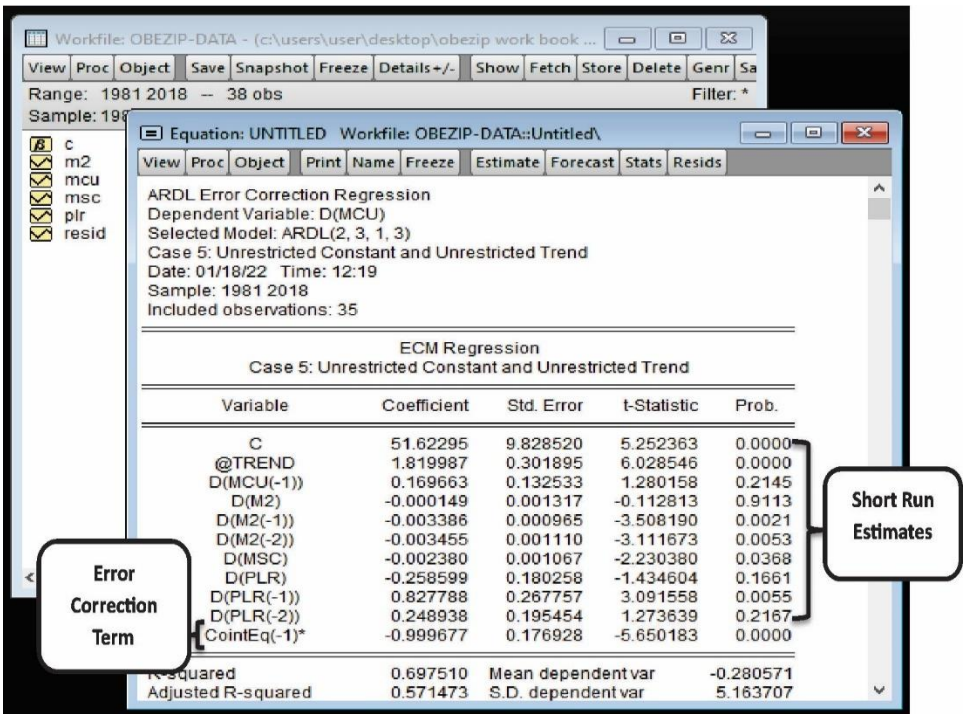
i. From the result, Click on **View** on the Menu Bar

ii. Click on **Coefficient Diagnostics**

iii. Select the **Error Correction Model** option



The diagram below shows the result window for the **ARDL Cointegrating and Short-run Form.**



1.6 Summary

You have learnt the various post post-estimation diagnostic tests in ARDL models. This test plays an indispensable role in validating the model's assumptions, hence affirming the reliability and validity of the estimation results. They serve to identify any potential issues like serial correlation, heteroskedasticity, non-normal errors, parameter instability, or lack of cointegration, which could undermine the accuracy and reliability of the model. Therefore, a rigorous post-estimation examination is as crucial as the model estimation itself in the econometric analysis.

1.7 References/Further Reading

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Narayan, P. K. (2005). The saving and investment nexus in China: evidence from cointegration tests. Applied Economics, 37, 1979 – 1990.

Pesaran, M., Shin, Y. & Smith, R.. (2001). Bound testing approaches to the analysis of level relationship. J. Appl. Econ. 16, 289–326.

Pesaran, H. and Shin, Y. (1999). An autoregressive distributed lag modeling approach to cointegration analysis. In: Strom, S. (Ed.), Econometrics and Economic Theory in 20th Century: The Ragnar–Frisch Centennial Symposium. Cambridge University Press: Cambridge.

1.8 Possible Answers to Self-Assessment Exercise(s) Within the Content

Answer to Self- Assessment 1

After estimating an Autoregressive Distributed Lag (ARDL) model, several post-estimation diagnostic tests are usually performed to ensure the model is well-specified and the assumptions of the model are met. Here are some of the common post-estimation tests:

1. **Serial Correlation Test**: This test checks if the residuals are autocorrelated, which violates one of the assumptions of the model. The Breusch-Godfrey Serial Correlation LM test is commonly used for this purpose.

2. **Heteroskedasticity Test**: This test checks if the residuals have constant variance (homoscedasticity), which is another assumption of the model. The Breusch-Pagan test or White test can be used.
3. **Normality Test**: This test checks if the residuals are normally distributed, which is another assumption of the model. The Jarque-Bera test is commonly used for this purpose.
4. **Model Specification Test**: This test checks if the model is correctly specified, that is, whether all relevant variables have been included and whether the functional form of the model is correct. The Ramsey RESET test is commonly used.
5. **Stability Test**: This test checks the stability of the model parameters over the sample period. The CUSUM (Cumulative Sum) and CUSUMSQ (Cumulative Sum of Squares) tests can be used.

MODULE 3: PANEL DATA ESTIMATION

Unit 1: Panel Data Regression Model

Unit 2: Fixed Versus Random Effects Panel Data

Unit 3: Testing Fixed and Random Effects

Unit 4: Panel Data Estimation in EViews

UNIT 1: PANEL DATA REGRESSION MODEL

1.1 Introduction

In the former unit, you learnt the answer to the question “What is Stationarity?” You also learnt how to conduct unit roots test in the AR(1) Models as well as testing for cointegration and stationarity of the variables using EViews software. The mastery of these concepts now prepared you for the study of panel data in this present unit. Panel data is a model which comprises variables that vary across time and cross section, in this

paper we will describe the techniques used with this model including a pooled regression, a fixed effect and a random effect.

1.2 Learning Outcomes

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

- Know the meaning of Panel Data Regression Model
- Explain panel Data Examples
- Discuss the advantages of Panel Data
- List the importance of Panel Data
- Design the format of a Panel Data
- List types of Panel Data

1.3 Meaning of Panel Data Regression Model

Panel (data) analysis is a statistical method, widely used in social science, epidemiology, and econometrics, which deals with two and "^"-dimensional (in and by the - cross sectional/times series time) panel data. The data are usually collected over time and over the same individuals and then a regression is run over these two dimensions. Multidimensional analysis is an econometric method in which data are collected over more than two dimensions (typically, time, individuals, and some third dimension). Panel data are repeated cross-sections over time, in essence there will be space as well as time dimensions. Panel data are a type of longitudinal data, or data collected at different points

in time. To put succinctly, panel data is a combination of both cross sectional data and time series data.

That is, Cross-sectional data (data collected on several individuals/units at one point in time) and time series data (data collected on one individual/unit over several time periods). Panel data are repeated cross-sections over time, in essence there will be space as well as time dimensions. Other names are pooled data, micropanel data, longitudinal data, event history analysis and cohort analysis.

Panel data allows you to control for variables you cannot observe or measure like cultural factors or difference in business practices across companies; or variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.). This is, it accounts for individual heterogeneity. With panel data you can include variables at different levels of analysis (i.e. students, schools, districts, states) suitable for multilevel or hierarchical modelling.

It is important to note: Time series data: Many observations (large t) on as few as one unit (small N). Examples: stock price trends, aggregate national statistics. Cross sectional data: analysis of data on n, several distinct entities at a given point of time (cross sectional data). Panel data: A data set with both a cross section and a time dimension.

Taken together, the repeated observation of one unit constitutes a “panel”. Other names of panel data are pooled data, micro panel data, longitudinal data, event history analysis and cohort analysis.

1.3.1 Panel Data Examples

The individuals/units can for example be workers, firms, states or countries. Annual unemployment rates of each state over several years. Quarterly sales of individual stores over several quarters. Wages for the same worker, working at several different jobs.

1. **Household Surveys:** Household surveys, such as the Panel Study of Income Dynamics (PSID) or the German Socio-Economic Panel (SOEP), collect data on income, expenditure, employment, health, and other socio-economic variables from the same households over time.
2. **Country-Level Data:** In macroeconomics and international economics, data on GDP, inflation, unemployment, and other macroeconomic variables are often collected for multiple countries over several years.
3. **Firm-Level Data:** In corporate finance and industrial organization, data on firm performance, profitability, market share, and other business metrics are collected for multiple firms across different periods.

1.3.2 Advantages of Panel Data

The use of panel data offers several significant advantages:

i. **Controlling for Individual Heterogeneity:** Panel data allows for the control of individual-specific characteristics that are time-invariant but differ across the cross-sectional units. These characteristics might be unobservable and thus result in omitted variable bias in cross-sectional or time series analysis.

ii. **Dynamic Analysis:** With multiple time periods, panel data allows for the study of dynamics, such as adjustment processes or lags in behavior.

iii. **Greater Data Variety:** Panel data provides more data points, enhancing degrees of freedom and reducing multicollinearity, thus improving the efficiency of econometric estimates.

iv. **Studying Temporal and Spatial Effects:** Panel data allows for the investigation of the impact of variables that change over time but are constant across entities, or vice versa. This enables the study of both temporal and spatial effects.

1.4 Importance of Panel Data

The importance of panel data rests on several advantages that set it apart from purely cross-sectional or time-series data:

i. **Control for Individual Heterogeneity:** A distinguishing feature of panel data is its ability to account for individual-specific, time-invariant unobservable characteristics. For instance, in an economic growth study, certain country-specific factors like culture or geography remain constant over time and can influence the variables of interest. These unobservable factors may cause bias in the estimated relationships if not properly controlled. Panel data allows the use of fixed or random effects models that can handle this unobserved heterogeneity, thereby providing more consistent and efficient estimates.

ii. **Study of Dynamics:** Panel data provides the temporal dimension necessary to analyze the dynamics of change and the impacts of past events on current situations. It allows the researcher to study lagged relationships and model dynamic behavior, leading to a more thorough understanding of the phenomena under study.

iii. **Enhanced Data Availability:** Panel datasets increase the number of observations by tracking the same units over time, thereby enhancing the power and efficiency of statistical tests. More data points lead to better model specifications, increase the degrees of freedom, reduce multicollinearity, and consequently improve the precision of the coefficient estimates.

iv. **Policy Analysis:** Panel data is indispensable in policy analysis. The temporal variation in policies and their impacts can be observed and analyzed across different cross-sectional units. This ability to study the impact of changes in policies over time and across different entities leads to more reliable policy implications.

v. **Investigation of Causal Relationships:** With its time-series component, panel data can better identify and measure effects that are simply not detectable in pure cross-sectional or time-series data. It can help establish the direction of causality in the relationships between variables, a feature of utmost importance in econometric analysis.

Self-Assessment Exercise 1

Discuss the importance of panel data

1.4.1 Format of a Panel Data

Below is a typical example of the format a Panel data usually takes.

Table M3.1.1: Panel Data Form

Firm	Year	Y	X_1	X_2
Firm 1	2019	231	3	18
Firm 1	2020	334	3	27
Firm 1	2021	436	3	32
Firm 2	2019	332	5	23
Firm 2	2020	401	5	33
Firm 2	2021	423	5	45
Firm 3	2019	304	2	22
Firm 3	2020	511	2	31
Firm 3	2021	634	2	48

Below is a typical example of a panel data equation.

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + \mu_{it}$$

$$i = 1...N$$

$$t = 1...T$$

1.5 Types of Panel Data

A panel data set contains n entities or subjects, each of which includes T observations measured at 1 through t time period. Thus, the total number of observations in the panel data is nT. Ideally, panel data are measured at regular time intervals (e.g., year, quarter, and month). Otherwise, panel data should be analyzed with caution. A panel may be long or short, balanced or unbalanced, and fixed or rotating.

1.5.1 Long Versus Short Panel Data

A short panel has many entities (large n) but few time periods (small T), while a long panel has many time periods (large T) but few entities (Cameron and Trivedi, 2009: 230). Accordingly, a short panel data set is wide in width (cross-sectional) and short in length (time-series), whereas a long panel is narrow in width. Both too small N (Type I error) and too large N (Type II error) problems matter. Researchers should be very careful especially when examining either short or long panel.

1.5.2 Balanced Versus Unbalanced Panel Data

In a balanced panel, all entities have measurements in all time periods. In a contingency table (or cross-table) of cross-sectional and time-series variables, each cell should have only one frequency. Therefore, the total number of observations is nT. This tutorial document assumes

that we have a well-organized balanced panel data set. When each entity in a data set has different numbers of observations, the panel data are not balanced. Some cells in the contingency table have zero frequency. Accordingly, the total number of observations is not nT in an unbalanced panel. Unbalanced panel data entail some computation and estimation issues although most software packages are able to handle both balanced and unbalanced data.

1.5.3 Fixed Versus Rotating Panel Data

If the same individuals (or entities) are observed for each period, the panel data set is called a fixed panel (Greene 2008: 184). If a set of individuals changes from one period to the next, the data set is a rotating panel. This document assumes a fixed panel.

1.6 Summary

In this unit you learnt panel data model which comprises variables that vary across time and cross section, and the types of panel data, as well as its importance for analysis. In summary, panel data offers a unique blend of cross-sectional and time-series observations, providing a multifaceted perspective on the phenomena under investigation. Its capacity to control for individual heterogeneity, enable dynamic analysis, offer greater data variety, and facilitate the study of temporal and spatial effects makes it a powerful tool in econometric analysis. Examples of panel data usage in household surveys, country-level macroeconomic data, and firm-level performance metrics underscore its versatility and applicability across fields. Despite challenges with model specification, measurement error, and missing data, the benefits of panel data make it a cornerstone of modern econometric analysis.

Tutor Marked Assignment

Discuss the importance of panel data

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th - 30th April, 2015 in University of Illorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH,

USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(s) Within the Content

Answer to Self- Assessment 1

Panel data, also known as longitudinal data or cross-sectional time series data, is a dataset that combines cross-sectional and time series data. It consists of observations on multiple subjects (such as individuals, firms, countries) across multiple time periods. Panel data analysis is widely used in social sciences, epidemiology, finance, and many other fields. Here are some of the reasons why panel data is important:

1. **More Information and Variability**: Panel data provide more data points, leading to increased variability, less collinearity among variables, more degrees of freedom, and more efficiency. This can increase the power of statistical analyses.
2. **Control for Individual Heterogeneity**: Panel data allow for the control of individual-specific variables (observable and unobservable), which might be constant over time but vary across entities. This is a major advantage over cross-sectional or time series data, as ignoring individual-specific effects could lead to biased results.
3. **Study Dynamics of Change**: With panel data, you can study the dynamics of change, as you have data on the same individuals at different points in time. For example, you can analyze the impact of a policy change on the behavior of individuals or firms.
4. **Detect and Measure Effects that Cannot Be Observed in Pure Cross-Section or Time-Series Data**: For instance, the effects of variables that change over time but do not change across entities, or that change across entities but do not change over time.
5. **Address Certain Types of Endogeneity and Measurement Error**: Panel data with multiple time periods provide a platform for using instrumental variable or fixed effects methods to address endogeneity issues due to omitted variables.
6. **Test and Control for Non-stationarity**: Panel data allows for testing and controlling for unit root and cointegration, common issues in time-series analysis that can lead to spurious regression results.
7. **Better Understanding of Adjustment Processes**: Panel data provide information not only about the impact of changes in variables of interest but also about the speed at which the adjustments take place.

UNIT 2: FIXED VERSUS RANDOM EFFECTS PANEL DATA

1.1 Introduction

In the preceding unit you learnt panel data models. In the present unit which is Unit 2, we shall continue with our discussion on panel data by examining fixed versus random effects panel data.

1.1 Learning Outcomes

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

- Discuss fixed versus random effects model
- Explain the advantages of fixed effect model
- Discuss the advantages of random effect model
- Explain the difference between fixed and random effects model

1.3 Fixed Versus Random Effects

Panel data models examine group (individual-specific) effects, time effects, or both in order to deal with heterogeneity or individual effect that may or may not be observed. These effects are either fixed or random effect. A fixed effect model examines if intercepts vary across group or time period, whereas a random effect model explores differences in error variance components across individual or time period. A one-way model includes only one set of dummy variables (e.g., firm1, firm2, ...), while a two-way model considers two sets of dummy variables (e.g., cityl, city2, ... and yearl, year2, ...). Panel data models examine fixed and/or random effects of individual or time. The core difference between fixed and random effect models lies in the role of dummy variables.

A parameter estimate of a dummy variable is a part of the intercept in a fixed effect model and an error component in a random effect model. Slopes remain the same across group or time period in either fixed or random effect model. The functional forms of one-way fixed and random effect models are,

Fixed effect model: $Y_{it} = \alpha_i + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + \mu_{it}$

Random effect model: $Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + (v_i + \mu_{it})$

where u is a fixed or random effect specific to individual (group) or time period that is not included in the regression, and errors are independent identically distributed,

A fixed group effect model examines individual differences in intercepts, assuming the same slopes and constant variance across individual (group and entity). Since an individual specific effect is time invariant and considered a part of the intercept, u is

allowed to be correlated with other regressors; That is, OLS assumption 2 is not violated. This fixed effect model is effect estimated by least squares dummy variable (LSDV) regression (OLS with a set of dummies) and within effect estimation methods (Green, 2012).

Table M3.2.1: Fixed Versus Random Effects

	Fixed Effect Model	Random Effect Model
Functional form	$y_{it} = (\alpha + u_i) + X_{it}'\beta + v_{it}$	$y_{it} = \alpha + X_{it}'\beta + (u_i + v_{it})$
Assumption	-	Individual effects are not correlated with regressors
Intercepts	Varying across group and/or time	Constant
Error variances	Constant	Randomly distributed across group and/or time
Slopes	Constant	Constant
Estimation	LSDV, within effect estimation	GLS, FGLS (EGLS)
Hypothesis test	F test	Breusch-Pagan LM test

1.4 Fixed-Effects Model

(Covariance Model, Within Estimator, Individual Dummy Variable Model, Least Squares Dummy Variable Model). Use fixed-effects (FE) whenever you are only interested in analyzing the impact of variables that vary over time. FE explore the relationship between predictor and outcome variables within an entity (country, person, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables (for example, being a male or female could influence the opinion toward certain issue; or the political system of a particular country could have some effect on trade or GDP; or the business practices of a company may influence its stock price).

When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. This is the rationale

behind the assumption of the correlation between entity's error term and predictor variables. FE removes the effect of those time-invariant characteristics so we can assess the net effect of the predictors on the outcome variable.

Another important assumption of the FE model is that those time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics. Each entity is different therefore the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others. If the error terms are correlated, then FE is no suitable since inferences may not be correct and you need to model that relationship (probably using random-effects).

The equation for the fixed effects model becomes:

$$Y_{it} = \alpha_i + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + \mu_{it}$$

Where

- α_i (i= 1 ....n) is the unknown intercept for each entity (n entity-specific intercepts).
- Y_{it} is the dependent variable (DV) where i = entity and t = time.
- X_{kit} represents one independent variable (IV),
- β_k is the coefficient for that IV,
- μ_{it} is the error term

Fixed-effects will not work well with data for which within-cluster variation is minimal or for slow changing variables over time.

The main advantage of the fixed effects model is its ability to control for time-invariant unobservable characteristics, thereby eliminating any potential bias caused by these

omitted variables. However, this model cannot estimate the effects of time-invariant variables because these variables are absorbed by the fixed effects.

The fixed effects model is most appropriate when the researcher believes that the individual-specific effects are unique to each cross-sectional unit and therefore should not be generalized to other units. The FE model is often the model of choice when the focus is on analyzing the impact of variables that vary over time.

1.4.1 The Advantages of the Fixed Effects Model in Panel Data Analysis

The fixed effects model is commonly applied in circumstances where the objective is to analyze the impact of time-varying variables within specific cross-sectional units, like individuals, firms, or countries. The key advantages of the fixed effects model include:

i. **Controlling for Time-Invariant Characteristics:** One of the main advantages of the fixed effects model is its ability to control for time-invariant characteristics. These could be unobservable characteristics specific to each cross-sectional unit that don't change over time but could be correlated with the explanatory variables. In other words, if there are unique, constant characteristics for each unit that affect the dependent variable, the fixed effects model is adept at handling these, thereby reducing omitted variable bias.

ii. **Focus on Within Variation:** The fixed effects model is designed to examine the causes of changes within an entity. By focusing on the variation within each cross-sectional unit over time, the fixed effects model can identify the impact of

explanatory variables on the dependent variable within the same entity. This makes it an excellent choice when the interest lies in understanding the within-entity effect of a particular variable.

iii. **Eliminating Unobserved Heterogeneity:** The fixed effects model assumes that individual-specific effects are correlated with the regressors. Therefore, it can eliminate unobserved heterogeneity that is constant over time, which helps prevent potential biases in the estimated coefficients.

iv. **Handling of Large Samples:** The fixed effects model handles large samples well. This can be especially useful in studies involving large datasets, as the fixed effects model allows the control of individual-specific heterogeneity in large panels.

v. **Usefulness in Policy Evaluation:** The fixed effects model is particularly useful in policy evaluation where the researcher is interested in the impact of changes in policy (a time-varying variable) on a particular outcome within the same entity.

Self- Assessment 1

Discuss the advantages of fixed effect model

1.5 Random Effects Method

The rationale behind random effects model is that, unlike the fixed effects model,the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model:

"…the crucial distinction between fixed and random effects is whether the unobserved

individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not"

If you have reason to believe that differences across entities have some influence on your dependent variable then you should use random effects. An advantage of random effects is that you can include time invariant variables (i.e. gender). In the fixed effects model these variables are absorbed by the intercept.

Hence, the variability of the constant for each section comes from:

$$\alpha_i = \alpha + v_i$$

where v_i is a zero *mean* standard random variable. The random effects model therefore takes the following form:

$$Y_{it} = (\alpha + v_i) + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + \mu_{it}$$

$$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + ----+ \beta_k X_{kit} + (v_i + \mu_{it})$$

Random effects assume that the entity's error term is not correlated with the predictors which allows for time-invariant variables to play a role as explanatory variables.

In random-effects, you need to specify those individual characteristics that may or may not influence the predictor variables. The problem with this is that some variables may not be available therefore leading to omitted variable bias in the model.

RE allows to generalize the inferences beyond the sample used in the model.

The random effects model is typically chosen when the researcher believes that the individual-specific effects are drawn from a larger population and are uncorrelated with

the independent variables. The RE model is often used when the focus is not only on the impact of time-varying variables but also on time-invariant variables.

Choosing between these two models usually depends on the nature of the data and the research questions being addressed. The Hausman test is a common statistical procedure used to decide between the FE and RE models. It tests the null hypothesis that the individual-specific effects are uncorrelated with the other regressors in the model. If this null hypothesis is rejected, the FE model is preferred; otherwise, the RE model is suitable.

1.5.1 The Advantages of the Random Effects Model in Panel Data Analysis

The random effects model assumes that the individual-specific effect is a random variable uncorrelated with the independent variables. It essentially treats the individual-specific effect as a component of the error term. The key advantages of the random effects model include:

i. **Consideration of Between-Entity Variation:** Unlike the fixed effects model that focuses only on within-entity variation, the random effects model considers both within and between-entity variation. This ability to utilize the variation between entities provides additional information that can lead to more efficient estimates.
ii. **Estimation of Time-Invariant Variables:** A significant advantage of the random effects model over the fixed effects model is its ability to estimate the effects of

time-invariant variables. If the research interest includes the impact of variables that do not change over time, such as gender or geographic location, then the random effects model would be the more suitable choice.

iii. **Greater Efficiency:** Under the assumption that the individual-specific effects are uncorrelated with the regressors, the random effects model provides more efficient (i.e., smaller standard errors) estimates than the fixed effects model. This is because it uses more information by exploiting both the within and between variations.

iv. **Generalizability:** The random effects model views the individual-specific effects as a random sample from a larger population. This allows for the generalization of the results to a broader population, which can be particularly valuable in policy implications.

v. **Better Use of Degrees of Freedom:** In large datasets, both fixed effects and random effects models can be effective. However, in smaller samples, the random effects model often makes better use of degrees of freedom. This is because unlike the fixed effects model, it does not estimate a different intercept for each cross-sectional unit.

1.6 Summary

In this unit 2, you have learnt two techniques use to analyze panel data. They are fixed effects random effects fixed versus random effects. Both fixed and random effects

models play crucial roles in panel data analysis, each providing unique insights depending on the research question and data at hand. While the fixed effects model controls for time-invariant characteristics unique to each cross-sectional unit, the random effects model exploits both within and between variations in the data and assumes the individual-specific effect is a random draw from a larger population. Understanding the assumptions and implications of each model is critical for accurate and meaningful econometric analysis.

Tutor Marked Assignment

Distinguish using examples between fixed effects and random effects in econometric analysis.

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Illorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(s) Within the Content

Answer to Self- Assessment 1

The fixed effects model is a popular approach for analyzing panel data where the interest lies in analyzing the impact of variables that vary over time. Here are some of the key advantages of the fixed effects model:

1. **Controls for Time-Invariant Characteristics**: The fixed effects model is designed to control for the effect of time-invariant characteristics of each individual. This means that it takes into account those characteristics that are unique to each individual but do not change over time, effectively controlling for individual heterogeneity.
2. **Eliminates Unobserved Heterogeneity**: The fixed effects model controls for unobserved variables that do not vary over time. Therefore, it eliminates the bias caused by time-invariant unobserved individual-specific effects when these are correlated with the independent variables.
3. **Examines the Impact of Variables that Change Over Time**: The fixed effects model is specifically good at dealing with variables that vary over time. This makes it suitable for studying the impact of policy changes, individual behaviors, or other factors that are expected to change over time.
4. **Avoids Omitted Variable Bias**: Because the fixed effects model accounts for individual-specific constant characteristics, it avoids omitted variable bias caused by time-invariant variables.
5. **Reflects within-individual Changes**: The fixed effects model emphasizes within-individual changes over time rather than differences between individuals. This makes it suitable for studies focused on individual dynamics.

UNIT 3: TESTING FIXED AND RANDOM EFFECTS

1.1 Introduction

One of the critical decisions in panel data analysis is choosing between the fixed effects (FE) and random effects (RE) models. Both models have different assumptions and interpretations, which can lead to differing results. Therefore, empirical tests are used to guide the decision between these two models. This unit discusses the methods for testing fixed and random effects.

1.2 Learning Outcomes

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

- Test fixed and random effects
- Discuss Breusch-Pagan LM Test for Random Effects
- Discuss Hausman Test for Comparing Fixed and Random Effects
- Understand guidelines of Model Selection

1.3 Testing Fixed and Random Effects

How do we know if fixed and/or random effects exist in panel data in hand? A fixed effect is tested by F-test, while a random effect is examined by Breusch and Pagan's (1980) Lagrange multiplier (LM) test. The former compares a fixed effect model and POLS to see how much the fixed effect model can improve the goodness-of-fit, whereas the latter contrast a random effect model with POLS. The similarity between random and fixed effect estimators is tested by a Hausman test.

1.3.1 F-Test for Fixed Effects

To run the test, one need to compare fixed effects to the simple common constant OLS method using the conventional F-test. The null hypothesis states that, all of the constants are the same (homogeneity) and that, the common constant approach can be used.

$H_0 : a_1 = a_2 = ...a_N$

The F-statistic is:

$$F = \frac{\left(R^2_{FE} - R^2_{CC}\right)/(N-1)}{\left(1-R^2_{FE}\right)/)NT-N-k)} \square F(N-1, NT-N-k)$$

If the null hypothesis is rejected (at least one group/time specific intercept u is not zero), you may conclude that there is a significant fixed effect or significant increase in goodness-of-fit in the fixed effect model; therefore, the fixed effect model is better than the pooled OLS.

Self-Assessment Exercise 1

Examine F-test for fixed effects of a panel data

1.3.2 Breusch-Pagan LM Test for Random Effects

Another diagnostic test used in this context is the Breusch-Pagan Lagrange Multiplier (LM) test. This test is specifically for determining the presence of random effects. The null hypothesis of the Breusch-Pagan LM test is that variances across entities are zero, which implies no significant difference across units (i.e., no panel effect). If this

hypothesis is rejected, it suggests the suitability of the RE model over the pooled ordinary least squares (OLS) model. However, if we fail to reject the null, then the panel data structure is not beneficial, and a simple OLS regression would suffice.

Breusch and Pagan's (1980) Lagrange multiplier (LM) test examines if individual (or time) specific variance components are zero, H0 : $\dot{O}U2$= 0. The LM statistic follows the chi-squared.

The LM statistic follows the chi-squared

$$LM_u = \frac{nT}{2(T-1)}\left[\frac{T^2\bar{e}'\bar{e}}{e'e} - 1\right] \square \chi^2(1),$$

Where $\bar{e}$ is the n × 1 vector of the group means of pooled regression residuals, and e'e is the SSE of the pooled OLS regression.

If the null hypothesis is rejected, you can conclude that there is a significant random effect in the panel data, and that the random effect model is able to deal with heterogeneity better than the pooled OLS.

1.3.3 Hausman Test for Comparing Fixed and Random Effects

How do we know which effect (fixed effect or random effect) is more relevant and significant in the panel data? The Hausman specification test compares fixed and random effect models under the null hypothesis that individual effects are uncorrelated with any regressor in the model (Hausman, 1978). If the null hypothesis of no correlation is not violated, LSDV and GLS are consistent, but LSDV is inefficient; otherwise, LSDV is

consistent but GLS is inconsistent and biased (Greene, 2008: 208). The estimates of LSDV and GLS should not differ systematically under the null hypothesis. The Hausman test uses that "the covariance of an efficient estimator with its difference from an inefficient estimator is zero".

The Hausman test is a standard procedure for deciding between the fixed and random effects models. The null hypothesis of the Hausman test is that the preferred model is the random effects, against the alternative that the fixed effects would be a better fit. In simpler terms, the Hausman test checks whether the unique errors (individual effects plus the idiosyncratic errors) are correlated with the regressors.

The test begins by estimating the parameters of interest under both FE and RE models and comparing the estimates. If there is no systematic difference between the FE and RE estimates, then it suggests that individual effects are not correlated with the regressors, and the RE model is preferred for its efficiency. If there is a systematic difference (i.e., the null hypothesis is rejected), then it suggests that the individual effects are correlated with the regressors, and the FE model is preferred for its consistency.

According to Ezie and Ezie (2021), the Hausman test uses the following test statistic:

$$H = (\hat{\beta}^{FE} - \hat{\beta}^{RE})'[Var(\hat{\beta}^{FE}) - Var(\hat{\beta}^{RE})]^{-1}(\hat{\beta}^{FE} - \hat{\beta}^{RE}) \sim \chi^2(k)$$

The formula says that a Hausman test examines if "the random effects estimate is insignificantly different from the unbiased fixed effect estimate" (Kennedy, 2008). If the null hypothesis of no correlation is rejected, you may conclude that individual effects u

are Significantly correlated with at least one regressors in the model and thus the random effect model is problematic. Therefore, you need to go for a fixed effect model rather than the random effect counterpart.

1.4 Model Selection: Fixed or Random Effect

When combining fixed vs. random effects, group vs. time effects, and one-way vs. two-way effects, we get 12 possible panel data models as shown in Table M3.3.1 In general, one-way models are often used mainly due to their parsimony, and a fixed effect model is easier than a random counterpart to estimate the model and interpret its result. It is not, however, easy to sort out the best one out of the following 12 models Nymoen (1991).

Table M3.3.1: Classification of Panel Data Analysis

	Type	Fixed Effect	Random Effect
One-way	Group	One-way fixed group effect	One-way random group effect
	Time	One-way fixed time effect	One-way random time effect
Two-way	Two groups*	Two-way fixed group effect	Two-way random group effect
	Two times*	Two-way fixed time effect	Two-way random time effect
	Mixed	Two-way fixed group & time effect	Two-way random group & time effect
		Two-way fixed time and random group effect	
		Two-way fixed group and random time effect	

*These models need two group (or time) variables (e.g., country and airline).

1.4.1 Substantive Meanings of Fixed and Random Effects

Specifically, the F-test compares a fixed effect model and (pooled) OLS, whereas the LM test contrasts a random effect model with OLS. The Hausman specification test compares fixed and random effect models. However, these tests do not provide substantive meanings of fixed and random effects. What does a fixed effect mean? How do we interpret a random effect substantively?

Here is a simple and rough answer. Suppose we are regressing the production of firms such as Apple, IBM, LG, and Sony on their R&D investment. A fixed effect might be interpreted as initial production capacities of these companies when no R&D investment is made; each firm has its own initial production capacity. A random effect might be viewed as a kind of consistency or stability of production. If the production of a company fluctuates up and down significantly, for example, its production is not stable (or its variance component is larger than those of other firms) even when its productivity (slope of R&D) remains the same across company.

Kennedy (2008: 282-286) provides theoretical and insightful explanation of fixed and random effects. Either fixed or random effect is an issue of unmeasured variables or omitted relevance variables, which renders the pooled OLS biased. This heterogeneity is handled by either putting in dummy variables to estimate individual intercepts of groups (entities) or viewing "the different intercepts as having been drawn from a bowl of possible intercepts, so they may be interpreted as random … and treated as though they were a part of the error term" (p. 284); they are fixed effect model and random effect model, respectively. A random effect model has a "composite error term" that consists of the traditional random error and a "random intercept" measuring the extent to which individual's intercept differs from the overall intercept (p. 284). He argues that the key difference between fixed and random effects is not whether unobserved heterogeneity is attributed to the intercept or variance components, but whether the individual specific error component is related to regressors.

1.5 Recommendations for Panel Data Modelling

The first recommendation, as in other data analysis processes, is to describe the data of interest carefully before analysis. Although often ignored in many data analyses, this data description is very important and useful for researchers to get ideas about data and analysis strategies. In panel data analysis, properties and quality of panel data influence model section significantly.

i. Clean the data by examining if they were measured in reliable and consistent manners. If different time periods were used in a long panel, for example, try to rearrange (aggregate) data to improve consistency. If there are many missing values, decide whether you go for a balanced panel by throwing away some pieces of usable information or keep all usable observations in an unbalanced panel at the expense of methodological and computational complication.

ii. Examine the properties of the panel data including the number of entities (individuals), the number of time periods, balanced versus unbalanced panel, and fixed versus rotating panel. Then, try to find models appropriate for those properties.

iii. Be careful if you have “long” or “short” panel data.

iv. Choosing Between Fixed and Random Effects Models: The choice between fixed effects (FE) and random effects (RE) models often hinges on the nature of the data and the research question being investigated. As discussed earlier, the Hausman test can guide this decision. Researchers should interpret the test results in conjunction with their understanding of the research question and the

underlying data. If the researcher suspects that unobserved, time-invariant characteristics may correlate with the other independent variables, then a fixed effects model might be preferred. If this correlation is not expected, a random effects model might be more efficient.

v. Addressing Autocorrelation and Heteroskedasticity: Just like in time-series and cross-sectional data, autocorrelation and heteroskedasticity can bias the standard errors in panel data models, leading to inaccurate inferences. To prevent this, it is recommended to perform diagnostic tests post-estimation, such as the Breusch-Pagan test for heteroskedasticity and the Wooldridge test for autocorrelation in panel data. If these issues are detected, they can often be addressed using robust standard errors or other corrective procedures.

vi. Taking into Account Non-Stationarity: The issue of non-stationarity, a common concern in time series analysis, can also arise in panel data. The presence of unit roots can lead to spurious regression results. Therefore, conducting unit root tests such as the Levin-Lin-Chu test or the Im-Pesaran-Shin test is important. If variables are found to be non-stationary, techniques like differencing or cointegration methods can be used.

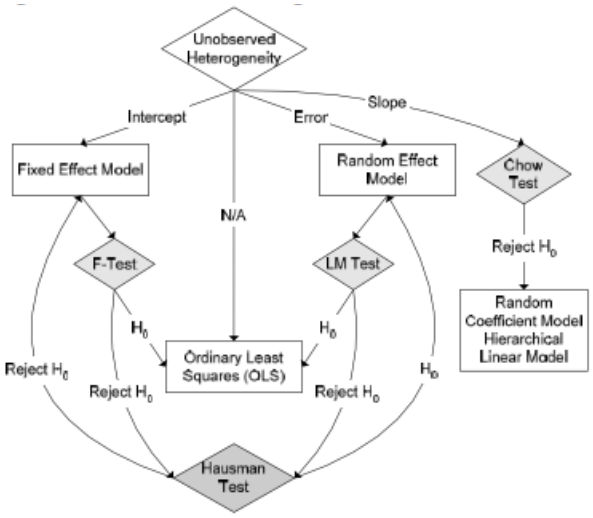
vii. Considering Cross-Sectional Dependence: If there are spill-over effects or interactions between cross-sectional units, ignoring them might lead to misleading results. It's important to test for contemporaneous correlation among the error terms of different cross-sections, and if detected, standard errors need to be corrected to make accurate inferences.

viii. Model Specification: Proper model specification is key to deriving accurate insights from panel data. Including relevant variables and considering potential interaction effects are essential. At the same time, it is necessary to beware of overfitting the model with too many variables relative to the number of observations.

1.5.1 Guidelines of Model Selection

On the modelling stage, let us begin with pooled OLS and then think critically about its potential problems if observed and unobserved heterogeneity (a set of missing relevant variables) is not taken into account. Also think about the source of heterogeneity (i.e., cross sectional or time series variables) to determine individual (entity Or group) effect or time effect. Figure M3.3.1 provides a big picture of the panel data modeling process Nymoen (1991).

Figure M3.3.1: Panel Data Modelling Process



If you think that the individual heterogeneity is captured in the disturbance term and the individual (group or time) effect is not correlated with any regressors, try a random effect model. If the heterogeneity can be dealt with individual specific intercepts and the individual effect may possibly be correlated with any regressors, try a fixed effect model. If each individual (group) has its own initial capacity and shares the same disturbance variance with other individuals, a fixed effect model is favored. If each individual has its own disturbance, a random effect will be better at figuring out heteroskedestic disturbances.

Next, conduct appropriate formal tests to examine individual group and/or time effects. If the null hypothesis of the LM test is rejected, a random effect model is better than the pooled OLS. If the null hypothesis of the F-test is rejected, a fixed effect model is favored over OLS.

If both hypotheses are not rejected, fit the pooled OLS.

Conduct the Hausman test when both hypotheses of the F-test and LM test are all rejected. If the null hypothesis of uncorrelation between an individual effect and regressors is rejected, go for the robust fixed effect model; otherwise, stick to the efficient random effect model.

If you have a strong belief that the heterogeneity involves two cross-sectional, two time series, or one cross-section and one time series variables, try two-way effect models. Double-check if your panel data are well-organized, and n and T are large enough; do not try a two-way model for a poorly organized, badly unbalanced, and/or too long/short

panel. Conduct appropriate F-test and LM test to examine the presence of two-way effects. Stata does not provide direct ways to fit two-way panel data models but it is not impossible. In Stata, two-way fixed effect models seem easier than two-way random effect models.

Finally, if you think that the heterogeneity entails slops (parameter estimates of regressors) varying across individual and/or time. Conduct a Chow test or equivalent to examine the poolability of the panel data. If the null hypothesis of poolable data is rejected, try a random coefficient model or hierarchical linear model.

1.6 Summary

In this unit, you learned how to conduct tests for fixed and random effects, Breusch-Pagan LM Test for Random Effect, Test for Comparing Fixed and Random Effect and guidelines of Model Selection. Fixed-effects models and alternatives are part of panel data model. The limitation or hidden assumption of each fixed-effects model, such as the individual fixed-effects model, time fixed-effects model, and the individual-time fixed-effects model, are discussed in detail.

Tutor Marked Assignment

What is fixed effect?

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Illorin, Nigeria.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.
Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.

1.8 Possible Answers to Self-Assessment Exercise(s) Within the Content

Answer to Self- Assessment 1

To run the test, one need to compare fixed effects to the simple common constant OLS method using the conventional F-test. The null hypothesis states that, all of the constants are the same (homogeneity) and that, the common constant approach can be used.

$$H_0 : a_1 = a_2 = ...a_N$$

The F-statistic is:

$$F = \frac{\left(R^2_{FE} - R^2_{CC}\right)/(N-1)}{\left(1-R^2_{FE}\right)/)NT-N-k)} \square F(N-1, NT-N-k)$$

Compare the F-statistic to the Critical Value: You compare your calculated F-statistic to the critical value from the F-distribution with (Number of groups - 1) and (Total number of observations - Number of groups*Number of time periods) degrees of freedom. If your calculated F-statistic is greater than the critical value, you reject the null hypothesis of no fixed effects.

Conclusion: If the null hypothesis is rejected, it indicates that the group-specific fixed effects are statistically significant, i.e., there are significant differences between the group-specific intercepts, and hence, using a fixed effects model would be more appropriate than a pooled OLS model.

UNIT4: PANEL DATA ESTIMATION IN EVIEWS

1.1 Introduction

In the previous unit of you learnt how to conduct tests for fixed and random effects, Breusch-Pagan LM Test for Random Effect, Test for Comparing Fixed and Random

Effect and guidelines of Model Selection. In the present unit we shall discuss panel data estimation in EViews.

1.2 Learning Outcomes

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

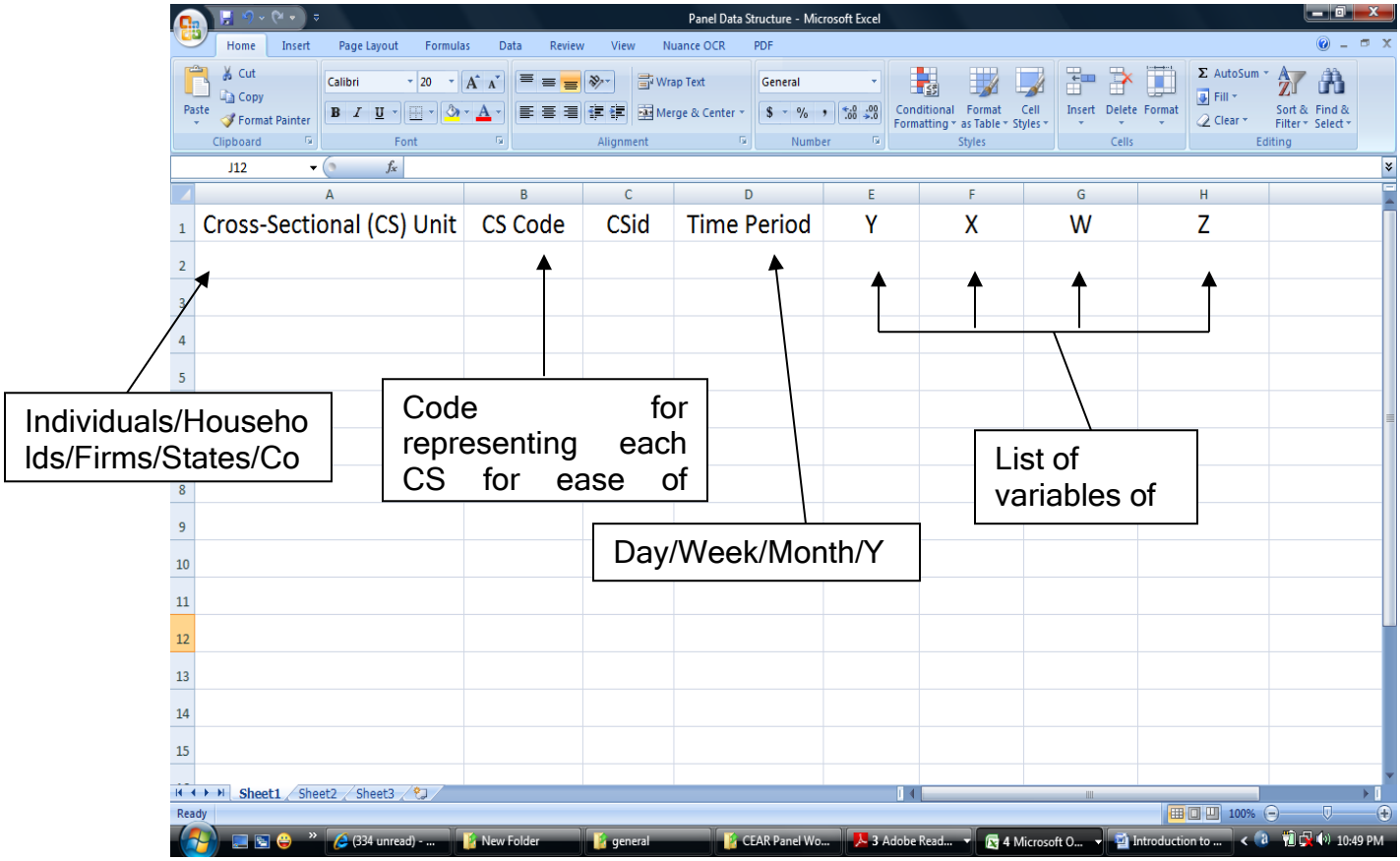
- List the steps in estimating a Panel Equation in Eviews
- Practicalise the Empirical Application: Pooled Regression
- Evaluate the empirical Application: Fixed Effects Regression
- Apply the empirical Application: Random Effects Regression

1.3 Steps in Estimating a Panel Equation in EViews 10

1.3.1 Data Structure

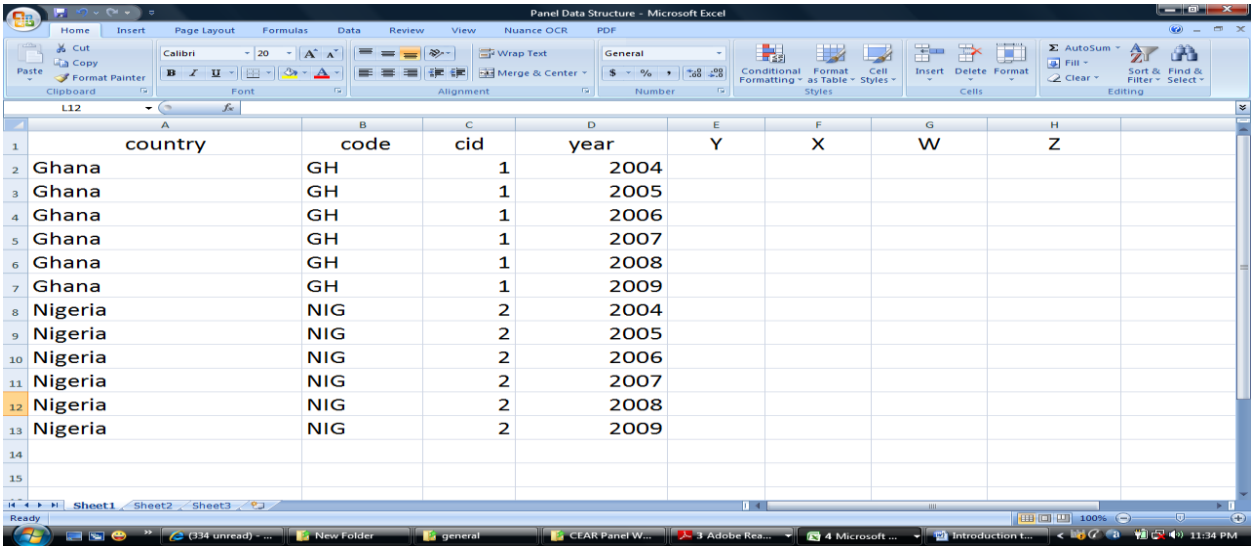
Step 1: It is convenient to start with an empty Excel Sheet. We provide below a figure showing a fresh excel sheet with appropriate labels.

Panel Data Structure in Excel Sheet



For example, if CS=countries= Ghana, Nigeria and Time period = years= 2004 – 2009 with the same set of variables as shown above, then we have:

Panel Data for CS=countries= Ghana, Nigeria and Time period = years= 2004 – 2009



country	code	cid	year	Y	X	W	Z
Ghana	GH	1	2004				
Ghana	GH	1	2005				
Ghana	GH	1	2006				
Ghana	GH	1	2007				
Ghana	GH	1	2008				
Ghana	GH	1	2009				
Nigeria	NIG	2	2004				
Nigeria	NIG	2	2005				
Nigeria	NIG	2	2006				
Nigeria	NIG	2	2007				
Nigeria	NIG	2	2008				
Nigeria	NIG	2	2009				

Let us arrange the following time series data in tables 1- 10 for selected SSA in panel structure in excel worksheet. Note that llr=Liquid Liabilities as a percentage of GDP; pcr = Private Credit by Deposit Money Banks as a percentage of GDP; bdr=Bank Deposits as a percentage of GDP; and rgdp=GDP per capita (constant 2000 US$). **Also, all the countries have the same units of measurement. This is very important in panel to make meaningful comparative analyses.**

Table M3.4.1: Time Series for Togo

Year	Llr	Pcr	bdr	Rgdp
1996	23.0315	16.0748	15.0802	250.4119
1997	20.0072	15.3494	13.713	276.0873
1998	21.3712	17.1686	14.4285	259.7281
1999	21.1455	15.905	13.4929	256.5784
2000	24.5442	15.6822	14.9995	245.9931
2001	26.1392	15.3362	16.2111	237.9157
2002	23.95	13.3844	16.2703	240.5463
2003	23.862	14.2132	18.3394	240.2094
2004	26.4359	15.8993	20.6205	240.7338
2005	28.478	16.9002	22.0857	237.1547

Table M3.4.2: Time Series for Tunisia

Year	Llr	pcr	bdr	rgdp
1996	46.0225	48.8067	36.0321	1743.74
1997	48.1615	47.5828	38.2282	1813.536
1998	49.4405	48.7962	39.3322	1876.205
1999	50.9677	48.8294	40.3478	1963.991
2000	54.605	53.1492	43.7798	2033.071
2001	56.1331	57.1981	45.5335	2108.913
2002	58.4117	59.5272	47.2955	2120.055
2003	57.1218	58.3255	46.1847	2224.77
2004	56.8293	58.2998	46.4354	2337.122

2005	58.9073	60.0528	48.4716	2412.397

Table M3.4.3: Time Series for Tanzania

Year	Llr	pcr	bdr	Rgdp
1996	21.9841	4.4241	14.678	253.0263
1997	19.5539	3.0814	13.1927	255.4859
1998	18.3521	3.7249	12.5247	258.7052
1999	18.0459	4.2732	12.2517	261.5302
2000	18.5975	4.4316	12.8029	268.2303
2001	18.9531	4.5071	13.6432	277.8862
2002	19.9563	5.1765	14.7567	290.4531
2003	21.5444	6.6206	16.1864	299.0575
2004	21.0916	7.4539	15.9761	311.0443
2005	24.6267	8.9208	18.9296	323.8417

Table M3.4.4: Time Series for Uganda

Year	Llr	pcr	bdr	Rgdp
1996	10.9502	4.4198	7.6369	214.9474
1997	11.7424	4.5096	8.5124	219.2205
1998	12.5713	4.5914	9.3281	223.2856
1999	13.6531	5.208	10.1499	234.2139
2000	14.2259	5.2279	10.6857	240.0347
2001	15.31	5.0121	11.6458	244.2106
2002	16.7173	4.9609	12.8386	251.7746
2003	17.658	5.319	13.6737	255.3492
2004	17.6943	5.5014	13.7752	260.6697
2005	17.8651	5.3874	13.8932	269.2797

Table M3.4.5: Time Series for South Africa

Year	Llr	pcr	bdr	Rgdp
1996	47.5964	57.0558	45.0357	3019.966
1997	50.2232	60.0383	47.6385	3029.774
1998	53.2823	63.5497	50.7814	2974.682
1999	54.4287	65.1732	51.8763	2972.204
2000	52.7026	64.996	50.1436	3019.947
2001	48.3147	69.8244	50.8579	3046.309
2002	42.4572	63.777	50.0018	3127.809
2003	41.9225	59.6852	51.3644	3186.26
2004	40.384	62.1936	51.3092	3301.071

2005	41.4266	65.3924	53.2743	3428.973

Table M3.4.6: Time Series for Niger

Year	Llr	pcr	bdr	Rgdp
1996	12.8354	4.2251	6.9918	167.9679
1997	10.5011	3.7052	5.8092	166.455
1998	7.4821	3.5422	4.6794	177.246
1999	6.914	3.8822	4.5292	169.9728
2000	8.425	4.6702	5.5463	161.666
2001	9.3728	4.9679	6.182	167.1001
2002	9.7845	5.084	6.5798	166.144
2003	10.827	5.2062	6.7173	167.4735
2004	13.8676	5.9907	7.9314	160.7414
2005	14.0486	6.4895	8.0553	166.3869

Table M3.4.7: Time Series for Nigeria

Year	Llr	pcr	bdr	Rgdp
1996	12.8536	8.4201	8.1133	367.982
1997	13.8668	9.7126	9.097	367.7034
1998	16.7043	11.8227	11.1183	364.6282
1999	18.7888	12.4598	13.157	358.9527
2000	17.5213	10.416	12.6448	368.5392
2001	25.1358	14.8838	18.0648	370.2718
2002	26.7725	16.09	19.3215	366.5663
2003	24.7666	14.5882	17.2996	395.7558
2004	24.407	15.3438	17.1914	409.3429
2005	17.3979	12.1546	13.121	428.4006

Table M3.4.8: Time Series for Kenya

Year	Llr	pcr	bdr	Rgdp
1996	37.9709	19.5837	27.572	422.6506
1997	39.4461	21.9409	30.3111	413.4464
1998	39.0445	23.3882	30.5958	416.0453
1999	38.4162	24.9332	29.9474	414.6931
2000	37.4519	25.5872	29.2896	406.5445
2001	36.3671	24.0622	28.9476	411.1291
2002	37.7525	23.503	30.2175	402.8055
2003	39.4971	23.0859	31.6685	403.9767

2004	39.0288	23.2237	31.7099	413.5642
2005	38.448	23.8282	31.6453	425.8653

Table M3.4.9: Time Series for Morocco

Year	Llr	pcr	bdr	Rgdp
1996	70.75	27.2908	46.3059	1280.803
1997	72.3646	37.7556	52.9304	1233.059
1998	69.4816	46.5368	55.5144	1308.17
1999	74.1824	50.5895	57.9418	1296.227
2000	79.5742	54.1697	62.6492	1301.872
2001	81.045	52.7253	64.4227	1382.973
2002	86.9772	53.3611	69.1398	1411.385
2003	88.3486	53.2476	70.6133	1480.456
2004	90.6856	54.2893	72.8972	1540.83
2005	97.1707	57.7634	78.3015	1561.895

Table M3.4.10: Time Series for Botswana

Year	Llr	pcr	bdr	Rgdp
1996	20.882	11.3541	18.8731	2725.958
1997	20.7326	9.7002	18.9318	2938.997
1998	24.4001	10.5208	22.6142	3185.879
1999	27.409	12.6244	25.3974	3354.679
2000	26.0662	14.0234	23.9447	3572.956
2001	24.8806	13.9019	23.2918	3706.873
2002	27.9743	16.6799	26.5515	3867.929
2003	27.6	17.9923	26.1646	4060.682
2004	29.4949	19.6398	27.353	4264.321
2005	28.1966	19.3561	26.1213	4382.497

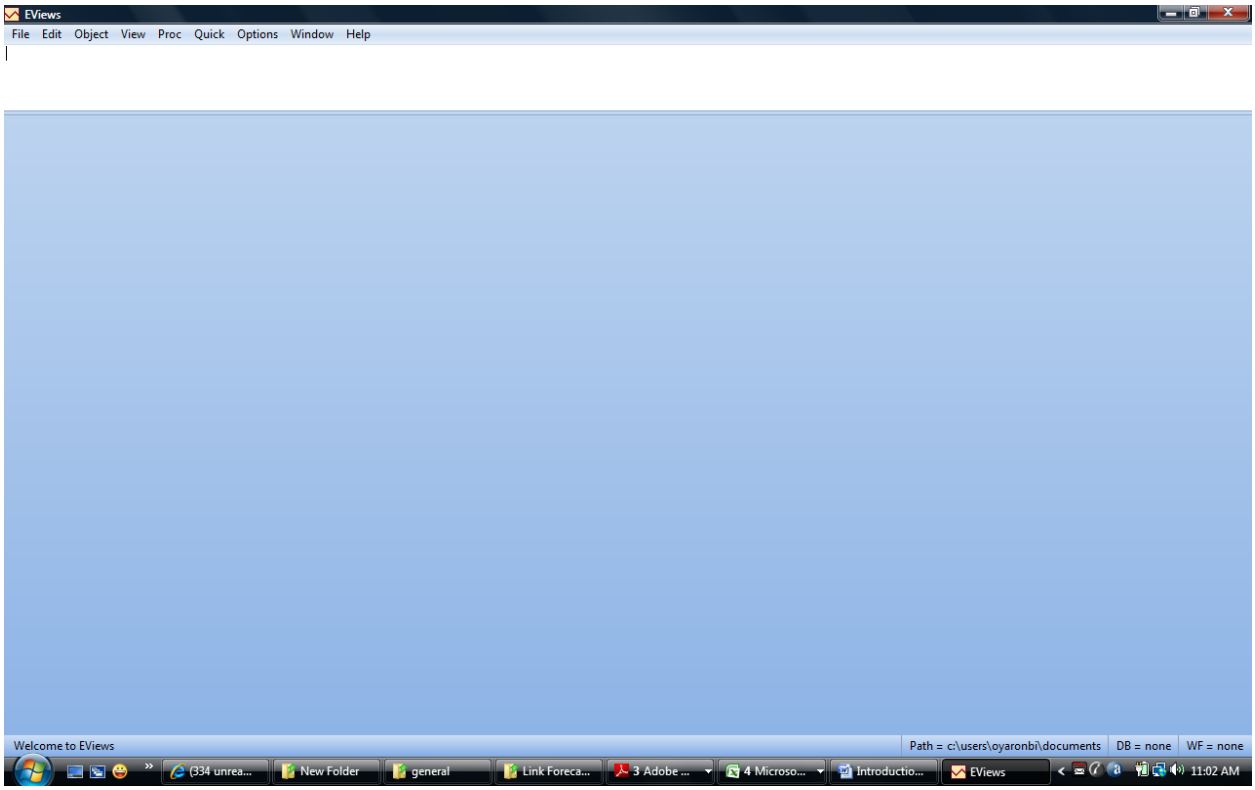
1.3.2 Data Importation

Step 2: Save your panel data generated from tables 1 – 10 above in a directory you can easily remember (say Desktop) and ensure the excel worksheet is in 2003 version for convenience. Kindly note the file name for your worksheet. This is important for loading data into Eviews for estimation purpose.

Step 3: Load data into Eviews. This requires the following sub-steps:

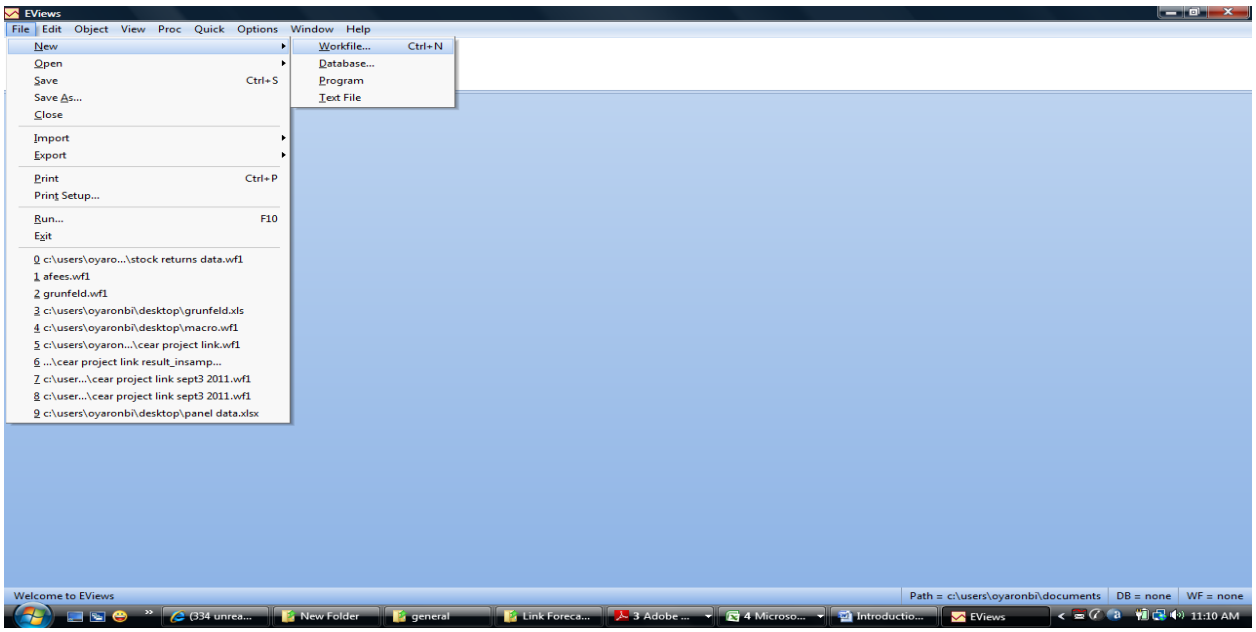
a. View your data in the excel sheet to ensure they are structured in panel data form.

b. Click on your Eviews icon. You will find the Eviews window as shown below:

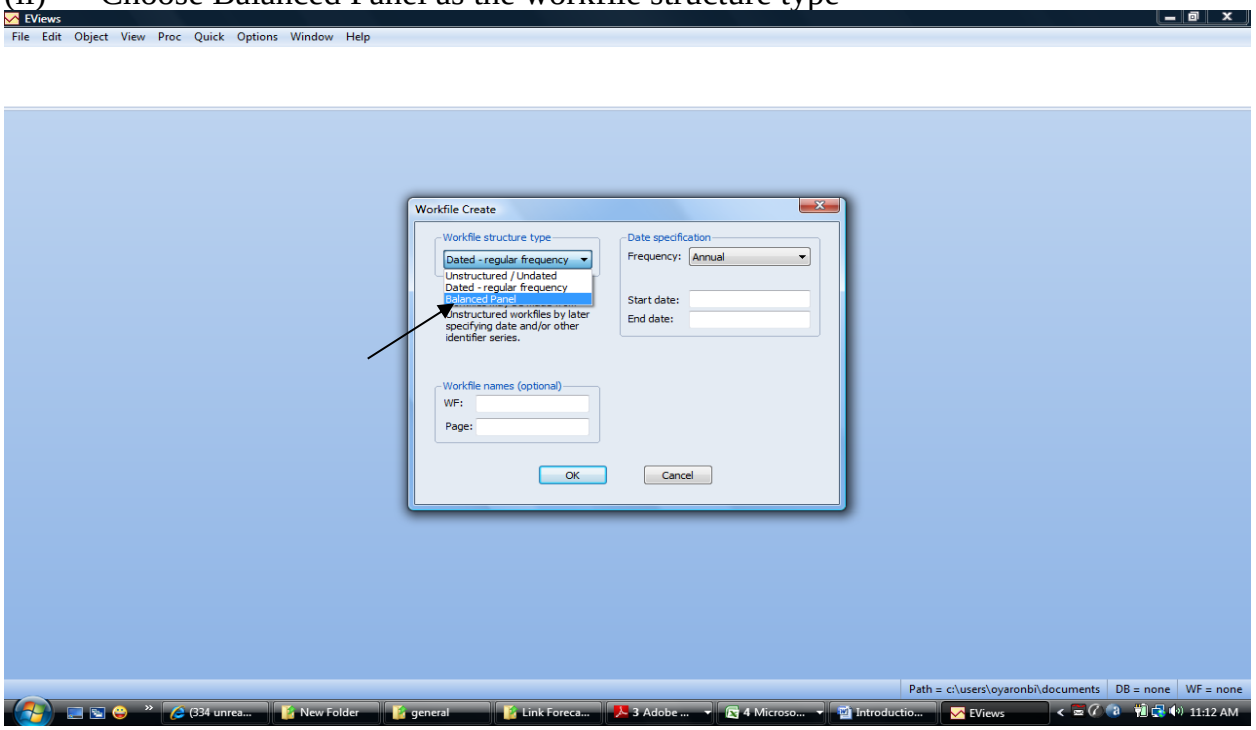


c. Define the structure of the Panel Data in Eviews. The following pictures show step-by-step procedure on how to do this.

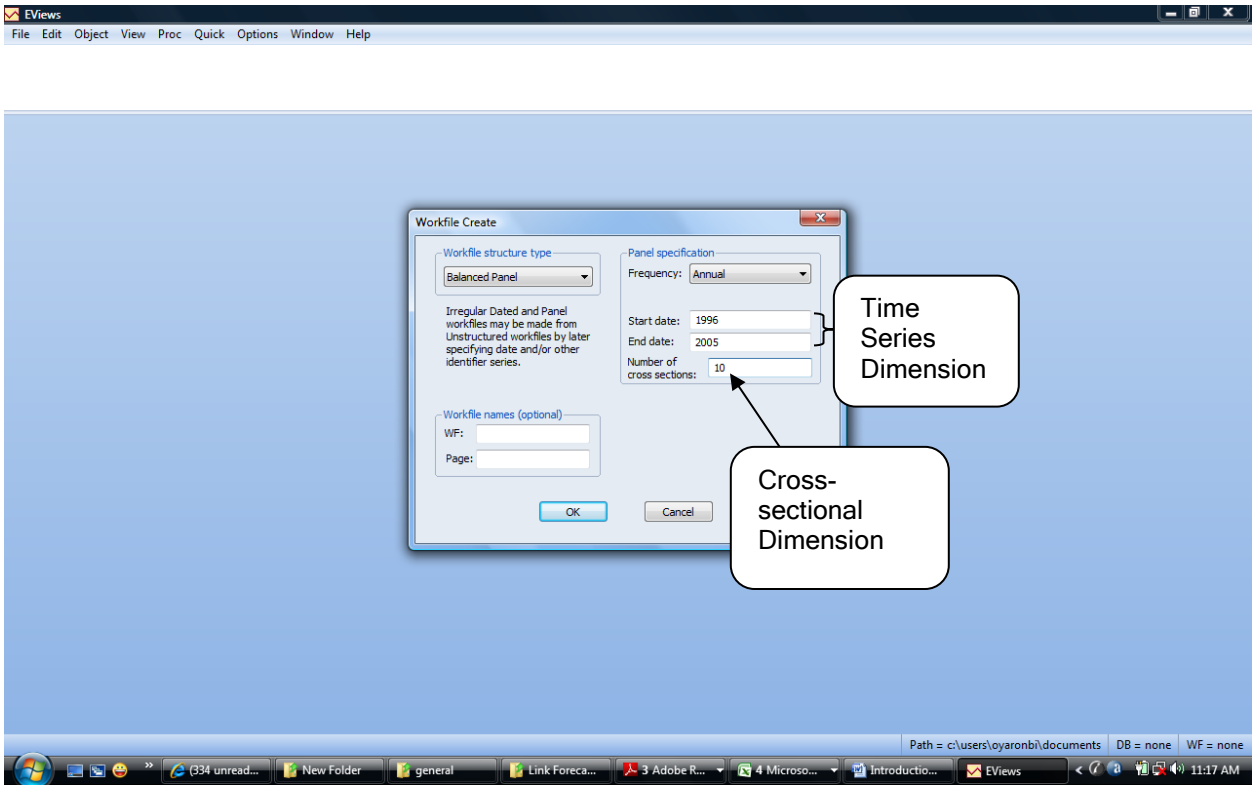
(i) Opening the Eviews Workfile



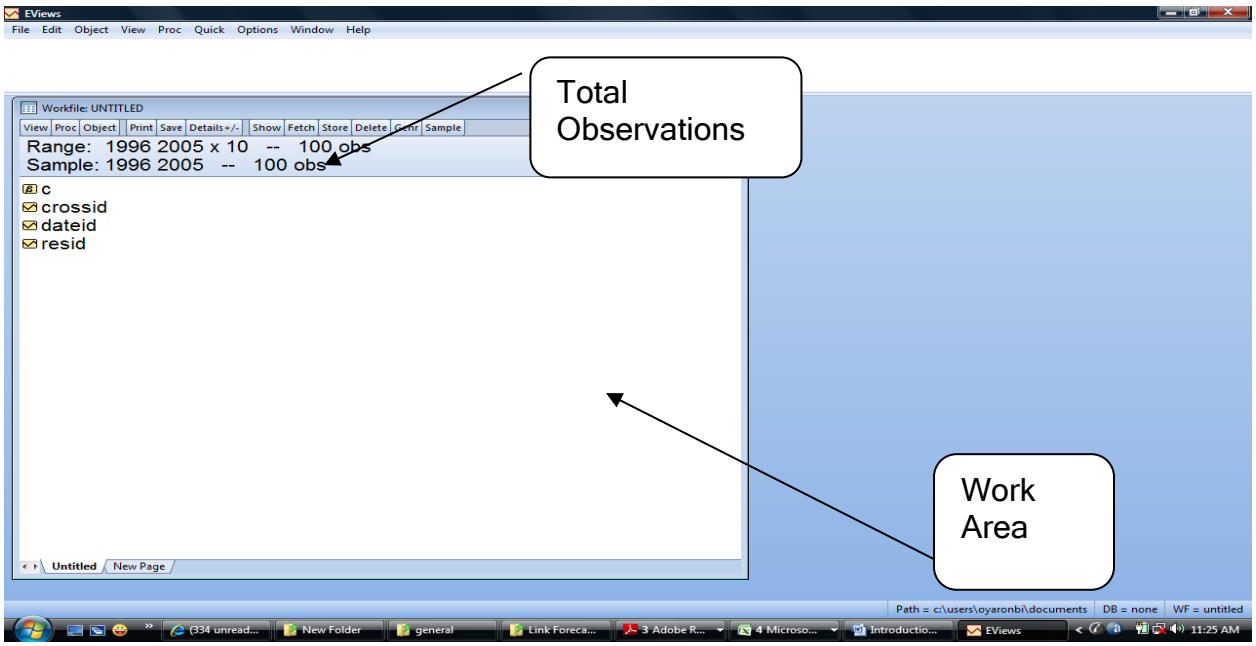
(ii) Choose Balanced Panel as the workfile structure type



(iii) By choosing balanced panel it allows you to indicate the time and cross-section dimensions for your panel data as shown below:

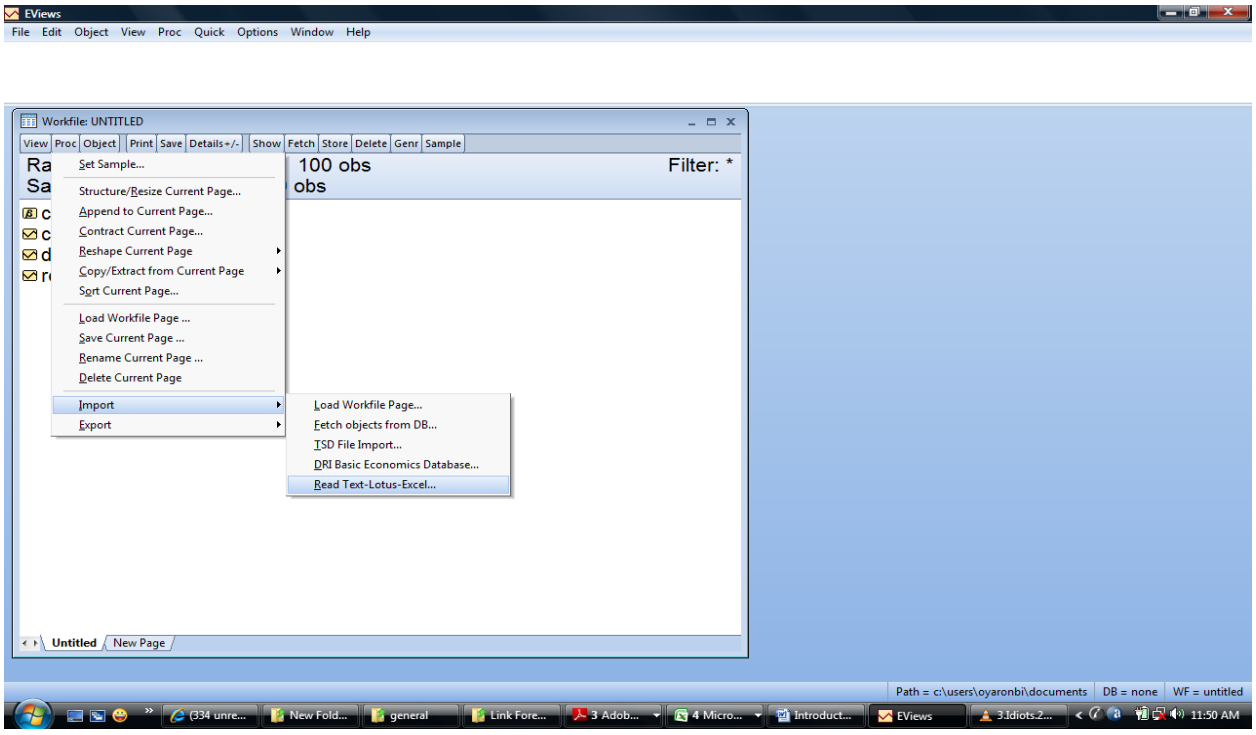


(iv) By clicking *OK* in (iii) above, you will find the Eviews work area in panel data form

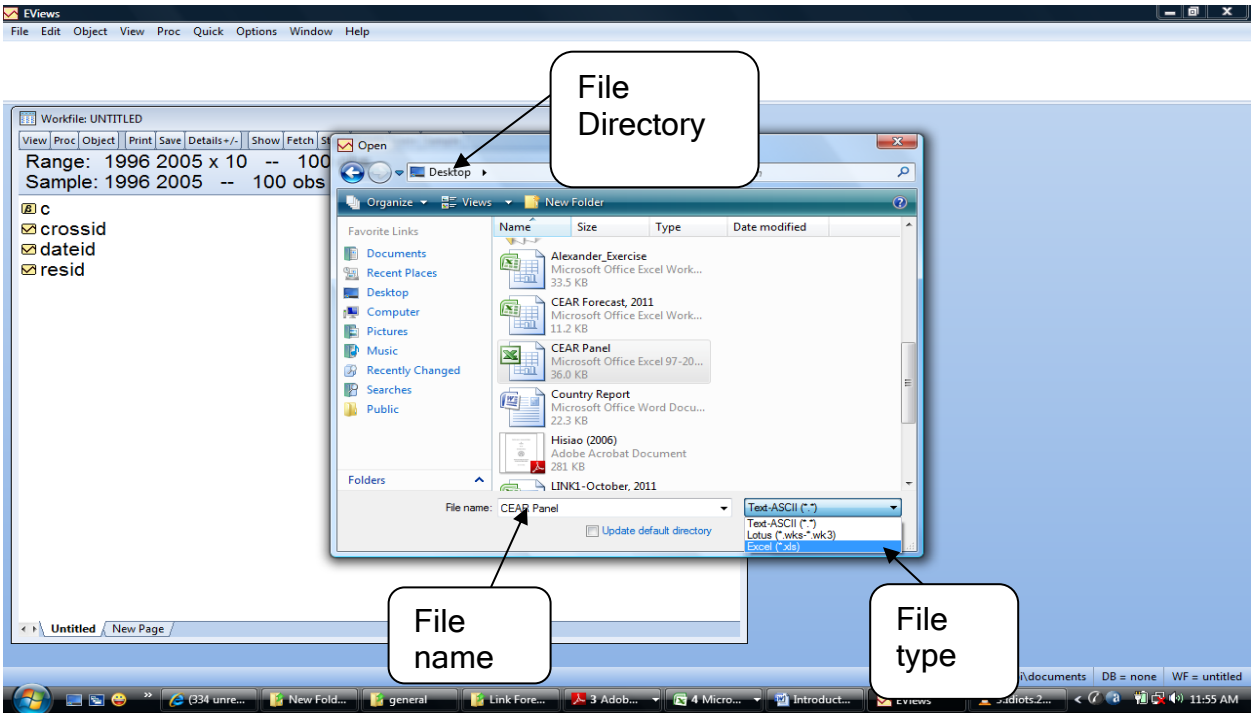


d. Import panel data from the excel file: Although, there are several methods of loading data into Eviews, for the purpose of this exercise however, we will consider the import method. Participants may explore other methods later. This also involves a number of sub-steps as follows:

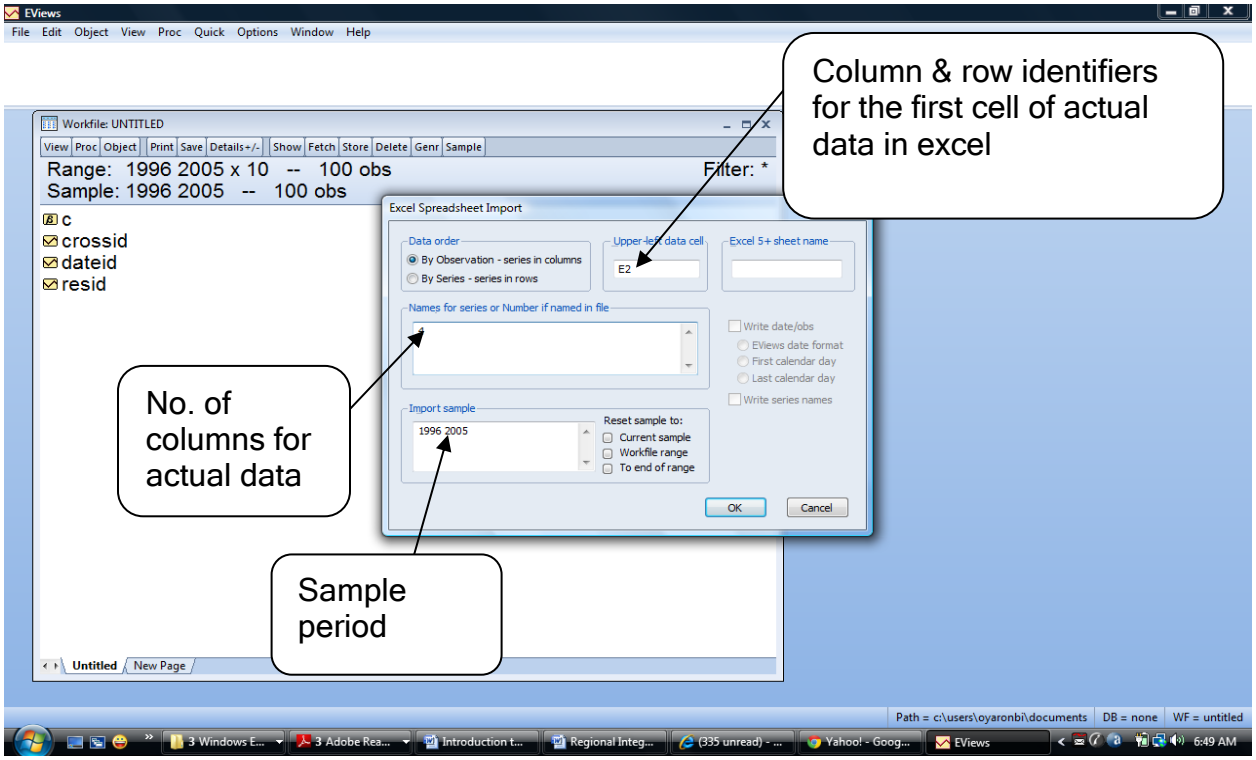
(i) Choose import method as shown below



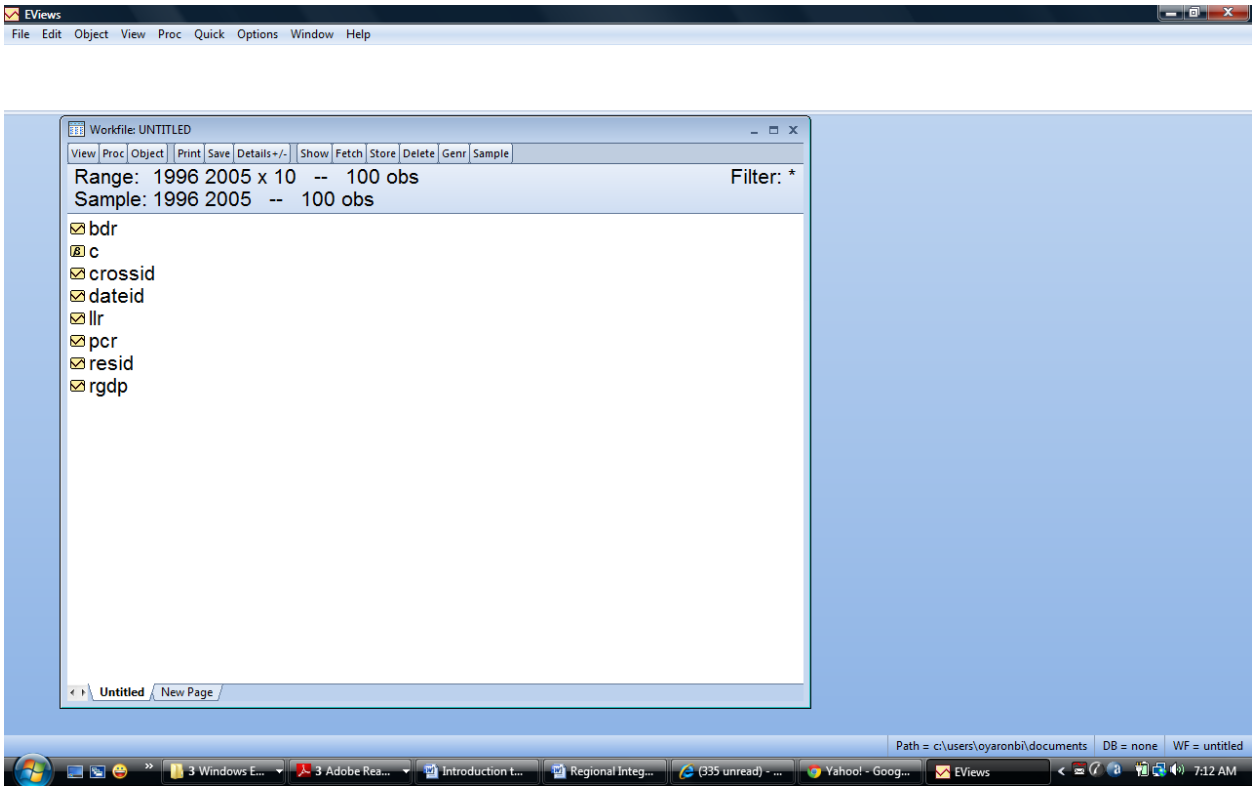
(ii) By clicking on *Read Text-Lotus-Excel* in (di) above, you will find a box prompting you to indicate the excel file name containing the panel data. Ensure the excel file is not opened before you import the data.



(iii) Following the information provided in (dii) above and clicking *ok*, you will find the import box below:

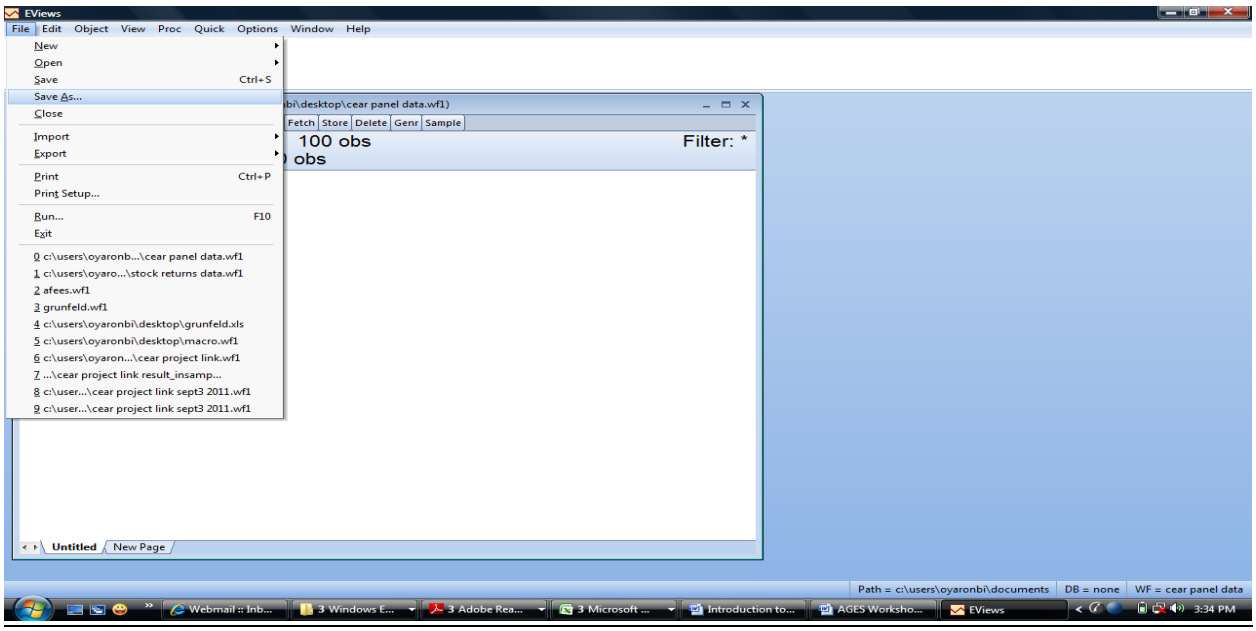


(iv) By clicking *ok* after providing necessary information in (diii) above, the Eviews work area is updated to reflect the data imported. This is shown below:



When you compare (div) with (civ), you will find that the former contains data on variables intended to be modelled in panel data form while the latter is an empty Eviews workfile.

(e) At this juncture, save your Eviews workfile to secure the data



Also note the file name for your Eviews workfile and its directory. This workfile will be recalled under Panel data models section.

1.4 Empirical Application: Pooled Regression

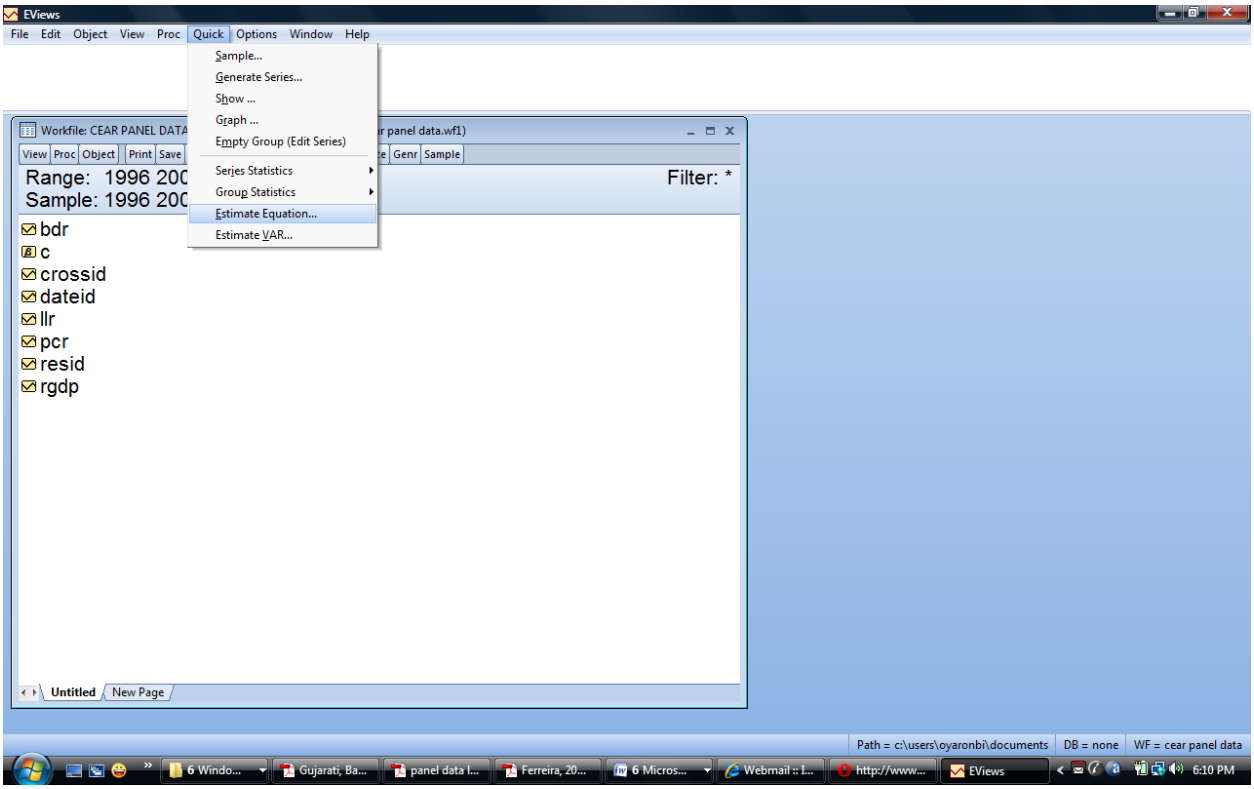
We are using the panel data in the saved Eviews workfile to illustrate the pooled regression. The Panel data regression model is given as:

$$rgdp_{it} = \alpha + \beta_1 llr_{it} + \beta_2 pcr_{it} + \beta_3 bdr_{it} + u_{it}$$

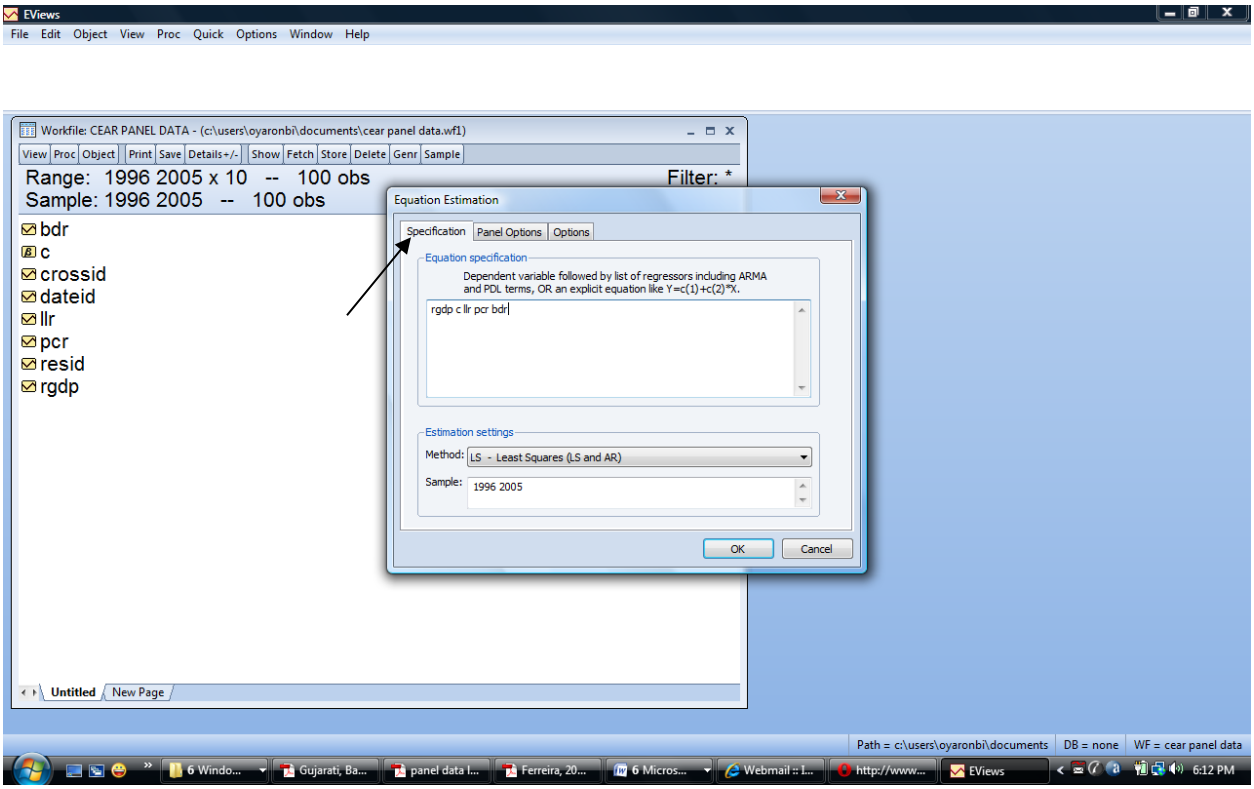
We have carefully structured the procedure for estimating the model in Eviews in the following pictures below:

Model specification and estimation

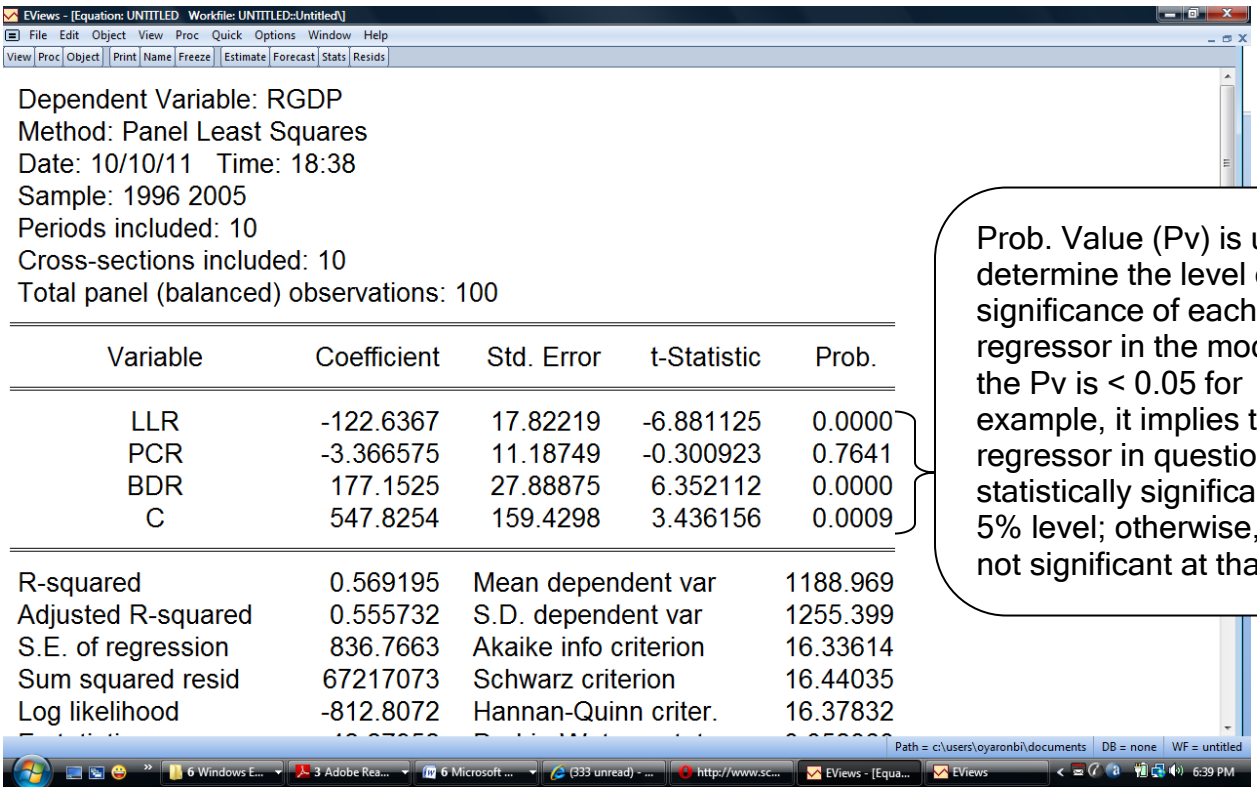
(a) Equation Editor



(b) Pooled Regression



(c) Pooled Regression Results



Interpretation of Regression Results

Based on the probability values, llr and bdr are statistically significant while pcr is not. Since the regressors in the model are in percentages, therefore, a 1% increase in llr leads to a significant reduction in rgdp by 122.64USD on the average. Conversely, a 1% increase in bdr leads to a significant rise in bdr by 177.15USD on the average. Although, the effect of pcr is not statistically significant, its increase by 1% has the potential effect of reducing rgdp by 3.37USD.

Note that llr is a typical measure of financial depth in economy and thus of the overall size of the financial sector; pcr is a measure of the activity of financial intermediaries (i.e. ability to channell savings to investors) and bdr is a measure of size and efficiency of the financial intermediaries as it captures the ability of the banks to mobilize funds from the public (the surplus unit). Based on these definitions, the economic interpretations can also be offered.

The two dominant approaches used when the unobserved specific effects are significant in a panel data model are the fixed effects regressions and the random effects regression models. We consider these two in the next sections.

1.5 Empirical Application: Fixed Effects and Random Effect Regression

The Hausman test is a statistical hypothesis test that determines whether the coefficients in a model are consistent, particularly in the context of choosing between a fixed effects model and a random effects model in panel data analysis.

The null and alternative hypotheses for the Hausman test are as follows:

1. **Null Hypothesis (H0)**: The random effects estimator and the fixed effects estimator are statistically equivalent. This essentially implies that the unique errors are not correlated with the regressors, meaning that the more efficient random effects estimator can be preferred over the fixed effects estimator.
2. **Alternative Hypothesis (HA)**: The random effects estimator and the fixed effects estimator are not statistically equivalent, suggesting that the unique errors are correlated with the regressors. In this case, the fixed effects model should be used as it provides consistent estimates while the random effects estimator does not.

In practice, if the p-value from the Hausman test is small (e.g., less than a significance level of 0.05), we reject the null hypothesis and conclude that we should use a fixed effects model. If the p-value is large, we fail to reject the null hypothesis, which means that it's safe to use the random effects model.

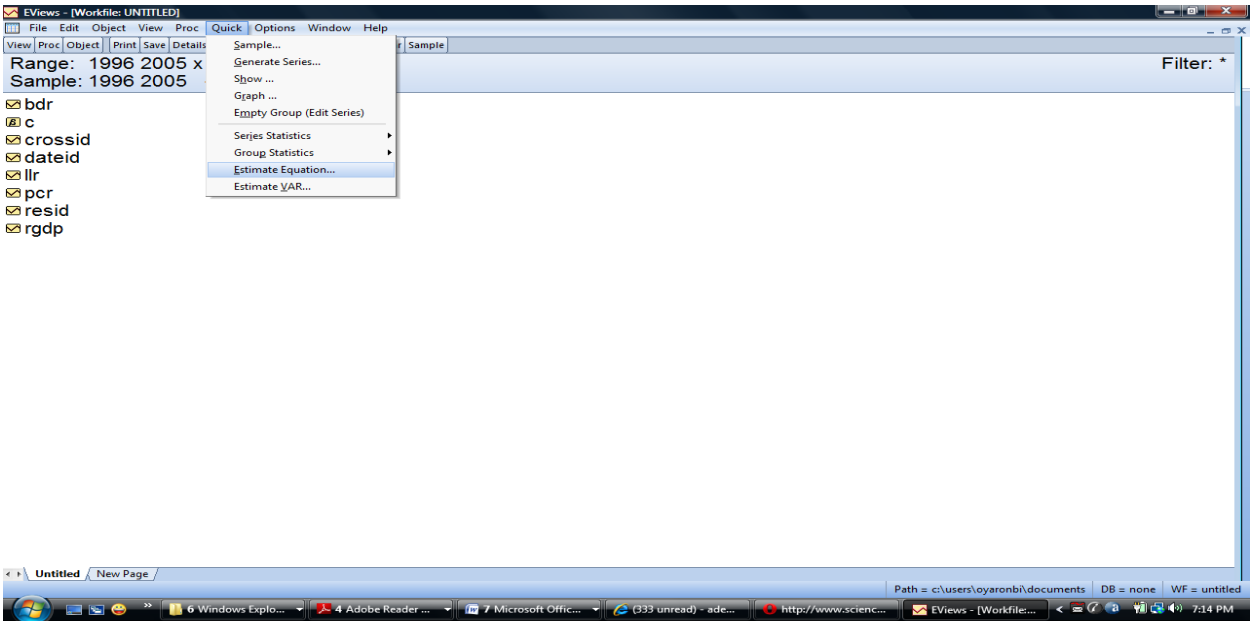
1.5.1 Empirical Application: Fixed Effects Regression

We are still using the panel data in the Eviews workfile created to illustrate the Fixed

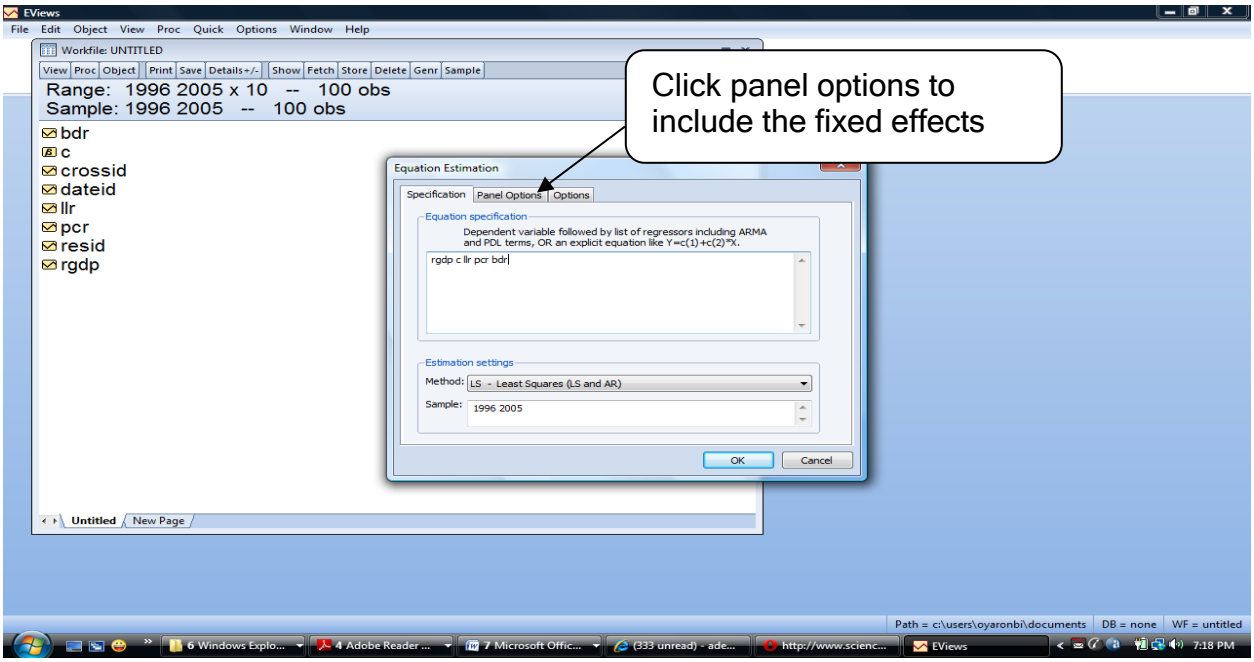
Effects Regression. The Panel data regression model with fixed effects is given as:

$$rgdp_{it} = \alpha + \beta_1 llr_{it} + \beta_2 pcr_{it} + \beta_3 bdr_{it} + \mu_i + \varepsilon_{it}$$

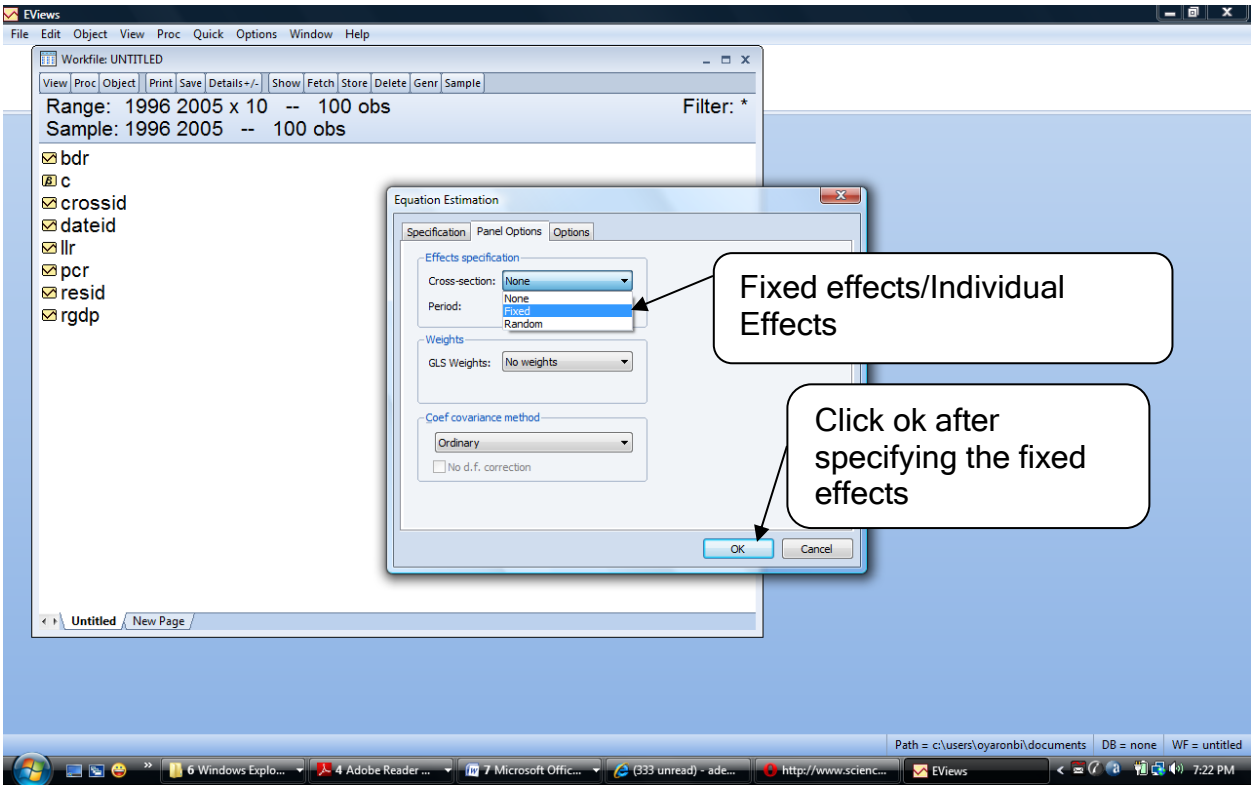
(a) Equation editor



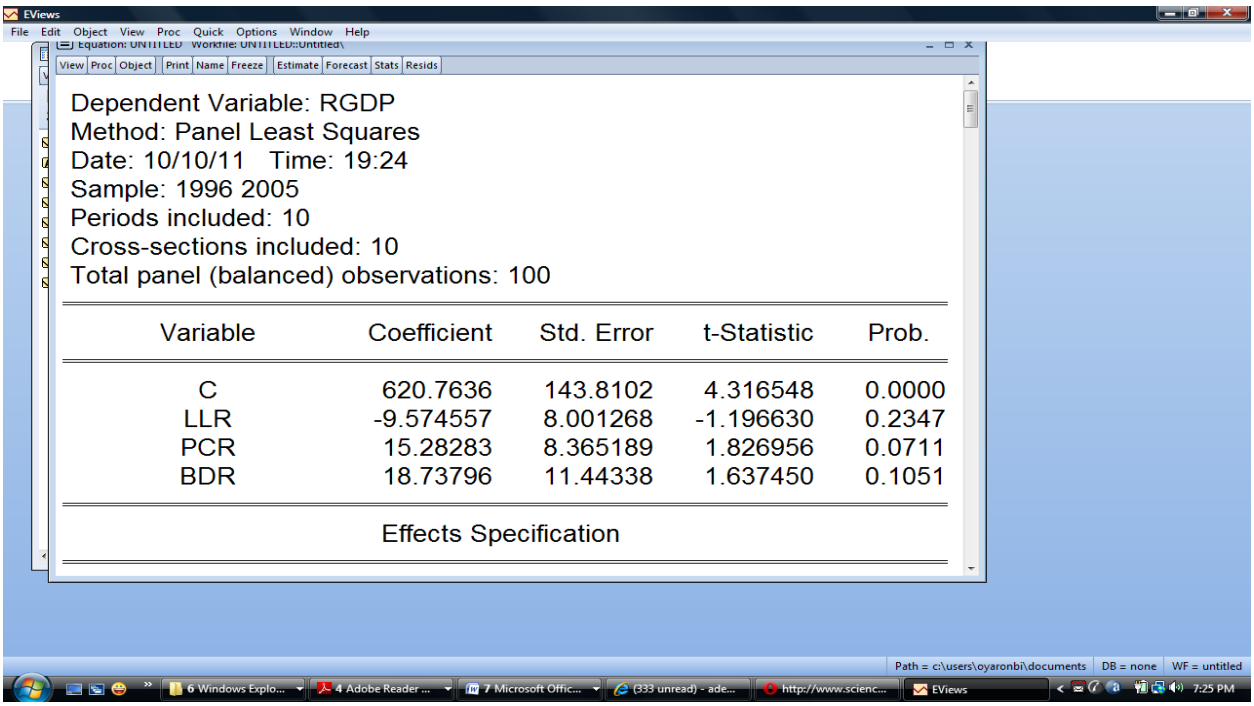
(b) Model Specification



(c) Specifying the fixed effects

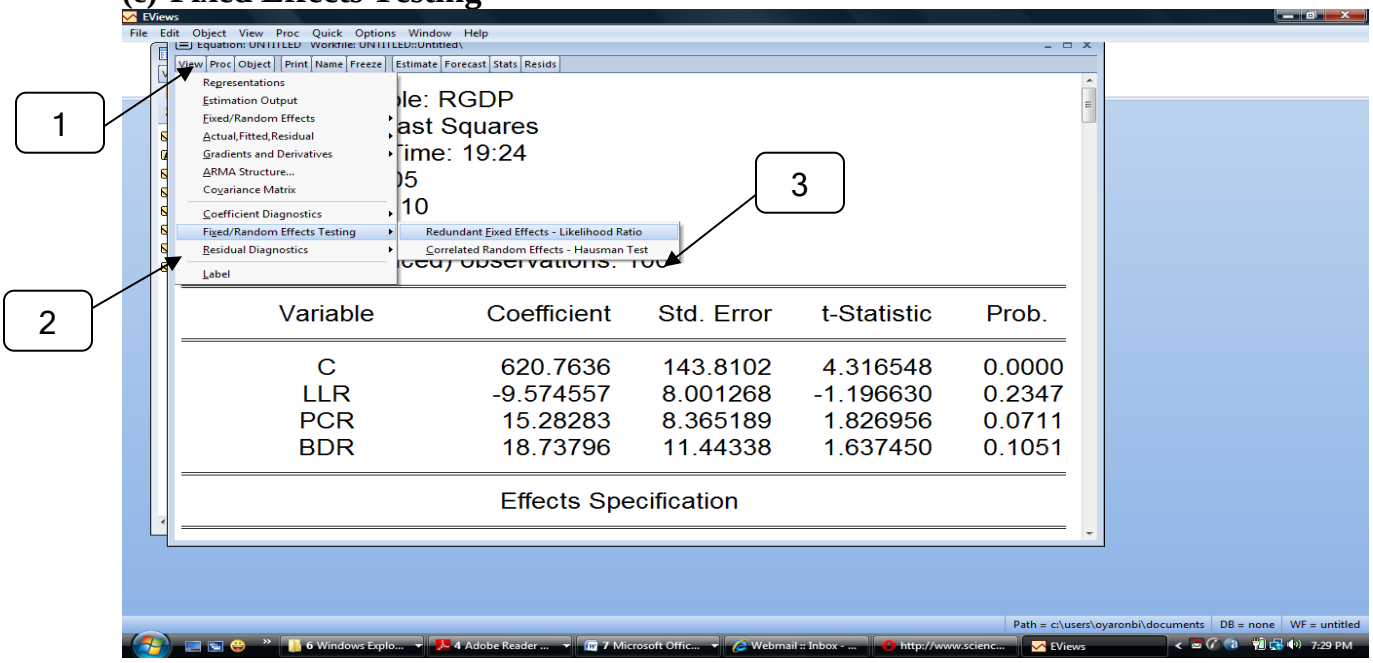


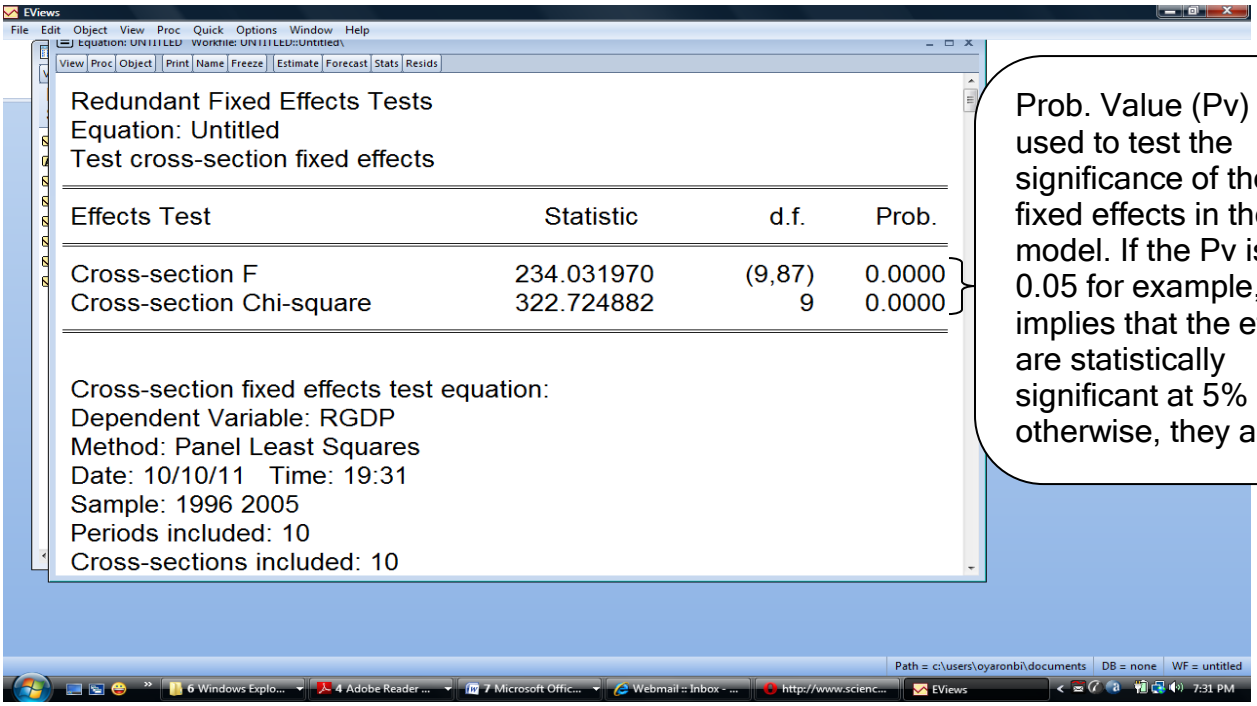
(d) Fixed Effects Regression



The interpretation of regression results here is not different from the style used under the pooled regression. However, we need to further test whether the inclusion of the effects is significant or not.

(e) Fixed Effects Testing





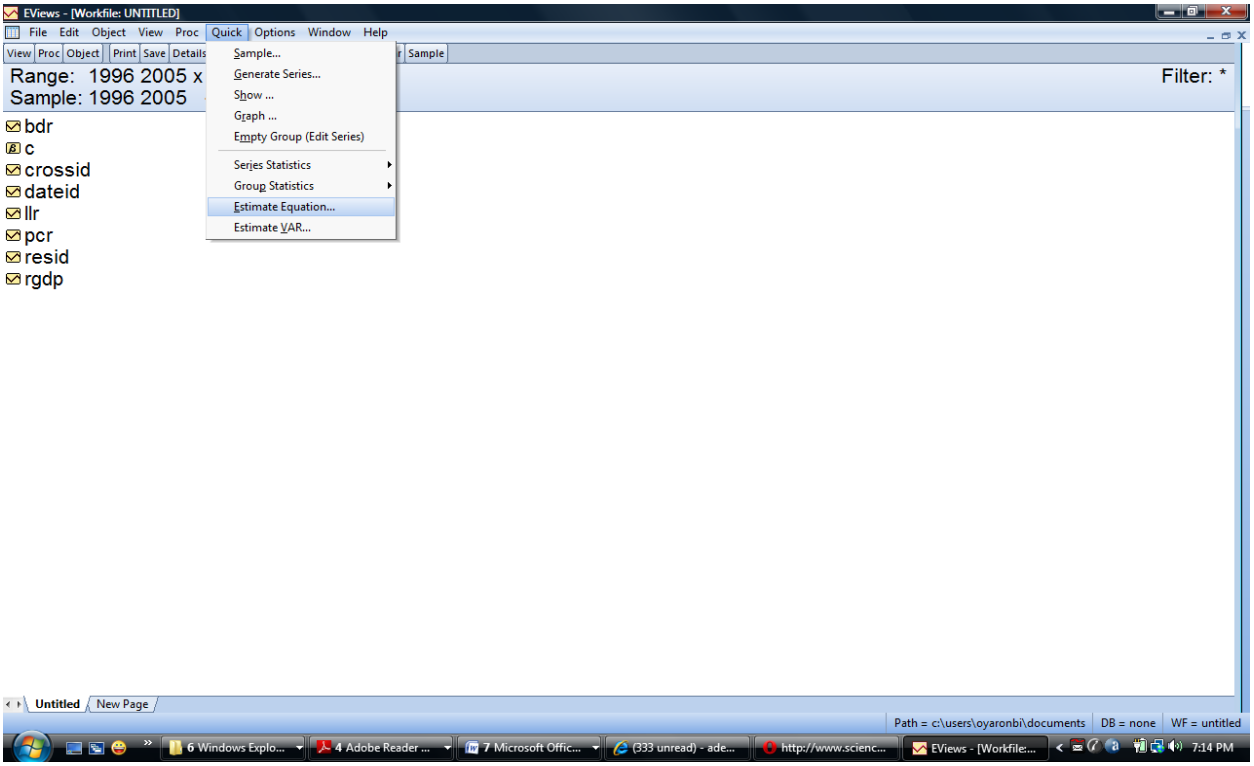
Recall that the null hypothesis for the fixed effect test is that the fixed effects are not important in the model. Based on the Pv, we can conclude that the fixed effects are significant in the model and, therefore, estimating without these effects which is the case in pooled regression will yield biased standard errors and policy prescriptions drawn from the analysis will be invalid.

1.5.2 Empirical Application: Random Effects Regression

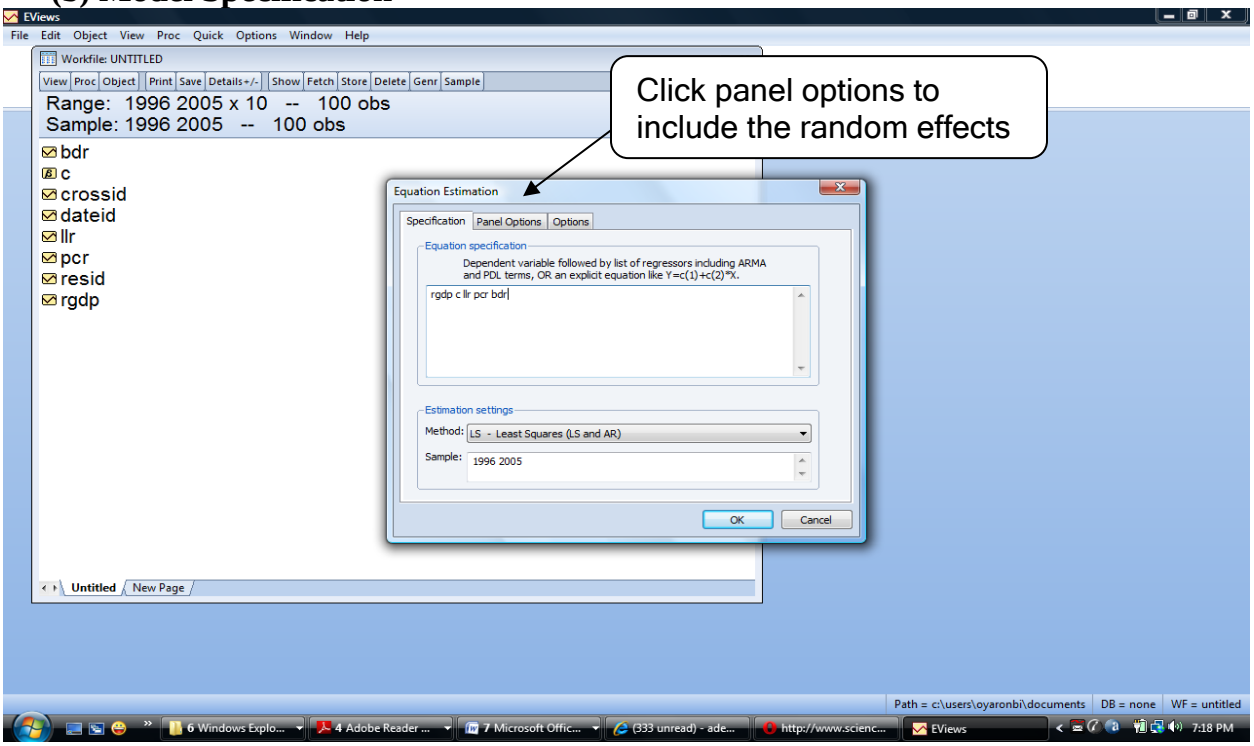
Again, we are still using the panel data in the Eviews workfile created to illustrate the random effects regression. The Panel data regression model with random effects is given as:

$$rgdp_{it} = \alpha + \beta_1 llr_{it} + \beta_2 pcr_{it} + \beta_3 bdr_{it} + \mu_i + v_{it}$$

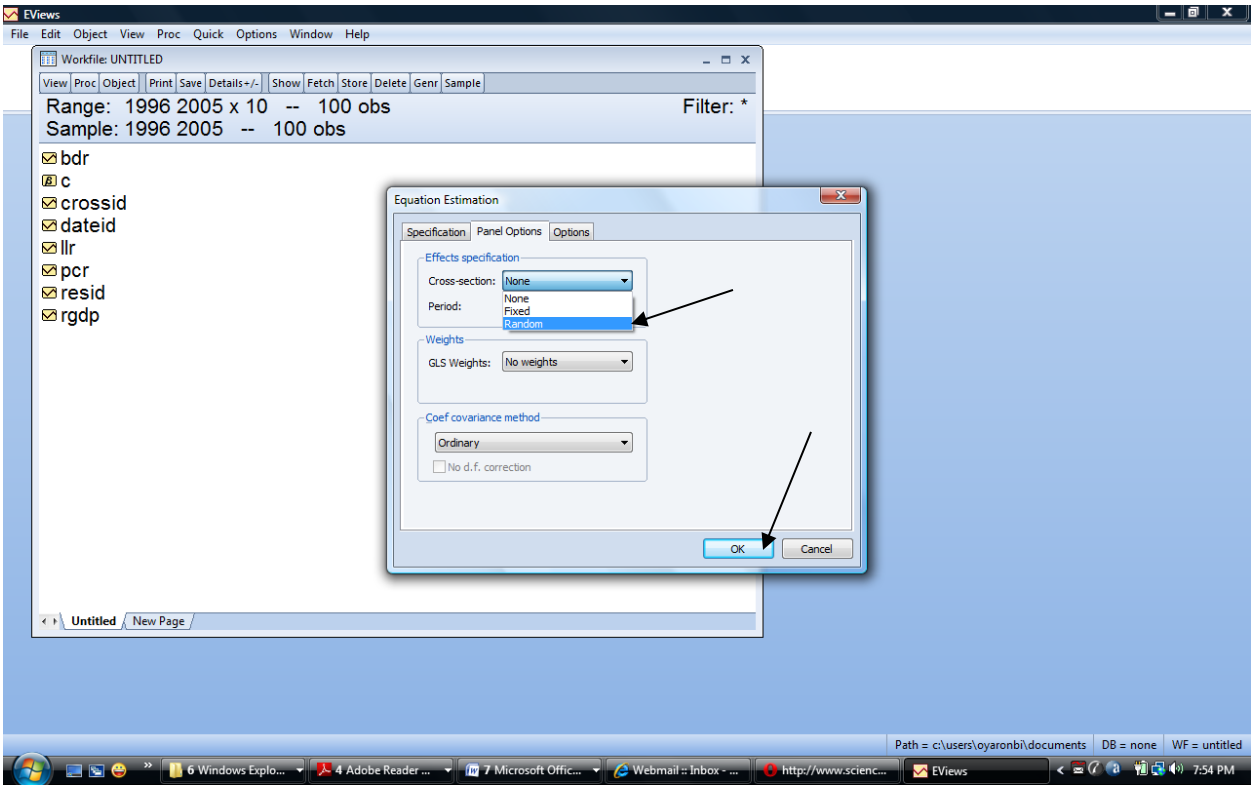
(a) Equation editor



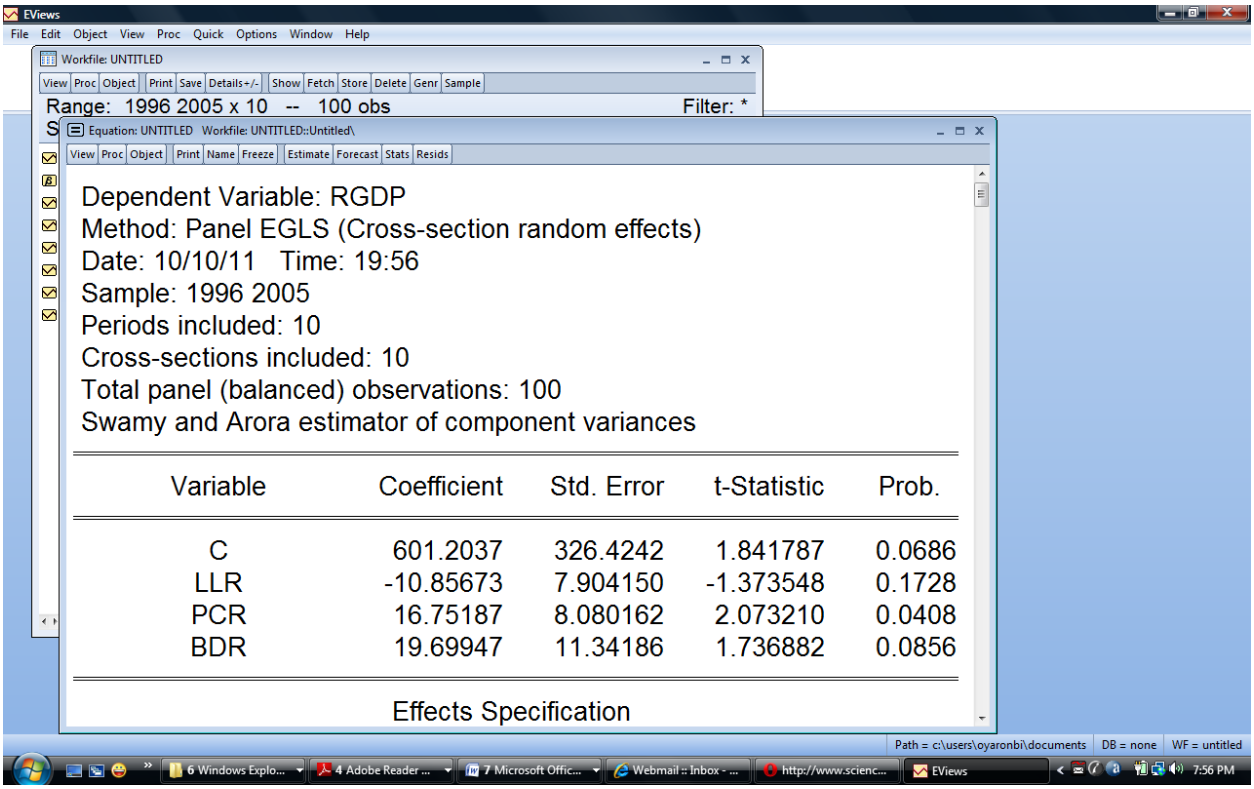
(b) Model Specification



(c) **Specifying the random effects**

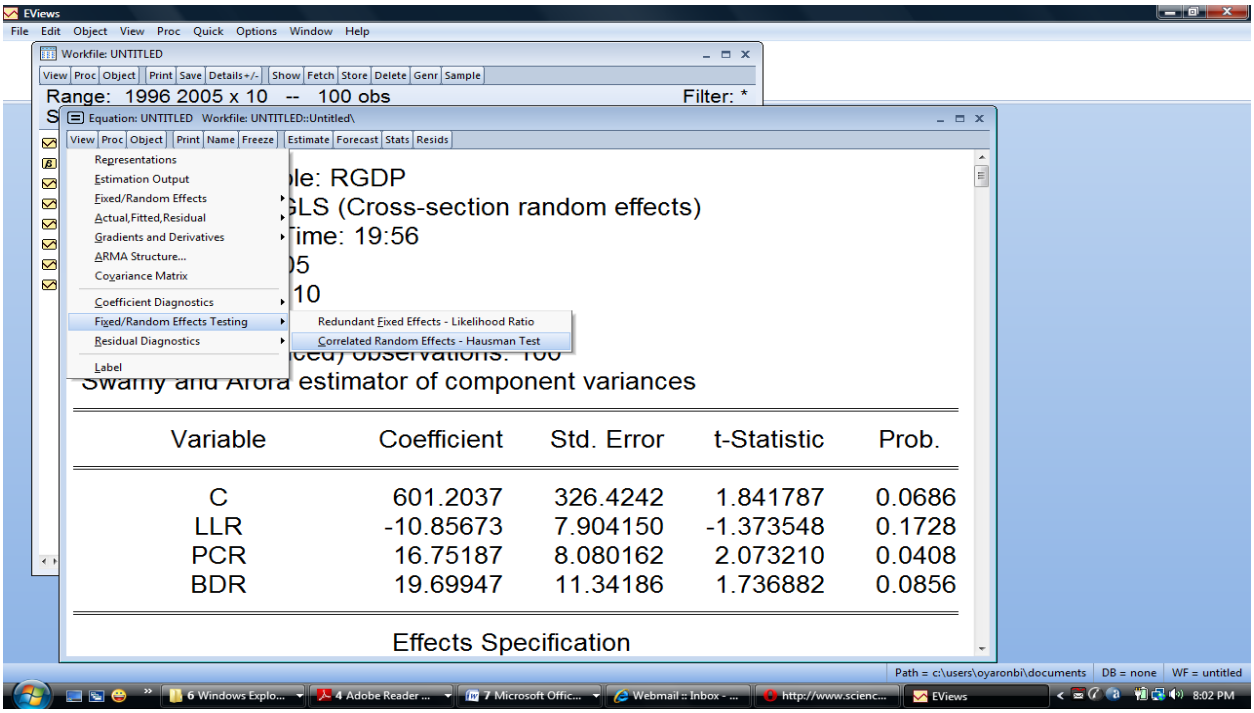


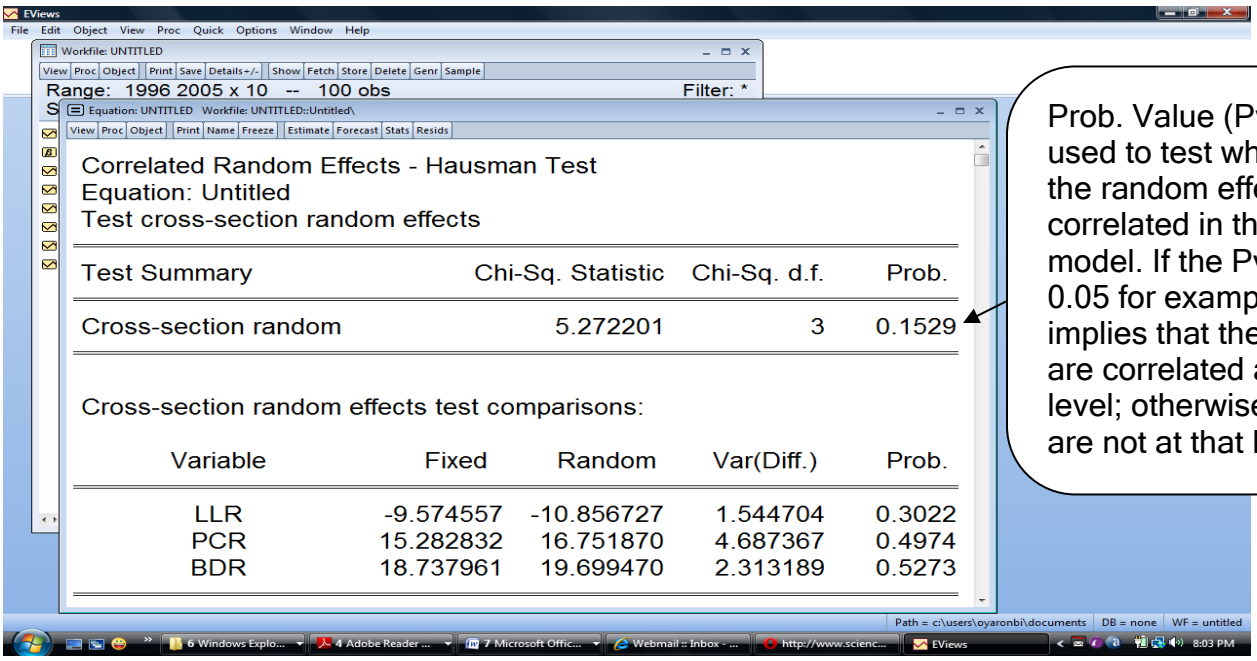
(d) Random Effects Regression



The interpretation of regression results here is not different from the style used under the pooled regression. However, we need to further test whether the inclusion of the random effects is significant and also whether these effects are correlated or not. As earlier explained, the Hausman test is used in this regard.

(e) Random Effects Testing





Prob. Value (Pv) is used to test whether the random effects are correlated in the model. If the Pv is < 0.05 for example, it implies that the effects are correlated at 5% level; otherwise, they are not at that level.

It should be noted that one of the assumptions underlying the use of random effects model is that specific effects are independently and identically distributed. To confirm whether the specific effects estimated are actually random effects and are uncorrelated,

the Hausman test is often used. As earlier emphasized, the rejection of the null hypothesis (when the statistic is statistically significant) implies an adoption of the fixed effects model and non-rejection is considered as an adoption of the random effects model. The rejection of the null hypothesis also implies correlated specific effects are better captured with fixed effects model. Based on the pv of the test summary as shown above, the effects are not correlated and therefore the random effects regression may give a better fit than the fixed effects model.

Self- Assessment 1

State the null and alternative hypothesis for Hausman test

1.6 Summary

In this unit, you learned the steps in estimating a Panel Equation in Eviews. In which Pooled OLS, Random effect and Fixed effect were discussed.

In estimating panel data model in Eviews, certain steps must be followed. You should make certain that your workfile is structured as a panel workfile. EViews will detect the presence of your panel structure and in place of the standard equation dialog will open the panel Equation Estimation dialog.

Tutor Marked Assignment

Estimate Random effect using the data provided in the example.

1.8 Possible Answers to Self-Assessment Exercise(s) Within the Content

Answer to Self- Assessment 1

The null and alternative hypotheses for the Hausman test are as follows:

1. **Null Hypothesis (H0)**: The random effects estimator and the fixed effects estimator are statistically equivalent. This essentially implies that the unique errors are not correlated with the regressors, meaning that the more efficient random effects estimator can be preferred over the fixed effects estimator.
2. **Alternative Hypothesis (HA)**: The random effects estimator and the fixed effects estimator are not statistically equivalent, suggesting that the unique errors are correlated with the regressors. In this case, the fixed effects model should be used as it provides consistent estimates while the random effects estimator does not.

1.7 References/Further Reading

Adewara, S. O. & Kilishi, A. A. (2015). Analysis of survey data using stata. A workshop lecture presented on 27th – 30th April, 2015 in University of Illorin, Nigeria.

Ezie, O., & Ezie, K.P. (2021). Applied Econometrics: Theory and Empirical Illustrations. Kabod Limited Publisher, Kaduna.

Cameron, A. C. & Trivedi, P. K. (2009). Microeconometrics using stata. Texas, USA: Stata Press.

Gujarati, D. N. & Porter, D. C. (2009). Basic econometrics (5th ed.). New York, NY: McGraw-Hill/Irwin.

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). OH, USA: Cengage.