INTRODUCTION TO ECONOMETRICS II

ECO 356

FACULTY OF SOCIAL SCIENCES

COURSE GUIDE

Course Developers:

Dr. Adesina-Uthman G. A.

Faculty of Social Sciences
Department of Economics
National Open University of Nigeria.

and

Okojie, Daniel Esene
E-mail: danelektrik@gmail.com
School of Postgraduate Studies (SPGS)
University of Lagos (UNILAG)
Akoka, Yaba, Lagos State
Nigeria.

Content Editor:

Prof. Ismael Ogboru
Department of Economics,
University of Jos,
Jos, Plateau State.

COURSE CONTENT:

Main Introduction
Course Outline
Aims
Course Objectives
Working through the Course
Course Materials
Study Units
Textbooks and Reference Resources
Assignment Folder
Presentation Plan
Assessment
Tutor-Marked Assignments (TMAs)
Concluding Examination and Grading
Marking Scheme
Overview
Making the Most of this Course
Tutors and Tutorials
Summary
Main Introduction

ECO 306 is a logical extension of the first-semester course on regression analysis. As such, it introduces the concept of the simultaneous equation and their estimation. Essentially, this course examines the possible solutions to problems arising from the breakdown of the ordinary least squares assumptions and sampling theories. To this end, it covers topics like multicollinearity, heteroscedasticity, autocorrelation and Econometrics Modeling: Specification and Diagnostic Testing. It also examines the use of regression analyses, correlation, variance and dummy variables. For this reason, experiential case studies that apply the techniques to real-life data are stressed and discussed throughout the course, and students are required to get acquainted with their several models and theories that deal with the measurement of economic relationships.

The course would be a very useful material to you in your academic pursuit and could help to broaden your understanding further in this case. Once this understanding and application are established, you are then able to have a broadened knowledge of econometrics while distinguishing it from mathematical economics.

This course is therefore developed in a manner to guide you further on what econometrics entails, what course materials in line with a course learning structure you will be using. The learning structure suggests some general guidelines for a time frame required of you on each unit to achieve the course aims and objectives effectively. Further work in this course would expose you to introductory levels of topics like; vector autoregressions, unit roots, cointegration, time-series analysis and errors in variables.

Course Outline

ECO 306 is made up of five modules with seventeen units spread across twelve lectures weeks. The modules cover areas such as the concept of the simultaneous equation and their estimation, ordinary least squares assumptions, multicollinearity, heteroscedasticity, autocorrelation and econometrics modeling: Specification and Diagnostic Testing, use of dummy variables and time-lags as independent variables.

Aims

The aim of this course is to give you thorough understanding and an appreciative importance of econometrics being concerned with more than measurement in economics. But more importantly, how econometrics as a method of causal inference is applied to economics. That is, this method of causal inference is a statistical inference combined with the logic of causal order; which is to infer or learn something about the real world by analysing a sample of data.

Specifically, the aims of the course are to:

- Equip you with the application of statistical methods to the measurement and critical assessment of assumed economic relationships using data.
- Provide an improved introductory understanding of how the economy works, at either the microeconomic or macroeconomic level.
Course Objectives

To achieve the aims mentioned above also to the overall stated course objectives. Each unit, in the beginning, has its specific objectives. You should read them before you start working through the unit. You may want to refer to them during your study of the unit to check on your progress and should always take a look back at the objectives after completion. In this way, you can be certain you have done what was necessary to you by the unit. The course objectives are set below for you to achieve the aims of the course. On successful conclusion of the course, you should be able to:

- Know the basic principles of econometric analysis
- Express relationships between economic variables using mathematical concepts and theories
- Understand both the fundamental techniques and wide array of applications involving linear regression estimation
- Analyse the strengths and weaknesses of the basic regression model.
- Outline the assumptions of the normal linear regression model and discuss the significance of these assumptions
- Explain the method of ordinary least squares
- Test hypotheses of model parameters and joint hypotheses concerning more than one variable
- Discuss the consequences of multicollinearity, the procedures for identifying multicollinearity, and the techniques for dealing with it
- Explain what is meant by heteroscedasticity, and the consequences for ordinary least square (OLS) estimators and prediction based on those estimators
- Assess the methods used to identify heteroscedasticity, including data plots and more formal tests, and the various techniques to deal with heteroscedasticity, including model transformations and estimation by weighted least squares
- Explain autocorrelation, and discuss the consequences of autocorrelated disturbances for the properties of OLS estimator and prediction based on those estimators
- Outline and discuss the methods used to identify autocorrelated disturbances, and what can be done about it, including estimation by generalised least squares
- Discuss the consequences of disturbance terms not being normally distributed, tests for non-normal disturbances, and methods to deal with non-normal disturbances, including the use of dummy variables
- Discuss the consequences of specifying equations incorrectly
- Discuss the tests used to identify correct model specification and statistical criteria for choosing between models

Working through the Course

This course highlights on critical thinking and the application of both logical and quantitative skills. It also stresses on the application of econometric methods to economic theory and practical problems. Therefore, to complete this course, you are required to read the study units, referenced books and other materials on the course. Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course, there is a final examination. This
course should take about twelve weeks to complete, and some components of the course are outlined under the course material subsection.

Course Material

The major component of the course, what you have to do and how you should allocate your time to each unit to complete the course successfully and on time are as follow:

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

Study Unit

In this course, there are five modules that subdivided into 17 units which should be studied thoroughly.

Module 1: Sampling Theory, Variance, and Correlation
Unit 1: Random Variables and Sampling Theory
Unit 2: Covariance and Variance
Unit 3: Correlation

Module 2: Regression Models, Hypotheses Testing, and Dummy Variables
Unit 4: Simple Regression Analyses
Unit 5: Properties of the Regression Coefficients and Hypothesis Testing
Unit 6: Multiple Regression Analysis and Multicollinearity
Unit 7: Transformations of Variables
Unit 8: Dummy Variables
Unit 9: Specification of regression variables: A preliminary skirmish

Module 3: Heteroscedasticity/Heteroskedasticity
Unit 10: Heteroscedasticity and Its Implications
Unit 11: Solution to Heteroscedasticity Problem
Unit 12: Other Tests

Module 4: Autocorrelation, Error, and Econometric Modelling
Unit 13: Stochastic Regression and measurement errors
Unit 14: Autocorrelation
Unit 15: Econometric Modelling and Models Using Time Series Data

Module 5: Simultaneous Equation, Binary Choice, and Maximum Likelihood Estimation
Unit 16: Simultaneous Equations
Unit 17: Binary Choice and Maximum Likelihood Estimation.

The general aim of module 1 (units 1-3) is to provide you with a thorough understanding of the basic statistical tools needed for regression analyses in the subsequent modules. The Random variables and sampling theory, covariance,
variance, and correlation are demystified for proper understanding. By the end of this module, you would have been able to understand the basics of regression analysis.

Module 2 (units 4-9) explains single-equation regression models. It shows how a hypothetical linear relationship between two variables can be quantified using appropriate data. The principle of least squares regression analysis explained, and expressions for the coefficients are derived. Multicollinearity and multiple regression analysis looked at in units 6. Transformations of Variables discussed in unit 7 while dummy variables as well preliminary skirmish of the specification of regression variables are the topics in units 8 and 9. An exploration of what happens when there is a violation of one of the classical assumptions; equal variances (homoscedastic) is carried out in module 3. It demonstrates how properties of estimators of the regression coefficients depend on the properties of the disturbance term in the regression model. Also, in this module, we shall look at some of the problems that arise when violations of the Gauss–Markov conditions; the assumptions relating to the disturbance term, are not satisfied. Basic understanding of heteroscedasticity (unequal-variances) will gain thorough explanation.

The module 4 (unit 13-15) covers an understanding of the basics of econometric modelling. It goes further to give some details on stochastic regression and measurement errors, autocorrelation, econometric modelling and models using time series data. More detail description of an introduction to Consequences of Measurement Errors. Intercorrelation among the Explanatory Variables and Measurement Errors in the Dependent Variable are brought to the students’ knowledge here. Also, possible causes of Autocorrelation and Detection of First-Order Autocorrelation using the Durbin–Watson Test are presented in units 14 and 15 of the same module 4. While module 5 with units 16 and 17 provide you with a thorough understanding of the basic rudiments of Simultaneous Equation, Binary Choice, and Maximum Likelihood Estimation.

Respectively, study unit will take at least two hours which include an introduction, objective, main content, examples, In-Text Questions (ITQ) and their solutions, self-assessment exercise, conclusion, summary, and reference. Additional areas border on the Tutor-Marked Assessment (TMA) questions. Some of the ITQ and self-assessment exercise will require you free-associating and solve with some of your colleagues. You are advised to do so to grasp and get familiar with how significant econometrics is in being concerned with measurement and also as a method of causal inference application to economics.

There are also econometrics materials, textbooks under the reference and other (on-line and off-line) resources for further studies. These are intended to give you extra facts whenever you allow yourself of such prospect. You are required to study the materials; practise the ITQ, self-assessment exercise and TMA questions for better and thorough understanding of the course. In doing these, the identified learning objectives of the course would have been attained.

For further reading in this course, the following reference texts and materials are suggested:
Textbooks and References

Robert D. Coleman, 2006, The Aims and Methodology of Econometrics Harvard Business School, USA


Assignment Folder

The assignments given in this course are for you to attempt all of them by following the timetable recommended regarding when to do them and submission of same for grading by your lecturer. The marks you obtain for these assignments will count toward the final mark you obtain for this course. Further information on assignments will be found in the Assignment File itself and later in this Course Guide in the section on Assessment.

There are five assignments in this course:
Assignment 1 - All TMAs’ question in Units 1 – 3 (Module 1)
Assignment 2 - All TMAs’ question in Units 4 – 8 (Module 2)
Assignment 3 - All TMAs’ question in Units 9 – 11 (Module 3)
Assignment 4 - All TMAs’ question in Units 12 – 14 (Module 4)
Assignment 5 - All TMAs’ question in Units 15 – 17 (Module 5)

Presentation Plan

The presentation plan included in your course materials gives you the important dates in the year for the completion of tutor-marking assignments and tutorial attendance. Remember, you are required to submit all your assignments by the due date. You should guide against dropping behind in your assignments submission.

Assessment

Two types of assessments are available in this course; Tutor-Marked Assignment and a written examination at the end of the course.

For the assignments, you are expected to apply lessons learnt during the course. The assignments must be submitted to your lecturer for proper valuation in agreement with the deadlines stated in the Presentation Schedule and the Assignments File. The assignment works you are to submit to your lecturer for evaluation would count for 30% of your total course grade.
At the end of the course, you will need to sit for a final written examination of three hours duration. This examination will also count for 70% of your total course grade.

**Tutor-Marked Assignments (TMAs)**

There are six tutor-marked assignments in this course. You will submit all the assignments. You are encouraged to attempt all the questions carefully. The TMAs constitute 30% of the total marks.

Assignment questions for the units contained in this course are in the Assignment File. You will be able to complete your assignments from the information and materials contained in your textbooks and study units. However, it is desirable that you demonstrate that you have read and solved a lot of problems relating to each topic in a module. You could use other reference materials to have a broader viewpoint of each subject in this course.

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

**Concluding Examination and Grading**

Final examination on the course will be for three hours duration and has 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. There is an evaluation of all areas of the course.

You are advised to use the time between finishing the last unit and sitting for the examination to revise the entire course materials. You might find it useful to review your In-Text Questions (ITQ) and self-assessment exercises, tutor-marked assignments and comments on them before the examination. The final examination covers the entire course outline.

**Marking Scheme**

Table 1 presents the total marks (100%) allocation.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment (Best three assignment out of the five marked)</td>
<td>30%</td>
</tr>
<tr>
<td>Final Examination</td>
<td>70%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Overview

Table 2 shows the units, number of weeks and assignments to be taken by you to complete the course successfully: Introduction to Econometrics (ECO 306).
Table 2: Assignment Schedule

* Comprise of a single module (Module 3) not broken into the unit.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Unit Title</th>
<th>Week’s Activity</th>
<th>Assessment (end of unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Course Guide</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sampling Theory, Variance, and Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Random variables and sampling theory</td>
<td>Week 1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Covariance and Variance</td>
<td>Week 2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Correlation</td>
<td>Week 3</td>
<td>Assignment 1</td>
</tr>
<tr>
<td></td>
<td>Regression Models, Hypotheses Testing, and Dummy Variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Simple Regression Analyses</td>
<td>Week 4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Properties of the regression coefficients and hypothesis testing</td>
<td>Week 5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Multiple regression analysis and Multicollinearity</td>
<td>Week 6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Transformations of Variables</td>
<td>Week 7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Dummy Variables</td>
<td>Week 8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Specification of regression variables: A preliminary skirmish</td>
<td>Week 9</td>
<td>Assignment 2</td>
</tr>
<tr>
<td></td>
<td>Heteroscedasticity/Heteroskedasticity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*10</td>
<td>Heteroscedasticity and its Effects</td>
<td>Week 10</td>
<td></td>
</tr>
<tr>
<td>*11</td>
<td>Solution to Heteroscedasticity Problem</td>
<td>Week 11</td>
<td></td>
</tr>
<tr>
<td>*12</td>
<td>Other Tests</td>
<td>Week 12</td>
<td>Assignment 3</td>
</tr>
<tr>
<td></td>
<td>Autocorrelation, Error and Econometric Modelling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Stochastic Regression and measurement errors</td>
<td>Week 13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Autocorrelation</td>
<td>Week 14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Econometric Modelling and Models Using Time Series Data</td>
<td>Week 15</td>
<td>Assignment 4</td>
</tr>
<tr>
<td></td>
<td>Simultaneous Equation, Binary Choice, and Maximum Likelihood Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Simultaneous Equation</td>
<td>Week 16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Binary Choice and Maximum Likelihood Estimation</td>
<td>Week 17</td>
<td>Assignment 5</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>17 Weeks</td>
<td>Examination</td>
</tr>
</tbody>
</table>

Making the Most of this Course

An advantage of the distance learning is that the study units replace the university lecturer. You can read and work through specially designed study materials at your tempo and at a time and place that goes well with you. Consider doing it yourself insolving and providing solutions to econometric problems in the lecture instead of listening and copying solution being provided by a lecturer. In the same way, that a lecturer might set you some practice exercises and ITQ to do, the study units tell you when to solve problems and read your books or other material, and when to embark on a discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provide exercises for you to do at appropriate points.
Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the unit, you must go back and check whether you have achieved the objectives. If you make a habit of doing this, you will significantly improve your chances of passing the course and getting the best grade.

The main body of the unit guides you through the required understanding from other sources. This will usually be either from your textbooks or a reading section. Some units require you to undertake a practical overview of real life econometric events. You will find when you need to embark on discussion and guided through the tasks you must do. The purpose of the practical overview of real life econometric events is in twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience and skills to evaluate economic arguments, and understand the roles of econometric in guiding current economic problems, measurements, analysis, solutions and debates outside your studies. In any event, most of the critical thinking skills you will develop during studying are applicable in normal working practice, so it is important that you encounter them during your studies.

Self-assessments are available throughout the units, and answers are at the ends of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each self-assessment exercises as you come to it in the study unit. Also, ensure to master some major econometricstheorems and models while studying the material.

The following is a practical strategy for working through the course. If students encounter any difficulties they may consult the lecturer concerned. Remember that part of the lecturers’ responsibility is to help the students when necessary. When in need of help, do not hesitate to contact the lecturer though the available channels, for the required assist.

1. Read this Course Guide thoroughly.
2. Organize a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your dates for working breach unit.
3. Once you have created your study schedule, do everything you can to stick to it. The major reason students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.
5. Assemble the study materials. Information about what you need for a unit is available in the 'Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your textbooks on your desk at the same time.

6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit, you will be instructed to read sections from your textbooks or other articles. Use the unit to guide your reading.

7. Up-to-date course information will be delivered continuously to you at the study centre.

8. Work before the relevant due date (about four weeks before due dates) get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.

9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your lecturer.

10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.

11. When you have submitted an assignment to your tutor for marking, do not wait for its return before starting on the next units. Keep to your schedule. When returning the assignment, pay particular attention to your lecturer's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your lecturer as soon as possible if you have any questions or problems.

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

**Tutors and Tutorials**

There are some hours of tutorials (2-hours sessions) provided in support of this course. You should get notifications of dates, times, and location for these tutorials. Together with the name and phone number of your lecturer, as soon as the tutorial group allocated are made.

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

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

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

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

Summary

The course, Introduction to Econometrics II (ECO 306) presents you with general background and applications of the concept of Random Variables, Sampling Theory and how to be able to identify functions and problems associated with estimation. This course also examines ordinary least squares assumptions and sampling theories. Topics like, multicollinearity, heteroscedasticity, autocorrelation and Econometrics modeling had illustrative examples used for further explanations. For this reason, use of regression analyses, correlation, variance and dummy variables with experiential case studies that apply the techniques to real-life data are stressed and discussed throughout the course.

This course is therefore developed in a manner to guide you further on what econometrics entails, what course materials in line with a course learning structure you will be using. The learning structure suggested some general guidelines for a time frame required of you on each unit to achieve the course aims and objectives.

Conclusively, you would have developed critical thinking skills with the material necessary for an efficient introductory understanding of econometrics. Nevertheless, to achieve a lot more from the course, please try to solve econometrics problems independently, do presentation and interpretation of findings in any assignment given both in your academic programme and other spheres of life. Further work in this course would expose you to introductory levels of topics like; vector autoregressions, unit roots, cointegration, and time-series analysis.

We wish you the very best in your schoolwork.
Course Code: ECO 356

Course Title: Introduction to Econometrics II

Course Developer/Writer:

Dr. Adesina-Uthman G. A.
Faculty of Social Sciences
Department of Economics
National Open University of Nigeria.

and

OKOJIE, Daniel Esene
School of Post Graduate Studies (SPGS)
University of Lagos, Akoka-Yaba
Lagos.

Content Editor:

Prof. Ismael Ogboru
Department of Economics,
University of Jos, Jos,
Plateau State.

December, 2017
CONTENTS

PAGES

Module 1: Sampling Theory, Variance and Correlation

Unit 1: Random variables and sampling theory.................................................4

Unit 2: Covariance and Variance..................................................................13

Unit 3: Correlation Coefficient.....................................................................20

Module 2: Simple Equation Regression Models

Unit 1: Simple Regression Analyses.........................................................25

Unit 2: Properties of the regression coefficients and hypothesis testing........36

Unit 3: Multiple regression analysis and Multicollinearity..........................54

Unit 4: Transformations of Variables.......................................................66

Unit 5: Dummy Variables........................................................................69

Unit 6: Specification of regression variables: A preliminary skirmish...........74

Module 3: Heteroscedasticity/Heteroskedasticity

- Heteroscedasticity and Its Implications..................................................76
- Solution to Heteroscedasticity Problem....................................................84
- Other Tests/Consequences of Heteroscedasticity................................84
Module 4: Autocorrelation, Error and Econometric Modelling
Unit 1: Stochastic Regression and measurement errors..........................86

Unit 2: Autocorrelation........................................................................93

Unit 3: Econometric Modelling and Models Using Time Series Data........99

Module 5: Simultaneous Equation, Binary Choice, and Maximum Likelihood Estimation
Unit 1: Simultaneous Equations..........................................................103

Unit 2: Binary Choice and Maximum Likelihood Estimation...............107

MODULE 1 SAMPLING THEORY, VARIANCE, AND CORRELATION

The general aim of this module is to provide the student with a thorough understanding of the basic statistical tools that will be needed for regression analyses in the subsequent module. As well as random variables and sampling theory, Covariance, variance, and correlation are to be demystified for proper understanding. By the end of this module, students would have been able to understand the basic parts of regression analysis.

The units to be studied in this module are:

Unit 1: Random variables and sampling theory
Unit 2: Covariance and Variance
Unit 3: Correlation

UNIT 1: RANDOM VARIABLES AND SAMPLING THEORY
CONTENTS
1.1.1.0 INTRODUCTION

Random variable or stochastic variable is a variable whose possible values are numerical products of a chance occurrence. As a function, a random variable is required to be quantifiable, which rules out certain uncontrolled circumstances where the quantity which the random variable returns is considerably sensitive to small changes in the outcome. It is common that these outcomes depend on some physical variables that are not well understood. For example, when you toss a coin, which outcome will be observed is not certain. The domain of a random variable is the set of possible outcomes. In the case of the coin, there are only two possible outcomes, namely heads or tails. The domain of the random variable leads us into the concept of sampling theory which is concerned with the theory involved in the selection of a subset of individuals from within a statistical population estimates the characteristics of the whole population.
1.1.2.0 OBJECTIVE

The main objective of this unit is to provide a broad understanding of the topic, Random Variables, and Sampling Theory, which is preparatory to the more widely used simple and multiple regression analyses.

1.1.3.0 MAIN CONTENTS

1.1.3.1 Random Variables and Sampling Theory

A variable $X$ is said to be a random variable if for every real number $a$ there exist a probability $P(X \leq a)$ that $X$ takes on a value less than or equal to $a$. That is, a Random variable is a variable whose value cannot be predicted exactly. It can assume any value. Random variables could be discrete or continuous. A discrete random variable is one that has a specific set of possible values or a finite set of values. An example is a total score when two dice are thrown. A continuous variable, e.g. the temperature in a particular room, is a variable that can assume any value in the certain range. It can take any form of the continuing range of values.

The set of all possible values of a random variable is known as a population where the sample or a random variable can be drawn for inferential analysis.

1.1.3.2 Expected values of discrete random variable

The expected value of a discrete random variable is the weighted average of all its possible values, taking the probability of each outcome as its weight. It can be calculated by multiplying each possible value of the random variable by its probability and adding. In mathematical terms, if $X$ denotes the random variable, its expected value is denoted by $E(X)$.

Let us suppose that $X$ can take $n$ particular values of $x_1, x_2, ..., x_n$ and that the probability of $x_i$ is $p_i$. Then,

$$E(X) = x_1p_1 + ... + x_np_n = \sum_{i=1}^{n} x_ip_i \quad \text{[1.01]}$$

Table 1.0 shows an example of expected value of variable $X$ with two dice.

*Table 1.0: Expected value of variable $X$ with two dice*
In the case of the two dice, the values \( x_1 \) ... \( x_n \) were the numbers 2 ... 12: \( x_1 = 2 \), \( x_2 = 3 \) ... \( x_{11} = 12 \), and \( p_1 = \frac{1}{36} \), \( p_2 = \frac{2}{36} \) ... \( p_{11} = \frac{1}{36} \). As shown in table 1.0, the expected value is 7. Also, the expected value of a random variable is described as population mean. In the case of the random variable \( X \), the population mean is given as \( \mu_X \).

### 1.1.3.3 Expected value rules

There are three main rules of expected values that are equally valid for both discrete and continuous random variables. These are:

**Rule 1:** The expected value of the sum of several variables is equal to the sum of their respective expected values. For example, if you have three random variables \( X \), \( Y \), and \( Z \),

\[
E(X + Y + Z) = E(X) + E(Y) + E(Z)
\]

\[\ldots[1.02]\]
**Rule 2:** If you multiply a random variable by a constant, you multiply its expected value by the same constant. If $X$ is a random variable and $b$ is a constant,

$$E(bX) = bE(X) \quad \text{...}[1.03]$$

**Rule 3:** The expected value of a constant is that constant. For example, if $b$ is a constant.

$$E(b) = b \quad \text{...}[1.04]$$

Putting the three rules together; suppose we wish to calculate $E(Y)$, where we have

$$Y = b_1 + b_2X \quad \text{...}[1.05]$$

and $b_1$ and $b_2$ are constants.

Then,

$$E(Y) = E(b_1 + b_2X) \quad \text{...}[1.06]$$

$$= E(b_1) + E(b_2X) \text{ (using rule 1)} \quad \text{...}[1.07]$$

$$= b_1 + b_2E(X) \text{ (using rule 2 & 3)} \text{...}[1.08]$$

### 1.1.3.4 Sampling theory

The goals of a sample survey and an experiment are very different. The role of randomisation also differs. In both cases, without randomisation, there can be no inference. Without randomisation, the analyst can only describe the observations and cannot generalise the results. In the sample survey, randomisation is used to reduce bias and to allow the results of the sample to be generalised to the population from which the sample was drawn. In an experiment, randomisation is used to balance the effects of confounding variables. The objective of a sample survey is often to estimate a population mean and variance.

### 1.1.3.4.1 Some terminology
i. **Element:** An element is an object on which a measurement is made, which could be a voter in an area, a product as it comes off the assembly line or a plant in a field that has either flowered or not.

ii. **Population:** A population is a collection of elements about which we wish to make an inference. The population must be clearly defined before the sample is taken.

iii. **Sampling units:** These are some overlapping collections of elements from the population that covers the entire population. The sampling units partition the population of interest. The sampling units could be households or individual voters.

iv. **Frame:** Is a list of sampling units.

v. **Sample:** This is a collection of sampling units drawn from a frame or frames. Data are obtained from the sample and are used to describe characteristics of the population. Two advantages of sampling are that the cost is lower and data collection is faster than measuring the entire population.

vi. **Census:** The enumeration of the total element of the population.

**Example 1:** Suppose we are interested in what voters in a particular area think about the drilling of oil in the national wildlife preserves. The elements are the registered voters in the area. The population is the collection of registered voters. The sampling units will likely be households in which there may be several registered voters. The frame is a list of households in the area.

**1.1.3.4.2 Reasons for sampling**

Information could be obtained by taking a complete enumeration of the whole population or aggregate. This is usually difficult as information on every element is rarely available. Therefore, it is better to employ sampling method to obtained information than complete enumeration for the following reasons:

i. **Reduce cost:** if data are secured from only a small fraction of the aggregate, expenditures are smaller than if a complete census is attempted. With large populations result accurate enough to be useful can be obtained from samples that represent only a small fraction of the population.
ii. **Greater speed:** for the same reason the data can be collected and summarized more quickly with a sample than with a complete count. This is a vital consideration when the information is urgently needed.

iii. **Greater Scope:** a complete census is impracticable; the choice lies between obtaining the information by sampling or not at all. Thus surveys that rely on sampling have more scope and flexibility regarding the kind of information that can be obtained.

iv. **Greater Accuracy:** here, personnel of higher quality can be employed and given intense training. This would allow for much more careful supervision of the field work. Processing and analysing of the results become feasible because the volume of work is now reduced. The sample would most likely produce a more accurate result than the complete enumeration.

1.1.3.4.3 Types of sampling technique

i. **Probability sampling technique:** Simple random sampling, systematic random sampling, stratified random sampling, cluster sampling, etc.

ii. **Non-probability sampling technique:** Snowball sampling, quota sampling technique, accidental or convenient sampling technique, etc.

Sample designs that utilise planned randomness are called **probability samples** while non-probability doesn’t apply randomness as it is based on the subjective dictate of the researcher since all elements are not given equal chance of being selected. The most fundamental probability sample is the simple random sample. In a simple random sample, a sample of size \( n \) is selected in such a way that each sample of size \( n \) has the same chance of being selected. In practice, other more sophisticated probability sampling methods are commonly used, but we would focus here on simple random sampling technique.

1.1.3.4.4 Simple Random Sampling technique

Suppose the observations \( y_1, y_2, \ldots, y_n \) are to be sampled from a population with mean, standard deviation, and size \( N \) in such a way that every possible sample of size \( n \) has an equal chance of being selected. Then the sample \( y_1, y_2, \ldots, y_n \) was selected in a simple random sample. If the sample mean is denoted by \( \bar{y} \) then we have;
\[ E(\bar{y}) = \mu \quad \text{...}[1.09] \]
\[ \text{V}(\bar{y}) = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right) \quad \text{...}[1.10] \]

The term \( \left( \frac{N-n}{N-1} \right) \) in the above expression is known as the \textbf{finite population correction factor}. For the sample variance \( s^2 \), it can be shown that

\[ E(s^2) = \left( \frac{N}{N-1} \right) \cdot \sigma^2 \quad \text{...}[1.11] \]

When using \( s^2 \) as an estimate of \( \sigma^2 \), we must adjust with

\[ \sigma^2 = \left( \frac{N}{N-1} \right) \cdot E(s^2) \quad \text{...}[1.12] \]

Consequently, an unbiased estimator of the variance of the sample mean is given by

\[ \hat{\text{V}}(\bar{y}) = \frac{s^2}{n} \left( \frac{N-n}{N} \right) \quad \text{...}[1.13] \]

As a rule of thumb, the correction factor \( \left( \frac{N-n}{N-1} \right) \) can be ignored if it is greater than 0.9, or if the sample is less than 10% of the population.

**Example 2:** Consider the finite population with \( N = 4 \) elements \( \{0, 2, 4, 6\} \). For this population \( \mu = 3 \) and \( \sigma^2 = 5 \). Simple random samples without replacement of size \( n = 2 \) are selected from the population. All possible samples along with their summary statistics are listed in table 1.1.1.

\begin{table}[h]
\centering
\caption{Simple Random Sampling}
\begin{tabular}{|c|c|c|c|}
\hline
Samples & Probability & mean & Variance \\
\hline
(0,2) & \frac{1}{6} & 1 & 2 \\
\hline
(0,4) & \frac{1}{6} & 2 & 8 \\
\hline
(0,6) & \frac{1}{6} & 3 & 18 \\
\hline
(2,4) & \frac{1}{6} & 3 & 2 \\
\hline
\end{tabular}
\end{table}
We see in this example that;

\[ V(\bar{y}) = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right) = \frac{5}{2} \left( \frac{4-2}{4-1} \right) = \frac{5}{2} \left( \frac{2}{3} \right) = \frac{5}{3} \]

Similarly,

\[ E(s^2) = \left( \frac{N}{N-1} \right) \sigma^2 \]

Could also be obtained from table 1.1.1

1.1.3.5 Estimation of Population Mean

If we are interested in estimating a population mean from a simple random sample, we have;

\[ \hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \]

If we are interested in estimating population variance from a simple random sample, we have;

\[ \hat{V}(\bar{y}) = \frac{s^2}{n} \left( \frac{N-n}{N} \right) \]

Where,

\[ s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \]

When the margin of error is two standard errors, we have;

\[ 2\sqrt{\hat{V}(\bar{y})} = 2 \sqrt{\frac{s^2}{n} \left( \frac{N-n}{N} \right)} \]

1.1.4.0 SUMMARY
In this unit, the students are presented with the essentials and applications of the concept of random variables and sampling theory and their estimations. Also, by now the students should be able to identify functions and problems associated with the estimation.

1.1.5 CONCLUSION

In this unit, the concepts of random variables and sampling theory have been introduced and discussed and the associated estimation explained. The students are made to understand that random variables are of two quantitative (discrete and continuous random) variable types, whose values may be determined through the outcome of a random trials. Discrete and continuous random variables were both explained with examples. Further definition of random variable as it relates to a population is explained. Furthermore, inference is made on how the student may apply the random variable in other statistical analysis like mean, variance, standard deviation, etc.

Sampling theory is introduced in this unit in form of comparison to random variable. Some terminologies and reasons associated with sampling are also discussed. Probability and non-

probability sampling techniques are presented as two sample designs methods that utilise planned randomness. Although, in practice, other more sophisticated probability sampling and estimation methods are commonly used, the unit focused on simple random sampling and population mean techniques as starting points.

1.1.6 TUTOR-MARKED ASSIGNMENT

1.) A random variable $X$ is defined to be the difference between the higher value and the lower value when two dice are thrown. If they have the same value, $X$ is defined to be 0. Find the probability distribution for $X$.

2.) A random variable $X$ is defined to be the larger of the two values when two dice are thrown, or the value if the values are the same. Find the probability distribution for $X$. 
UNIT 2: CO-VARIANCE AND VARIANCE

CONTENTS
1.2.1.0 Introduction
1.2.2.0 Objectives
1.2.3.0 Main Content
   1.2.3.1 CoVariance and Variance
   1.2.3.2 Some Basic Covariance rule
   1.2.3.3 Population CoVariance
   1.2.3.4 Sample Variance
   1.2.3.5 Variance Rule
1.2.4.0 Conclusion
1.2.5.0 Summary
1.2.6.0 Tutor-Marked Assignment
1.2.7.0 References/Further Reading

1.2.1.0 INTRODUCTION

The previous unit in this module introduced and discussed random variable and associated sampling theories. In order to further equip the students with the adequately understanding of more basic tools needed for regression analysis in the next module
and in general statistical analyses, this unit discusses covariance and variance. The unit explains that, variance and covariance are two measures used in statistics. While variance is an intuitive concept that measures the scatter of the data, covariance on the other hand, is not that intuitive at first but gives a mathematical indication of the degree of change of two random variables together.

1.2.2.0 OBJECTIVE

The main objective of this unit is to provide a broad understanding of the topics Covariance and Variance which is preparatory to the more widely used simple and multiple regression analyses.

1.2.3.0 MAIN CONTENTS

1.2.3.1 Covariance and Variance

Sample covariance is a measure of association between two variables. The sample covariance, Cov(X, Y), is a statistic that enables you to summarize this association with a single number. In general, given $n$ observations on two variables $X$ and $Y$, the sample covariance between $X$ and $Y$ is given by;

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad \ldots[2.19]$$

Where the bar over the variable signifies the sample mean. Therefore, a positive association would be summarized by a positive sample covariance while a negative sample covariance would summarise a negative association.

1.2.3.2 Some Basic Covariance rules

i. Co-variance Rule 1: If $Y = V + W$, Cov(X, Y) = Cov(X, V) + Cov(X, W)

ii. Co-variance Rule 2: If $Y = bZ$, where $b$ is a constant and $Z$ is a variable, Cov(X, Y) = $b$Cov(X, Z)

iii. Co-Variance Rule 3: If $Y = b$, where $b$ is a constant, Cov(X, Y) = 0

For example, Tables 1.2.0(a) and (b) show years of schooling $S$, and hourly earnings $Y$, for a subset of 20 households in the United States. We are required to calculate the covariance.
### Table 1.2(a) Covariance table

<table>
<thead>
<tr>
<th>Observation</th>
<th>S</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>17.24</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>15.00</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>14.91</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4.50</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>18.00</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6.29</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>19.23</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>18.69</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>7.21</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>42.06</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>15.38</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>12.70</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>26.00</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>7.50</td>
</tr>
<tr>
<td>Observation</td>
<td>S</td>
<td>Y</td>
</tr>
<tr>
<td>-------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>17.24</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>15.00</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>14.91</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4.50</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>18.00</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6.29</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>19.23</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>18.69</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>7.21</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>42.06</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>15.38</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>12.70</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>26.00</td>
</tr>
</tbody>
</table>

Table 1.2.0(b) Covariance table
14 | 9  | 7.50 | -1.45 | -6.725 | 28.579  
15 | 15 | 5.00 | 1.75  | -9.225 | -16.143 
17 | 16 | 12.10| 2.75  | -2.125 | -5.842  
18 | 12 | 5.55 | -1.25 | -8.675 | 10.843  
19 | 12 | 7.50 | -1.25 | -6.725 | 8.406   
20 | 14 | 8.00 | 0.75  | -6.225 | -4.668  
**Total** | **265** | **284.49** |  | **305.888**  
**Average** | **13.250** | **14.225** |  |  | **15.294**

Note from the above example that the association is positive. This is given by the positive covariance.

1.2.3.3 Population Covariance

If \( X \) and \( Y \) are random variables, the expected value of the product of their deviations from their means is defined to be the population covariance \( \sigma_{XY} \):

\[
\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \quad \text{...[2.20]}
\]

Where \( \mu_X \) and \( \mu_Y \) are the population means of \( X \) and \( Y \), respectively.

As you would expect, if the population covariance is unknown, the sample covariance will provide an estimate of it, given a sample of observations. However, the estimate will be biased downwards, for

\[
E[COV(X, Y)] = \frac{n-1}{n} \cdot \sigma_{XY} \quad \text{...[2.21]}
\]

The reason is that the sample deviations are measured from the sample means of \( X \) and \( Y \) and tend to underestimate the deviations from the true means. Therefore, we can construct an unbiased estimator by multiplying the sample estimate by \( n/(n-1) \).

1.2.3.4 Sample Variance
For a sample of $n$ observations, $X_1, \ldots, X_2$, the sample variance will be defined as the average squared deviation in the sample:

\[
\text{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \quad \text{...}[2.22]
\]

The sample variance, thus defined, is a biased estimator of the population variance. The reason for the underestimation is because it is calculated as the average squared deviation from the sample mean rather than the true mean. This is because the sample mean is automatically in the centre of the sample, the deviations from it will tend to be smaller than those from the population mean. Therefore, sample variance as an unbiased estimate of population variance is given as:

\[
s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \quad \text{...}[2.23]
\]

### 1.2.3.5 Variance Rule

Variance rule 1: If $Y = V + W$, $\text{Var}(Y) = \text{Var}(V) + \text{Var}(W) + 2\text{Cov}(V, W)$

Variance rule 2: If $Y = bZ$, where $b$ is a constant, $\text{Var}(Y) = b^2 \text{Var}(Z)$

Variance rule 3: If $Y = b$, where $b$ is a constant, $\text{Var}(Y) = 0$.

Variance rule 4: If $Y = V + b$, where $b$ is a constant, $\text{Var}(Y) = \text{Var}(V)$ since the variance of a constant is 0.

### 1.2.4.0 SUMMARY

While explaining the variance and covariance, the temptations to make comparison of the two concepts may not be completely overcome. The unit briefly describes variance as the measure of spread in a population while covariance is considered as a measure of variation of two random variables. Furthermore, the unit showed that variance and covariance are dependent on the magnitude of the data values and cannot be compared; therefore, regulated. This means, covariance is dividing by the product of the standard deviations of the two random variables and variance is normalised into the standard deviation by taking the square root of it.

### 1.2.5.0 CONCLUSION
Variance and covariance concepts as statistical tools are discussed in this unit. How they are estimated were also explained using some basic covariance and variance rules. The existence of population covariance and sample variance estimations were briefly introduced. The introductions and discussions of these two concepts point out that variance can be considered as a special case of covariance.

1.2.6.0 TUTOR-MARKED ASSIGNMENT

1.) In a large bureaucracy the annual salary of each, $Y$, is determined by the formula

$$ Y = 10,000 + 500S + 200T $$

Where, $S$ is the number of years of schooling of the individual and $T$ is the length of time, in years, of employment. $X$ is the individual’s age. Calculate $\text{Cov}(X, Y)$, $\text{Cov}(X, S)$, and $\text{Cov}(X, T)$ for the sample of five individuals shown below and verify that

$$ \text{Cov}(X, Y) = 500\text{Cov}(X, S) + 200\text{Cov}(X, T) $$

2.) In a certain country the tax paid by a firm, $T$, is determined by the rule

$$ T = -1.2 + 0.2P - 0.1I $$

Where, $P$ is profits, and $I$ is an investment, the third term being the effect of an investment incentive. $S$ is sales. All variables are measured in $ million at annual rates. Calculate $\text{Cov}(S, T)$, $\text{Cov}(S, P)$, and $\text{Cov}(S, I)$ for the sample of four firms shown below and verify that

$$ \text{Cov}(S, T) = 0.2\text{Cov}(S, P) - 0.1\text{Cov}(S, I) $$

1.2.7.0 REFERENCES /FURTHER READING


UNIT 3: CORRELATION CO-EFFICIENT

CONTENTS
1.3.1.0 Introduction
1.3.2.0 Objectives
1.3.3.0 Main Content
   1.3.3.1 Properties of the regression coefficients and hypothesis testing
1.3.4.0 Summary
1.3.5.0 Conclusion
1.3.6.0 Tutor-Marked Assignment
1.3.7.0 References/Further Reading

1.3.1.0 INTRODUCTION

This unit introduces a statistic called correlation coefficient. The correlation coefficient describes the direction, whether positive or negative and further measures the degree of relationship that exist between two different variables.

1.3.2.0 OBJECTIVE

The main objective of this unit is to provide ways for which the student may have a simpler understanding of the topic ‘correlation’.

1.3.3.0 MAIN CONTENTS

Correlation measures the degree of association between two or more variables.

1.3.3.1 Properties of the regression coefficients and hypothesis testing

Like variance and covariance, the correlation coefficient comes in two forms, population and sample. \( \rho \) traditionally denotes the population correlation coefficient,
the Greek letter that is the equivalent of “r”, and pronounced “row”, as in row a boat. For variables $X$ and $Y$ it is defined by

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \quad \ldots[3.24]$$

If $X$ and $Y$ are independent, $\rho_{XY}$ will be equal to 0 because the population covariance will be 0. If there is a positive association between them, then we have $\sigma_{XY}$, otherwise $\rho_{XY}$ will still be positive.

If there is an exact positive linear relationship, $\rho_{XY}$ will assume its maximum value of 1. Similarly, if there is a negative relationship $\rho_{XY}$ will be negative, with minimum value of $-1$.

The sample correlation coefficient, $r_{XY}$, is defined by replacing the population covariance and variances by their unbiased estimators. We have seen that these may be obtained by multiplying the sample variances and co-variances by $\frac{n}{n-1}$. Hence,

$$r_{XY} = \frac{n}{n-1} \frac{\text{cov}(XY)}{\sqrt{n \frac{\text{var}(X)}{n-1} \frac{\text{var}(Y)}}} \quad \ldots[3.25]$$

The factors $\frac{n}{n-1}$ could be cancelled out so we can conveniently define the sample correlation by

$$\hat{r}_{XY} = \frac{\text{cov}(XY)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad \ldots[3.26]$$

$$= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \ldots[3.27]$$

Like $\rho$, $r$ has maximum value 1, which is attained when there is a perfect positive association between the sample values of $X$ and $Y$ (when you plot the scatter diagram, the points lie exactly on an upward-sloping straight line). Similarly, it has minimum value $-1$, attained when there is a perfect negative association (the points lying exactly on a
downward-sloping straight line). A value of 0 indicates that there is no association between the observations on X and Y in the sample. Of course the fact that \( r = 0 \) does not necessarily imply that \( \rho = 0 \) or vice versa.

That is:

\[-1 \leq \text{corr}(X,Y) \leq 1\]

\( \text{corr}(X,Y) = 1 \) means perfect positive linear association

\( \text{corr}(X,Y) = -1 \) means perfect negative linear association

\( \text{corr}(X,Y) = 0 \) means no linear association

Figures 1.3(a) to (d) below give more graphical explanations;

Figures 1.3(a) to (d); Scattered diagrams of correlation coefficient as a measure of linear association

Example: For illustration, using the education and earning example, the sample correlation coefficient can be estimated. This is shown below:
Table 1.2.1 Sample correlation coefficient

<table>
<thead>
<tr>
<th>Observ.</th>
<th>S</th>
<th>Y</th>
<th>S̅ − S̅</th>
<th>Y − Y̅</th>
<th>(S − S̅)^2</th>
<th>(Y − Y̅)^2</th>
<th>(S − S̅)(Y − Y̅)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>17.24</td>
<td>1.75</td>
<td>3.016</td>
<td>3.063</td>
<td>9.093</td>
<td>5.277</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>15.00</td>
<td>2.75</td>
<td>0.775</td>
<td>7.563</td>
<td>0.601</td>
<td>2.133</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>14.91</td>
<td>-5.25</td>
<td>0.685</td>
<td>27.563</td>
<td>0.470</td>
<td>-3.599</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4.50</td>
<td>-7.25</td>
<td>-9.725</td>
<td>52.563</td>
<td>94.566</td>
<td>70.503</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>18.00</td>
<td>1.75</td>
<td>3.776</td>
<td>3.063</td>
<td>14.254</td>
<td>6.607</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6.29</td>
<td>-1.25</td>
<td>-7.935</td>
<td>1.563</td>
<td>62.956</td>
<td>9.918</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>19.23</td>
<td>-1.25</td>
<td>5.006</td>
<td>1.563</td>
<td>25.055</td>
<td>-6.257</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
<td>18.69</td>
<td>4.75</td>
<td>4.466</td>
<td>22.563</td>
<td>19.941</td>
<td>21.211</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>7.21</td>
<td>-1.25</td>
<td>-7.015</td>
<td>1.563</td>
<td>49.203</td>
<td>8.768</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>42.06</td>
<td>6.75</td>
<td>27.836</td>
<td>45.563</td>
<td>774.815</td>
<td>187.890</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>15.38</td>
<td>3.75</td>
<td>1.156</td>
<td>14.063</td>
<td>1.335</td>
<td>4.333</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>12.70</td>
<td>-1.25</td>
<td>-1.525</td>
<td>1.563</td>
<td>2.324</td>
<td>1.906</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>26.00</td>
<td>-1.25</td>
<td>11.776</td>
<td>1.563</td>
<td>138.662</td>
<td>-14.719</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>7.50</td>
<td>-1.45</td>
<td>-6.725</td>
<td>18.063</td>
<td>45.219</td>
<td>28.579</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>5.00</td>
<td>1.75</td>
<td>-9.225</td>
<td>3.063</td>
<td>85.091</td>
<td>-16.143</td>
</tr>
</tbody>
</table>
From column 6 and 7, you can see that Var (S) is 10.888 and Var (Y) is 77.108, therefore,

\[
r_{xy} = \frac{15.294}{\sqrt{10.888 \times 77.108}} = \frac{15.294}{28.975} = 0.55
\]

1.3.4.0 SUMMARY

This unit briefly introduced correlation coefficient and showed some properties of regression coefficients and hypothesis testing. Four figures and a table were used to further illustrate the correlation coefficient as a measure of linear association and how sample correlation may be estimated.

1.3.5.0 CONCLUSION

The brief discussion on correlation in this unit is to let the students aware that correlation may be approached as a statistical tool that precedes the basic introduction to econometrics. Although correlation measures may be clouded by relationships that exist with other variables, figures and table were however used to show correlation as the linear relationships between two variables.

1.3.6.0 TUTOR-MARKED ASSIGNMENT
1.) Demonstrate that, in general; the sample correlation coefficient is not affected by a change in the unit of measurement of one of the variables.

2.) Suppose that the observations on two variables $X$ and $Y$ lie on a straight line

$$Y = b_1 + b_2X$$

Demonstrate that $Cov(X, Y) = b_2Var(X)$ and that $Var(Y) = b_2^2Var(X)$, and hence that the sample correlation coefficient is equal to 1 if the slope of the line is positive, $-1$ if it is negative.

1.3.7.0 REFERENCES /FURTHER READING


MODULE 2: SIMPLE EQUATION REGRESSION MODELS

The general aim of this module is to provide students with a thorough understanding of the basic rudiments of simple equation regression models. It shows how a theoretical linear relationship between two variables can be quantified using appropriate data. The principle of least squares regression analysis is explained, and expressions for the coefficients are derived. By the end of this module, students should be able to understand the basic parts of regression analysis. The units to be studied are:

Unit 1: Simple Regression Analyses

Unit 2: Properties of the regression coefficients and hypothesis testing

Unit 3: Multiple regression analysis and Multicollinearity

Unit 4: Transformations of Variables

Unit 5: Dummy Variables


UNIT 1: SIMPLE REGRESSION ANALYSES

CONTENTS
2.1.1.0 Introduction
2.1.2.0 Objectives
2.1.3.0 Main Content
2.1.3.1 Simple Regression Analyses
2.1.3.2 Causes of the Existence of the Disturbance Term
2.1.3.3 Least Squares Regression
   2.1.3.3.1 Least Squares Regression with One Explanatory Variable
   2.1.3.3.2 Alternative Expressions for $b_2$
2.1.1.0 INTRODUCTION

In this unit, an analytic method to measure the association of one or more independent variables with a dependent variable is discussed. Simple regression analyses is a statistical method which can be used to summarize and study relationships existing between two continuous quantitative variables. This unit further introduces the simple regression as a concept that is better understood after a thorough understanding of the basic correlation procedures.

2.1.2.0 OBJECTIVE

The main objective of this unit is to acquaint students with the rudiments of identifying and differentiating simple equation model from multiple regression.

2.1.3.0 MAIN CONTENTS

2.1.3.1 Simple Regression Analyses

The correlation coefficient may indicate that two variables (bivariate regression model) are associated with one another, but it does not give any idea of the kind of relationship involved. While regression predicts the value of the dependent variable based on the known value of the independent variable. In this module further step is taken for cases which we are willing to hypothesize on, than one variable dependence on another. It must be stated immediately that one would not expect to find an exact relationship between any two economic variables unless it is true as a matter of definition. In textbook expositions of economic theory, the usual way of dealing with this awkward fact is to write down the relationship as if it were exact and to warn the reader that it is only an approximation. However, in statistical analysis, one acknowledges the fact that the relationship is not exact by explicitly including in it a
random factor known as the **disturbance term**. We shall start with the simplest possible model:

\[ Y_i = \beta_1 + \beta_2 X_i + \mu_i \]  

...[2.01]

\( Y_i \), the value of the dependent variable in observation \( i \), has two components: (1) the non-random (deterministic term) component \( \beta_1 + \beta_2 X_i \), \( X_i \) being described as the explanatory (or independent/descriptive) variable and the fixed quantities \( \beta_1 \) and \( \beta_2 \) as the parameters of the equation, and (2) the disturbance (stochastic term), \( \mu_i \).

Figure 2.0 illustrates how these two components combine to determine \( Y \). \( X_1, X_2, X_3, \) and \( X_4 \), which are four hypothetical values of the explanatory variable. If the relationship between \( Y \) and \( X \) were exact, the corresponding values of \( Y \) would be represented by the points \( Q_1 \) – \( Q_4 \) on the line. The disturbance term causes the actual values of \( Y \) to be different. In the diagram, the disturbance term has been assumed to be positive in the first and fourth observations and negative in the other two, with the result that, if one plots the actual values of \( Y \) against the values of \( X \), one obtains the points \( P_1 \) – \( P_4 \).

![Figure 2.0 Illustration of independent component combination to give a dependent variable](image)

In practice, the \( P \) points are all not what can be seen in Figure 2.0. The actual values of \( \beta_1 \) and \( \beta_2 \) and hence the location of the \( Q \) points, are unknown, as these are the values of the disturbance term in the observations. The task of regression analysis is to
obtain estimates of $\beta_1$ and $\beta_2$, and hence an estimate of the location of the line, given the $P$ points. As it is, it’s somehow curious. The question “Why then does the disturbance term exist”? would therefore arise. There are several reasons.

2.1.3.2 Reasons for inclusion of Disturbance Term

i. **The omission of explanatory variables**: The relationship between $Y$ and $X$ is almost certain to be a simplification. In reality, there will be other factors affecting $Y$ that have been left out of the non-random dependent component, and their influence will cause the points to lie on the line. It often happens that there are variables that you would like to include in the regression equation but cannot because you are unable to measure them. All of these other factors contribute to the disturbance term.

ii. **Aggregation of variables**: In many cases, the relationship is an attempt to summarise in aggregate some microeconomic relationships. For example, the aggregate consumption function is an attempt to summarize a set of individual expenditure decisions. Since the individual relationships are likely to have different parameters, any attempt to relate aggregate expenditure to aggregate income can only be an approximation. The discrepancy is attributed to the disturbance term.

iii. **Model misspecification**: The model may be misspecified regarding its structure. Just to give one of the many possible examples, if the relationship refers to time series data, the value of $Y$ may depend not on the actual value of $X$ but on the value that had been anticipated in the previous period. If the anticipated and actual values are closely related, there will appear to be a relationship between $Y$ and $X$, but it will only be an approximation, and again the disturbance term will pick up the discrepancy.

iv. **Functional misspecification**: The functional relationship between $Y$ and $X$ may be misspecified mathematically. For example, the true relationship may be non-linear instead of linear. Obviously, one should try to avoid this problem by using an appropriate mathematical specification, but even the most sophisticated specification is likely to be only an approximation, and the discrepancy contributes to the disturbance term.
v. *Measurement error:* If the measurement of one or more of the variables in the relationship is subject to error, the observed values will not appear to conform to an exact relationship, and the discrepancy contributes to the disturbance term.

The disturbance term is the collective outcome of all these factors. Obviously, if you were concerned only with measuring the effect of $X$ on $Y$, it would be much more convenient if the disturbance term did not exist. Were it not for its presence, the $P$ points in Figure 2.1 would coincide with the $Q$ points. Therefore, it would be known that every change in $Y$ from observation to observation was due to a change in $X$, and you would be able to calculate $\beta_1$ and $\beta_2$, exactly. However, part of each change in $Y$ is due to a change in $\mu$, and this makes life more difficult. For this reason, $\mu$ is sometimes described as *noise*.

### 2.1.3.3 Least Squares Regression

Suppose that you are given the four observations on $X$ and $Y$ represented in Figure 2.1 and you are asked to obtain estimates of the values of $\beta_1$ and $\beta_2$, in [2.01]. As a rough approximation, you could do this by plotting the four $P$ points and drawing a line to fit them as best you can, as shown in Figure 2.2. The intersection of the line with the $Y$-axis provides an estimate of the intercept $\beta_1$, which will be denoted $b_1$ and the slope provides an estimate of the slope coefficient $\beta_2$, which will be denoted $b_2$. The fitted line will be written as;
Figure 2.2 Plotting of Observations

Figure 2.3 fitting Plotted Observations

\[ \hat{Y}_i = b_1 + b_2 X_i \]  

The hat mark over \( Y \) in [2.02] indicates that it is the fitted value of \( Y \) corresponding to \( X \) and not the actual value. In Figure 2.3 the fitted points are represented by the points \( R_1 - R_4 \). One thing that should be accepted from the beginning is that however much care you take in drawing the line; you can never discover the true values of \( \beta_1 \) and \( \beta_2 \). \( b_1 \) and \( b_2 \) are only estimates, and they may be good or bad. Once in a while your estimates may be absolutely accurate, but this can only be by coincidence and even then you will have no way of knowing that you have hit the target exactly.

This remains the case even when you use more sophisticated techniques. Drawing a regression line by eye is all very well, but it leaves a lot to subjective judgment. Furthermore, as will become obvious, it is not even possible when you have a variable \( Y \) depending on two or more explanatory variables instead of only one. The question arises, is there a way of calculating good estimates of \( \beta_1 \) and \( \beta_2 \) algebraically? The answer is yes! The first step is to define what is known as a residual for each observation. This is the difference between the actual value of \( Y \) in any observation and the fitted value given by the regression line, that is, the vertical distance between \( P_i \) and \( R_i \) in observation \( i \). Which will be denoted by \( e_i \).
$e_i = Y_i - \hat{Y}_i \quad \ldots[2.03]$  

The residuals for the four observations are shown in Figure 2.3

Substituting [2.02] into [2.03], we obtain

$e_i = Y_i - b_1 - b_2 X_i \quad \ldots[2.04]$  

and hence the residual in each observation depends on our choice of $b_1$ and $b_2$. Obviously, we wish to fit the regression line, that is, choose $b_1$ and $b_2$, in such a way as to make the residuals as small as possible. Equally obvious, a line that fits some observations well will fit others badly and vice versa. We need to devise a criterion of fit that takes account of the size of all the residuals simultaneously. There are some possible criteria, some of which work better than others. It is useless minimizing the sum of the residuals, for example. The sum will automatically be equal to 0 if you make $b_1$ equal to $\bar{Y}$ and $b_2$ equal to 0, obtaining the horizontal line $Y = \bar{Y}$. The positive residuals will then exactly balance the negative ones but other than that, the line will not fit the observations.

One way of overcoming the problem is to minimize $RSS$ (sum of the squares of the residuals).

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2 \quad \ldots[2.05]$$

According to this criterion, the smaller one can make $RSS$ the better is the fit. If one could reduce $RSS$ to 0, one would have a perfect fit, for this would imply that all the residuals are equal to 0. The line would go through all the points, but of course, in general, the disturbance term makes this impossible. There are other quite reasonable solutions, but the least squares criterion yields estimates of $b_1$ and $b_2$ that are unbiased and the most efficient of their type, provided that certain conditions are satisfied. For this reason, the least squares technique is far and away the most popular in uncomplicated applications of regression analysis. The form used here is usually referred to as ordinary least squares and abbreviated OLS.
We shall assume that the true model is:

\[ Y_i = \beta_1 + \beta_2 X_i + \mu_i \]  

...[2.06]

And we shall estimate the coefficients \( b_1 \) and \( b_2 \) of the equation using:

\[ \hat{Y}_i = b_1 + b_2 X_i \]  

...[2.07]

When \( X \) is equal to 1, according to the regression line \( \hat{Y} \) is equal to \( (b_1 + b_2) \). When \( X \) is equal to 2, \( \hat{Y} \) is equal to \( (b_1 + 2b_2) \). Therefore, we can set up Table 2.1. So the residual for the first observation, \( e_1 \), which is given by \( (Y_1 - \hat{Y}_1) \), is equal to \( (3 - b_1 - b_2) \), and \( e_2 \), given by \( (Y_2 - \hat{Y}_2) \), is equal to \( (5 - b_1 - 2b_2) \). Hence

\[
\text{RSS} = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2
\]

\[
= 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2 + 25 + b_1^2 + 4b_2^2 - 10b_1 - 20b_2 + 4b_1b_2
\]

\[
= 34 + 2b_1^2 + 5b_2^2 - 16b_1 - 26b_2 + 6b_1b_2
\]

...[2.08]

Now we want to choose \( b_1 \) and \( b_2 \) so as to minimize \( \text{RSS} \). To do this, we use the calculus and find the values of \( b_1 \) and \( b_2 \) that satisfy

\[
\frac{\partial \text{RSS}}{\partial b_1} = 0 \quad \text{and} \quad \frac{\partial \text{RSS}}{\partial b_2} = 0
\]

...[2.09]

Taking partial differentials of [2.08];
\[ \frac{\partial \text{RSS}}{\partial b_2} = 4b_1 + 6b_2 - 16 \quad \ldots \text{[2.10]} \]

And

\[ \frac{\partial \text{RSS}}{\partial b_2} = 10b_2 + 6b_2 - 16 \quad \ldots \text{[2.11]} \]

And so we have

\[
2b_1 + 3b_2 - 8 = 0
\]

And

\[
3 + 5b_2 - 13 = 0
\]

Solving these two equations, we obtain \( b_1 = 1 \) and \( b_2 = 2 \), and hence the regression equation

\[ \hat{Y}_i = 1 + 2X_i \]

Just to check that we have come to the right conclusion, we shall calculate the residuals:

\[
e_1 = 3 - b_1 - b_2 = 3 - 1 - 2 = 0
\]

\[
e_2 = 5 - b_1 - 2b_2 = 5 - 1 - 4 = 0
\]

Thus both residuals are equal to 0, implying that the line passes exactly through both points.

### 2.1.3.3.1 Least Squares Regression with One Explanatory Variable

We shall now consider the general case where there are \( n \) observations on two variables \( X \) and \( Y \) and supposing \( Y \) to depend on \( X \); we will fit the equation

\[ \hat{Y}_i = b_1 + b_2X_i \quad \ldots \text{[2.12]} \]

The fitted value of the dependent variable in observation \( i \).

\( \hat{Y}_i \) will be \((b_1 + b_2X_i)\) and the residual \( e_i \) will be \((Y_i - b_1 - b_2X_i)\). We wish to choose \( b_1 \) and \( b_2 \) so as to minimize the residual sum of the squares RSS given by
\[ \text{RSS} = e_1^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 \] \[\ldots[2.13]\]

We will find that \(\text{RSS}\) is minimised when

\[ b_2 = \frac{\text{cov}(XY)}{\text{var}(X)} \] \[\ldots[2.14]\]

And

\[ b_1 = \bar{Y} - b_2 \bar{X} \] \[\ldots[2.15]\]

The derivation of the expressions for \(b_1\) and \(b_2\) will follow the same procedure as the derivation in the preceding example, and you can compare the general version with the examples at each step.

We will begin by expressing the square of the residual in observation \(i\) regarding \(b_1\), \(b_2\) and the data on \(X\) and \(Y\):

\[ e_i^2 = (Y_i - \bar{Y})^2 = (Y_i - b_1 - b_2 X_i)^2 = Y_i^2 + b_1^2 + b_2^2 X_i^2 - 2b_1 Y_i - 2b_2 X_i Y_i + 2b_1 b_2 X_i \] \[\ldots[2.16]\]

Summing over all the observations, we can write \(\text{RSS}\) as

\[ \text{RSS} = (Y_1 - b_1 - b_2 X_1)^2 + \cdots + (Y_n - b_1 - b_2 X_n)^2 = \sum_{i=1}^{n} Y_i^2 + nb_1^2 + nb_2^2 \sum_{i=1}^{n} X_i^2 - 2b_1 \sum_{i=1}^{n} Y_i - 2b_2 \sum_{i=1}^{n} X_i Y_i + 2b_1 b_2 \sum_{i=1}^{n} X_i \] \[\ldots[2.17]\]

Note that \(\text{RSS}\) is effectively a quadratic expression in \(b_1\) and \(b_2\), with numerical coefficients determined by the data on \(X\) and \(Y\) in the sample. We can influence the size of \(\text{RSS}\) only through our choice of \(b_1\) and \(b_2\). The data on \(X\) and \(Y\), which determine the locations of the observations in the scatter diagram and are fixed once we have taken the sample. This equation [2.17] is the generalized version of the equations.

The first order conditions for a minimum,

\[ \frac{\partial\text{RSS}}{\partial b_1} = 0 \quad \text{and} \quad \frac{\partial\text{RSS}}{\partial b_2} = 0 \] \[\ldots[2.18]\]

Yield the following equations:
\[
2nb_1 - 2 \sum_{i=1}^{n} Y_i + 2b_2 \sum_{i=1}^{n} X_i = 0
\]
\[
2b_2 \sum_{i=1}^{n} X_i^2 - 2 \sum_{i=1}^{n} Y_i X_i + 2b_1 \sum_{i=1}^{n} X_i = 0 \quad \text{...[2.19]}
\]

Noting that
\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i
\]
may be rewritten as
\[
2nb_1 - 2n\bar{Y} + 2b_2 n\bar{X} = 0 \quad \text{...[2.21]}
\]
and hence
\[
b_1 = \bar{Y} - b_2 \bar{X} \quad \text{...[2.22]}
\]

Substituting for \(b_1\) and again noting that \(\sum_{i=1}^{n} X_i = n\bar{X}\) we obtain
\[
2b_2 \sum_{i=1}^{n} X_i^2 - 2 \sum_{i=1}^{n} Y_i X_i + (\bar{Y} - b_2 \bar{X}) n \bar{X} = 0 \quad \text{...[2.23]}
\]
Separating the terms involving \(b_2\) and not involving \(b_2\) on opposite sides of the equation, we have
\[
2b_2 [\left( \sum_{i=1}^{n} X_i^2 \right) - n\bar{X}^2] = 2 \sum_{i=1}^{n} Y_i X_i - 2n\bar{Y}\bar{X} \quad \text{...[2.24]}
\]
Dividing both sides by \(2n\),
\[
\left[ \frac{1}{n} \left( \sum_{i=1}^{n} X_i^2 \right) - \bar{X}^2 \right] b_2 = \frac{1}{n} (\sum_{i=1}^{n} Y_i X_i) - \bar{Y}\bar{X} \quad \text{...[2.25]}
\]
Using the alternative expressions for sample variance and covariance, this may be rewritten as:
\[
b_2 \text{Var}(X) = \text{Cov}(X, Y)
\]
\[
b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{...[2.26]}
\]

\(b_2\) is from [2.23], \(b_1\) is equally from [2.22]. Those who know about the second-order conditions will have no difficulty confirming that we have minimized RSS.

### 2.1.3.3.2 Alternative Expressions for \(b_2\)
From the definitions of Cov($X$, $Y$) and Var($X$) one can obtain alternative expressions for $b_2$

$$b_2 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2} \quad \ldots [2.27]$$

where,

$$x = X - \bar{X}$$

$$y = Y - \bar{Y}$$

2.1.4.0 SUMMARY

In this unit, you are expected to have learnt the essentials and applications of the concept of simple regression analyses and its estimation.

2.1.5.0 CONCLUSION

In conclusion, the concept of simple regression analyses and its estimation are explained.

2.1.6.0 TUTOR-MARKED ASSIGNMENT

1.) A researcher obtained data on the aggregate expenditure on services $Y$, and aggregate disposable personal income $X$, both measured in N billion at constant prices, for each of the U.S. states and fits the equation

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t + \mu_t$$

The researcher initially fits the equation using OLS regression analysis. However, suspecting that tax evasion causes both $Y$ and $X$ to be substantially underestimated, the researcher
adopts
two alternative methods of compensating for the under-reporting:

a.) The researcher adds N90 billion to the data for Y in each state and N200 billion to the data for X.
b.) The researcher increases the figures for both Y and X in each state by 10 percent.

2.) Derive from first principles the least squares estimator of \( \beta_2 \) and \( \beta_1 \) in the model

\[
Y_i = \beta_1 + \beta_2 X_i + \mu_i
\]

2.1.7.0 REFERENCES /FURTHER READING


UNIT 2: PROPERTIES OF THE REGRESSION COEFFICIENTS AND HYPOTHESIS TESTING

CONTENTS
2.2.1.0 Introduction
2.2.2.0 Objectives
2.2.3.0 Main Content
2.2.3.1 The Random Components of the Regression Coefficients
2.2.3.2 Assumptions Concerning the Disturbance Term
  2.2.3.2.1 Gauss–Markov Condition 1: $E(\mu_i) = 0$ for All Observations
  2.2.3.2.2 Gauss–Markov Condition 2: Population Variance of $\mu_i$ Constant for All Observations
  2.2.3.2.3 Gauss–Markov Condition 3: $\mu_i$ Distributed Independently of $\mu_j$ ($i \neq j$)
  2.2.3.2.4 Gauss–Markov Condition 4: $u$ Distributed Independently of the Explanatory Variables
2.2.3.3 The Normality Assumption
2.2.3.4 Unbiasedness of the Regression Coefficients
2.2.3.5 Precision of the Regression Coefficients
2.2.3.6 Testing Hypotheses Relating to the Regression Coefficients
  2.2.3.6.1 Formulation of a Null Hypothesis
  2.2.3.6.2 Developing the Implications of a Hypothesis
  2.2.3.7 Compatibility, Freakiness, and the Significance Level
  2.2.3.8 What Happens if the Standard Deviation of $b_2$ is Not Known
2.2.4.0 Conclusion
2.2.5.0 Summary
2.2.6.0 Tutor-Marked Assignment
2.2.7.0 References/Further Reading

2.2.1.0 INTRODUCTION
This unit firstly attempts giving an appropriate explanation to the concept of GAUSS-MARKOV THEOREM before proceeding into the discussion of the properties of regression coefficients and hypothesis testing. However, to properly understand this unit, a brief discussion would be made of some knowledge areas like;

i. Estimators
ii. Assumptions underlying the Classical Linear Regression Model (CLRM)
iii. Properties of Ordinary Least Square (OLS) estimator
iv. Statistical inference (hypothesis testing) like; null and alternative hypotheses.

The basic knowledge of the aforementioned areas are what the students must be equipped with before proceeding in this unit.

- Estimators
  A rule for calculating an estimate of a given quantity based on observed data is referred to as an estimator. Hence in the calculation three quantities are distinguished; the quantity of interest, referred to as an estimand, the result (estimate) and the rule (estimator).
  The properties of estimators are the concerns of Estimation theory. The theory defines and determines properties that can be used under given circumstances to relate different estimators or different rules for creating estimates for the same quantity which are built on the same data.
  A simple example of an estimator is given by a sample mean equation of a population mean shown below.

\[
\bar{k} = \frac{1}{N} \sum_{i=1}^{N} k_i
\]

where, \( \bar{k} \) is an estimator for the population mean \( \mu \).

- Assumptions underlying Classical Linear Regression Model (CLRM)
  This model is the basis of most econometric theory and provides a description of the method of ordinary least squares (OLS). It also explains...
how the OLS, using sample data, can estimate unknown parameters of a regression equation. Furthermore, it makes available opportunity to ask whether statements about the true unknown parameters of the model, based on our estimated values can be made. In doing this, there is need to make a number of assumptions. These assumptions, if satisfied, ensure that the estimators being used are accurate and efficient. Precise predictions about the unknown model parameters can also be made through satisfactory assumptions. Here is a summary of the 10 assumptions in CLRM:

**Assumption 1: Linear regression model.** The regression model is linear in the parameters, as shown in [01]

$$Y_i = \beta_1 + \beta_2 X_i + \mu \quad \ldots [01]$$

where,

$Y$ is the regressand $Y$ and the regressor $X$ may be nonlinear.

**Assumption 2: X values are fixed in repeated sampling.** Values taken by the regressor $X$ are considered fixed in repeated samples. More precisely, $X$ is assumed to be nonstochastic.

**Assumption 3: Zero means value of disturbance $\mu$.** That is, given the value of $X$, the mean value of the random disturbance term $\mu$ is zero. This shows that the uncertain mean value of $\mu$ is zero, as shown in [02]

$$E(\mu \mid X_i) = 0 \quad \ldots [02]$$

**Assumption 4: Equal variance of $\mu$.** If given the value of $X$, the variance of $\mu$ is the same for all observations. Which means that the uncertain variances of $\mu$ are alike, as shown in [03].
\[ \text{var}(\mu | X_i) = E[\mu - E(\mu | X_i)]^2 = E(\mu^2 | X_i) \quad (\text{due to assumption 3}) \]

where,
\[ = \sigma^2 \quad ...[03] \]

Assumption 5: No autocorrelation between the disturbances.
If given any two \( X \) values, \( X_i \) and \( X_j (i \neq j) \), the autocorrelation between any two \( \mu_i \) and \( \mu_j (i \neq j) \) is zero, as shown in [04].

\[ \text{cov}(\mu, \mu | X_i, X_j) = E[(\mu - E(\mu))|X_i][\mu_j - E(\mu_j)] X_j] \]

\[ = E(\mu | X_i)(\mu_j | X_j) \quad (\text{shows } \mu \text{ and } \mu_j \text{ are uncorrelated}) \]

\[ ...[04] \]

\[ = 0 \]

where,
\( (i \text{ and } j) \) are different observations and \( \text{cov} \) is covariance.

Assumption 6: Zero covariance between \((\mu \text{ and } X_i)\). As earlier expressed;
\[ \text{cov}(\mu, X_i) = E[\mu - E(\mu)]X_i - E(X_i)] \]

since \( E(\mu) = 0 \)

\[ = E[\mu(X_i - E(X_i))] \]

\[ = E(\mu X_i) - E(X_i)E(\mu) \]

since \( E(X_i) \) is nonstochastic and \( E(\mu) = 0 \)
\[ E(\mu X_i) = 0 \quad \ldots[05]\]

Assumption 7: The number of observations \( n \) must be greater than the number of parameters to be estimated. On the other hand, the number of observations \( n \) must be greater than the number of descriptive (explanatory) variables.

Assumption 8: Variability in \( X \) values. The \( X \) values in a given sample must not all be the same. That is, \( var(X) \) must be a finite positive number.

Assumption 9: The regression model is correctly specified. On the other hand, there is no error in the model used in observed analysis.

Assumption 10: There is no perfect multicollinearity. That is, there are no perfect linear relationships among the descriptive variables.
Multicollinearity as a problem associated with CLRM is discussed in unit 3.

- Properties of Ordinary Least Square (OLS) estimator

The following properties are associated with OLS estimators:

1. Linearity
2. Unbiasedness
3. Efficient: it has the minimum variance
4. Consistency
5. Asymptotic Unbiasedness
The OLS estimator is sometimes referred to as the CLRM and in data analysis the best estimator is refer to as BLUE (best linear unbiased estimator). Therefore, the OLS estimator requires that the descriptive variables are received outside a data group and there is no perfect multicollinearity. Also, OLS is best in the class of linear unbiased estimators when the errors are vector of random variables and successively uncorrelated. Within these conditions, the OLS offers minimum-variance mean-unbiased estimation when the errors have fixed variances. Again, the OLS is a maximum likelihood estimator under the additional assumption that the errors are normally distributed. So, whenever students are planning to use a linear regression model by means of OLS, each time check for the OLS assumptions. In as much as the OLS assumptions are satisfied, the analysis becomes simpler. Through the Gauss-Markov theorem (as will be seen later in this unit) students can directly use OLS for the best results. When the OLS estimator is asymptotically normal and a consistent estimator of the asymptotic covariance matrix is available to carry out hypothesis tests on the coefficients of a linear regression model.

2.2.2.0 OBJECTIVE

The main objective of this unit is to provide basic understanding of the topic, properties of regression coefficients and hypothesis testing. As well as how these properties form the basis for prediction and forecasting analyses. Focus will also be on the use of regression analysis to recognise which among the independent variables are related to the dependent variable and to explore the forms of these relationships.

2.2.3.0 MAIN CONTENTS

With the aid of regression analysis, we can obtain estimates of the parameters of a relationship. However, they are only estimates. The next question to ask is, how reliable are they? We shall answer this first in general terms, investigating the conditions for unbiasedness and the factors governing their variance. Secondly, building on those conditions for unbiasedness and their variances, we shall develop a means of testing whether a regression estimate is compatible with a specific prior hypothesis concerning the true value of a parameter. Hence, we shall derive a confidence interval for the true value,
that is, the set of all hypothetical values not contradicted by the experimental result. We shall also see how to test whether the goodness of fit of a regression equation is better than what might be expected by pure chance.

2.2.3.1 The Random Components of the Regression Coefficients

The least squares regression coefficient is a special form of a random variable whose properties depend on those of the disturbance term in the equation. This will be demonstrated first theoretically and then using a controlled experiment. In particular, we will investigate the implications for the regression coefficients of certain assumptions concerning the disturbance term. Throughout the discussion, we shall continue to work with the simple regression model where \( Y \) depends on \( X \) according to

\[
Y_i = \beta_1 + \beta_2 X_i + \mu_i
\]

And we fit the regression equation \( \hat{Y}_i = b_1 + b_2 X_i \) given a sample of \( n \) observations.

We shall also continue to assume that \( X \) is a non-stochastic exogenous (not external randomly determined) variable; that is, that its value in each observation may be considered to be predetermined by factors unconnected with the present relationship.

First, note that \( Y_i \) has two components. It has non-random component \( (\beta_1 + \beta_2 X_i) \), which owes nothing to the laws of chance (\( \beta_1 \) and \( \beta_2 \) may be unknown, but nevertheless they are fixed constants) and it has the random component \( \mu_i \). This implies that, when we calculate \( b_2 \) according to the usual formula;

\[
b_2 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}
\]  

\[\text{...}[2.28]\]

\( b_2 \) would also have a random component \( \text{Cov}(X,Y) \). \( \text{Cov}(X,Y) \) depends on the values of \( Y \), and the values of \( Y \) depend on the values of \( \mu \). If the values of the disturbance term had been different in the \( n \) observations, we would have obtained different values of \( Y \), hence of \( \text{Cov}(X,Y) \), and hence of \( b_2 \). Thus we have shown that the regression coefficient \( b_2 \) obtained from any sample consists of (1) a fixed component, equal to the true value \( \beta_2 \), and (2) a random component dependent on \( \text{Cov}(X, \mu) \), which is responsible for its variations around this central tendency. Similarly, one may easily show that \( b_1 \) has a fixed component equal to the true value \( \beta_1 \), plus a random component that depends on the random factor \( \mu \).

2.2.3.2 Assumptions Concerning the Disturbance Term
It is thus obvious that the properties of the regression coefficients depend critically on the properties of the disturbance term. Indeed the latter has to satisfy four conditions, known as the Gauss–Markov conditions, if ordinary least squares regression analysis is to give the best possible results. If they are not satisfied, the user should be aware of the fact. If remedial action is possible, he or she should be capable of taking it. If it is not possible, he or she should be able to judge how seriously the results may have been affected.

2.2.3.2.1 Gauss–Markov Condition 1: \( E(\mu_i) = 0 \) for All Observations

The first condition is that the expected value of the disturbance term in any observation should be 0. Sometimes it will be positive, sometimes negative, but it should not have a systematic tendency in either direction. If an intercept is included in the regression equation, it is usually reasonable to assume that this condition is satisfied automatically since the role of the intercept is to pick up any systematic but constant tendency in \( Y \) not accounted for by the explanatory variables included in the regression equation.

2.2.3.2.2 Gauss–Markov Condition 2: Population Variance of \( \mu_i \) Constant for All Observations

The second condition is that the population variance of the disturbance term should be constant for all observations. Sometimes the disturbance term will be greater, sometimes smaller, but there should not be any a priori reason for it to be more erratic in some observations than in others. The constant is usually denoted by \( \sigma^2_\mu \), often abbreviated to \( \sigma^2 \), and the condition is written as, \( \sigma^2_{\mu i} = \sigma^2 \) for all \( i \). Since \( E(\mu_i) \) is 0, the population variance of \( \mu_i \) is equal to \( E(\mu_i^2) \), so the condition can also be written

\[
E(\mu_i^2) = \sigma^2 \text{ for all } i, \quad \sigma_\mu \text{ of course is unknown. One of the tasks of regression analysis is to estimate the standard deviation of the disturbance term. If this condition is not satisfied, the OLS regression coefficients will be inefficient, but you should be able to obtain more reliable results by using a modification of the regression technique.}

2.2.3.2.3 Gauss–Markov Condition 3: \( \mu_i \) Distributed Independently of \( \mu_j \) (\( i \neq j \))
This condition states that there should be no systematic association between the values of the disturbance term in any two observations. For example, just because the disturbance term is large and positive in one observation, there should be no tendency for it to be large and positive in the next (or large and negative, for that matter, or small and positive, or small and negative). The values of the disturbance term should be independent of one another. The condition implies that $\sigma_{\mu_i\mu_j}$, the population covariance between $\mu_i$ and $\mu_j$, is 0, because:

$$\sigma_{\mu_i\mu_j} = E[(\mu_i - \mu)(\mu_j - \mu)] = E(\mu_i)E(\mu_j) = 0$$

...[2.29]

where, $u$ is a value in $\mu$ as shown in $(u_i)$ of Figure 2.0

Note that the population means of $\mu_i$ and $\mu_j$ are 0, by the first Gauss–Markov condition, and that $E(\mu_i)E(\mu_j)$ can be decomposed as $E(\mu_i)E(\mu_j)$ if $\mu_i$ and $\mu_j$ are generated independently. If this condition is not satisfied, OLS will again give inefficient estimates.

### 2.2.3.2.4 Gauss–Markov Condition 4: $u$ Distributed Independently of the Explanatory Variables

The final condition comes in two versions, weak and strong. The strong version is that the explanatory variables should be non-stochastic, that is, not have random components. This is very unrealistic for economic variables, and we will eventually switch to the weak version of the condition, where the explanatory variables are allowed to have random components provided that they are distributed independently of the disturbance term. However, the strong version is usually used because it simplifies the analysis of the properties of the estimators.

$$\sigma X_i u_i = E[(X_i - E(X_i))(u_i - \mu u)] = (X_i - E(X_i))E(u_i) = 0$$

...[2.30]

### 2.2.3.3 The Normality Assumption
In addition to the Gauss–Markov conditions, one usually assumes that the disturbance term is normally distributed. The reason is that if $u$ is normally distributed, so will be the regression coefficients, and this is useful when performing tests of hypotheses and constructing confidence intervals for $\beta_1$ and $\beta_2$ using the regression results. The justification for the assumption depends on the Central Limit Theorem; that, if a random variable is the composite result of the effects of a large number of other random variables, it will have an approximately normal distribution even if its components do not, provided that none of them is dominant. The disturbance term $u$ is composed of a number of factors not appearing explicitly in the regression equation so, even if we know nothing about the distribution of these factors (or even their identity), we are entitled to assume that they are normally distributed.

2.2.3.4 Unbiasedness of the Regression Coefficients

We can show that $b_2$ must be an unbiased estimator of $\beta_2$ if the fourth Gauss–Markov condition is satisfied:

$$E(b_2) = E\left[\beta_2 + \frac{\text{Cov}(X,u)}{\text{Var}(X)}\right] = \beta_2 + E\left[\frac{\text{Cov}(X,u)}{\text{Var}(X)}\right] \quad \ldots[2.31]$$

since $\beta_2$ is a constant. If we adopt the strong version of the fourth Gauss–Markov condition and assume that $X$ is non-random, we may also take $\text{Var}(X)$ as a given constant, and so

$$E(b_2) = \beta_2 + \frac{1}{\text{Var}(X)}E[\text{Cov}(X,u)] \quad \ldots[2.32]$$

To demonstrate that $E[\text{Cov}(X,u)]$ is 0:

$$E[\text{Cov}(X,u)] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})\right] = \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \bar{X})(u_i - \bar{u})]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})E[(u_i - \bar{u})] = 0 \quad \ldots[2.33]$$
In the second line, the second expected value rule has been used to bring \((1/n)\) out of the expression as a common factor, and the first rule has been used to break up the expectation of the sum into the sum of the expectations. In the third line, the term involving \(X\) has been brought out because \(X\) is non-stochastic. By virtue of the first Gauss–Markov condition, \(E(u_i)\) is 0, and hence \(E(u)\) is also 0. Therefore \(E[Cov(X, u)]\) is 0 and

\[ E(b_2) = \beta_2 \quad \ldots[2.34] \]

In other words, \(b_2\) is an unbiased estimator of \(\beta_2\). We can obtain the same result with the weak version of the fourth Gauss–Markov condition (allowing \(X\) to have a random component but assuming that it is distributed independently of \(u\)), unless the random factor in the \(n\) observations happens to cancel out exactly, which can happen only by coincidence. \(b_2\) will be different from \(\beta_2\) for any given sample, but in view of unbiased regression coefficient, there will be no systematic tendency for it to be either higher or lower. The same is true for the regression coefficient \(b_1\).

Using [2.22]

\[ b_1 = \bar{Y} - b_2 \bar{X} \quad \ldots[2.35] \]

Hence

\[ E(b_1) = E(\bar{Y}) - \bar{X}E(b_2) \quad \ldots[2.36] \]

Since \(Y_i\) is determined by

\[ Y_i = \beta_1 + \beta_2 X_i + \mu_i \]

We have

\[ E(Y_i) = \beta_1 + \beta_2 X_i + E(\mu_i) = \beta_1 + \beta_2 X_i \quad \ldots[2.37] \]

because \(E(\mu_i)\) is 0 if the first Gauss–Markov condition is satisfied. Hence

\[ E(\bar{Y}) = \beta_1 + \beta_2 \bar{X} \quad \ldots[2.38] \]

Substituting this into [2.36], and using the result that \(E(b_2) = \beta_2\).
Thus $b_1$ is an unbiased estimator of $\beta_1$ provided that the Gauss–Markov conditions 1 and 4 are satisfied. Of course in any given sample the random factor will cause $b_1$ to differ from $\beta_1$.

### 2.2.3.5 Precision of the Regression Coefficients

Now we shall consider $\sigma^2_{b_1}$ and $\sigma^2_{b_2}$, the population variances of $b_1$ and $b_2$ about their population means.

The following expressions give these

\[
\sigma^2_{b_1} = \frac{\sigma^2_u}{n} \left[ 1 + \frac{X^2}{\text{Var}(X)} \right] \quad \text{and} \quad \sigma^2_{b_2} = \frac{\sigma^2_u}{n\text{Var}(X)} \quad \ldots[2.40]
\]

Equation [2.40] has three obvious implications. First, the variances of both $b_1$ and $b_2$ are directly inversely proportional to the number of observations in the sample. This makes good sense. The more information you have, the more accurate your estimates are likely to be.

Second, the variances are proportional to the variance of the disturbance term. The bigger the variance of the random factor in the relationship, the worse the estimates of the parameters are likely to be.

Third, the variance of the regression coefficients is inversely related to the variance of $X$. What is the reason for this? Remember that (1) the regression coefficients are calculated on the assumption that the observed variations in $Y$ are due to variations in $X$, but (2) they are in reality partly due to variations in $X$ and partly to variations in $u$. The smaller the variance of $X$, the greater is likely to be the relative influence of the random factor in determining the variations in $Y$ and the more likely is regression analysis give inaccurate estimates.

### 2.2.3.6 Testing Hypotheses Relating to the Regression Coefficients
Which comes first, theoretical hypothesizing or empirical research? In practice, theorizing and experimentation feed on each other, and questions of this type cannot be answered. For this reason, we will approach the topic of hypothesis testing from both directions. On the one hand, we may suppose that the theory has come first and that the purpose of the experiment is to evaluate its acceptability. This will lead to the execution of significance tests. Alternatively, we may perform the experiment first and then consider what theoretical hypotheses would be consistent with the results. This will lead to the construction of confidence intervals.

Students would already have encountered the logic underlying significance tests and confidence intervals in an introductory statistics course. Students will thus be familiar with most of the concepts in the following applications to regression analysis. There is, however, one topic that may be new: the use of one-tailed tests. Such tests are used very frequently in regression analysis. Indeed, they are, or they ought to be, more common than the traditional textbook two-tailed tests. It is, therefore, important that you understand the rationale for their use, and this involves a sequence of small analytical steps. None of this should present any difficulty, but be warned that, if students attempt to use a shortcut or, worse, try to reduce the whole business to the mechanical use of a few formulae, you will be asking for trouble.

2.2.3.6.1 Formulation of a Null Hypothesis

We will start by assuming that the theory precedes the experiment and that you have some hypothetical relationship in your mind. For example, you may believe that the percentage rate of price inflation in an economy, $p$, depends on the percentage rate of wage inflation, $w$, according to the linear equation

$$p = \beta_1 + \beta_2 w + u$$

...[2.41]
where $\beta_1$ and $\beta_2$ are parameters and $u$ is a disturbance term. You might further hypothesize that, apart from the effects of the disturbance term, price inflation is equal to wage inflation. Under these circumstances you would say that the hypothesis that you are going to test, known as your null hypothesis and denoted $H_0$, is that $\beta_2$ is equal to 1. We also define an alternative hypothesis, denoted $H_1$, which represents your conclusion if the experimental test indicates that $H_0$ is false. In the present case $H_1$ is simply that $\beta_2$ is not equal to 1. The two hypotheses are stated using the notation

$$H_0: \beta_2 = 1$$
$$H_1: \beta_2 \neq 1$$

In this particular case, if we believe that price inflation is equal to wage inflation, we are trying to establish the credibility of $H_0$ by subjecting it to the strictest possible test and hoping that it emerges intact. In practice, however, it is more usual to set up a null hypothesis and attack it with the objective of establishing the alternative hypothesis as the correct conclusion. For example, consider the simple earnings function

$$EARNINGS = \beta_1 + \beta_2 S + u$$

Where $EARNINGS$ is hourly earnings in dollars and $S$ is years of schooling. On very reasonable theoretical grounds, you expect earnings to be dependent on schooling, but your theory is not strong enough to enable you to specify a particular value for $\beta_2$. You can nevertheless establish the dependence of earnings on schooling by the inverse procedure in which you take as your null hypothesis the assertion that earnings does not depend on schooling, that is, that $\beta_2$ is 0. Your alternative hypothesis is that $\beta_2$ is not equal to 0, that is, that schooling does affect earnings. If you can reject the null hypothesis, you have established the relationship, at least in general terms. Using the conventional notation, your null and alternative hypotheses are $H_0: \beta_2 \neq 0$ and $H_1: \beta_2 = 0$, respectively.

The following discussion uses the simple regression model

$$Y_t = \beta_1 + \beta_2 X_t + \mu_t$$
It will be confined to the slope coefficient, $\beta_2$, but exactly the same procedures are applied to the constant term, $\beta_1$. We will take the general case, where you have defined a null hypothesis that $\beta_2$ is equal to some specific value, say $\beta_2^0$, and the alternative hypothesis is that $\beta_2$ is not equal to this value ($H_0: \beta_2 = \beta_2^0$, $H1: \beta_2 \neq \beta_2^0$); you may be attempting to attack or defend the null hypothesis as it suits your purpose. We will assume that the four Gauss–Markov conditions are satisfied.

**2.2.3.6.2 Developing the Implications of a Hypothesis**

If $H_0$ is correct, values of $b_2$ obtained using regression analysis in repeated samples will be distributed with mean $\beta_2^0$ and $\frac{\sigma_u^2}{nVar(X)}$, we will now introduce the assumption that $u$ has a normal distribution. If this is the case, $b_2$ will also be normally distributed. In view of the structure of the normal distribution, most values of $b_2$ will lie within two standard deviations of $\beta_2^0$ (if $H0: \beta_2 = \beta_2^0$ is true).

**2.2.3.7 Compatibility, Freakiness, and the Significance Level**

Now, suppose that we take an actual sample of observations on average rates of price inflation and wage inflation over the past five years for a sample of countries and estimate $\beta_2$ using regression analysis. If the estimate is close to 1.0, we should almost certainly be satisfied with the null hypothesis, since it and the sample result are compatible with one another. But suppose, on the other hand, that the estimate is a long way from 1.0. Suppose that it is equal to 0.7. This is three standard deviations below 1.0. If the null hypothesis is correct, the probability of being three standard deviations away from the mean, positive or negative, is only 0.0027, which is very low. You could come to either of two conclusions about this worrisome result:

You could continue to maintain that your null hypothesis $H_0: \beta_1 = 1$ is correct, and that the experiment has given a freak result. You concede that the probability of such a low value of $b_2$ is very small, nevertheless it does occur 0.27 percent of the time and you reckon that this is one of those times.
Or you could conclude that the regression result contradicts the hypothesis. You are not convinced by the explanation in (1) because the probability is so small and you think that a much more likely explanation is that $\beta_2$ is not really equal to 1. In other words, you adopt the alternative hypothesis $H_1: \beta_2 \neq 1$ instead.

We can summarize this decision rule mathematically by saying that we will reject the null hypothesis if

$$z > 1.96 \text{ or } z < -1.96 \quad \cdots [2.43]$$

where $z$ is the number of standard deviations between the regression estimate and the hypothetical value of $\beta_2$:

$$z = \frac{\text{distance between regression estimate and hypothetical value}}{\text{standard deviation of } b_2} = \frac{b_2 - \beta_2^0}{s.d.(b_2)} \quad \cdots [2.44]$$

The null hypothesis will not be rejected if

$$-1.96 \leq z \leq 1.96$$

This condition can be expressed by substituting for $z$ from

$$-1.96 \leq \frac{b_2 - \beta_2^0}{s.d.(b_2)} \leq 1.96 \quad \cdots [2.45]$$

Multiplying through by the standard deviation of $b_2$, one obtains

$$-1.96 \cdot s.d.(b_2) \leq b_2 - \beta_2^0 \leq 1.96 \cdot s.d.(b_2) \quad \cdots [2.46]$$

from which one obtains

$$\beta_2^0 - 1.96 \cdot s.d.(b_2) \leq b_2 \leq \beta_2^0 + 1.96 \cdot s.d.(b_2) \quad \cdots [2.47]$$

[2.47] gives the set of values of $b_2$ which will not lead to the rejection of a specific null hypothesis $2 = \beta_2^0$. It is known as the acceptance region for $b_2$, at the 5 percent significance level.

2.2.3.8 What Happens if the Standard Deviation of $b_2$ is Not Known
So far we have assumed that the standard deviation of $b_2$ is known, which is most unlikely in practice. It has to be estimated by the standard error of $b_2$. This causes two modifications to the test procedure. First, $z$ is now defined using $s.e.(b_2)$ instead of $s.d.(b_2)$ and it is referred to as the $t$ statistic:

$$t = \frac{b_2 - \beta_2^0}{s.d.(b_2)} \quad \text{...}[2.48]$$

Second, the critical levels of $t$ depend on what is known as a $t$ distribution instead of a normal distribution. We will not go into the reasons for this, or even describe the $t$ distribution mathematically. But enough to say that it is a partner of the normal distribution. Its exact shape depends on the number of degrees of freedom in the regression and approximates the normal distribution increasingly closely as the number of degrees of freedom increases. You will certainly have encountered the $t$ distribution in your introductory statistics course.

The estimation of each parameter in a regression equation consumes one degree of freedom in the sample. Hence the number of degrees of freedom is equal to the number of observations in the sample minus the number of parameters estimated. The parameters are constant (assuming that this is specified in the regression model) and the coefficients of the explanatory variables. In the present case of simple regression analysis, only two parameters, $\beta_1 + \beta_2$, are estimated and hence the number of degrees of freedom is $n - 2$. It should be emphasized that a more general expression will be required when we come to multiple regression analysis.

The critical value of $t$, which we will denote $t_{crit}$, replaces the number 1.96 in [2.43], so the condition that a regression estimate should not lead to the rejection of a null hypothesis $H_0$: $\beta_2 = \beta_2^0$ is

$$-t_{crit} \leq \frac{b_2 - \beta_2^0}{s.d.(b_2)} \leq t_{crit} \quad \text{...}[2.49]$$

Hence we have the decision rule:
reject $H_0$ if $\left| \frac{b_2 - \beta_2^0}{s.d.(b_2)} \right| > t_{crit}$,

do not reject if $\left| \frac{b_2 - \beta_2^0}{s.d.(b_2)} \right| < t_{crit}$

Where $\left| \frac{b_2 - \beta_2^0}{s.d.(b_2)} \right|$ is the absolute value (numerical value, neglecting the sign) of $t$.

2.2.4.0 SUMMARY

In this unit, basic understanding of the properties of regression coefficients and hypotheses testing has been explained. As introductory step to understanding this unit, some knowledge areas like estimators, assumptions underlying CLRM and properties of OLS estimators were briefly discussed. These knowledge areas acquainted the students of what is expected to be known for better understanding of the different aspect of regression Coefficients, hypotheses testing and the assumptions associated with studying these topics in this unit.

2.2.5.0 CONCLUSION

Assumptions and basic knowledge areas needed for the students to be acquainted with the properties of regression coefficients and hypotheses testing have been introduced and discussed in this unit. Most especially, the respective Gauss-Markov conditions in the assumptions concerning disturbance term of regression analyses were discussed in a manner that the students would be able to properly understand. Also, formulation of null hypothesis and developing implications of a hypothesis were discussed as testing hypotheses relating to the regression coefficients.

2.2.6.0 TUTOR-MARKED ASSIGNMENT

1.) Where performance on a game of skill is measured numerically, the improvement that comes with practice is called a learning curve. This is especially obvious with some arcade-type games. The first time players try a new one; they are likely to score very little. With more attempts, their scores should gradually improve as they become
acustomed to the game, although, obviously, there will be variations caused by the luck factor. Suppose that the learning curve determines their scores

\[ Y_i = 500 + 100X_i + \mu_i \]

where, \( Y \) is the score, \( X \) is the number of times that they have played before, and \( \mu \) is a disturbance term. The following table gives the results of the first 20 games of a new player. \( X \) automatically goes from 0 to 19; \( \mu \) was set equal to 400 times the numbers generated by a normally distributed random variable with 0 mean and unit variance, and \( X \) and \( \mu \) determined \( Y \) according to the learning curve.

<table>
<thead>
<tr>
<th>Observation</th>
<th>( X )</th>
<th>( \mu )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-236</td>
<td>264</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-96</td>
<td>504</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-332</td>
<td>368</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>12</td>
<td>812</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>-152</td>
<td>748</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>-876</td>
<td>124</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>412</td>
<td>1,512</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>96</td>
<td>1,296</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>1,012</td>
<td>2,312</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>-52</td>
<td>1,348</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>636</td>
<td>2,136</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>-368</td>
<td>1,232</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>-284</td>
<td>1,416</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>-100</td>
<td>1,700</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>676</td>
<td>2,576</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>60</td>
<td>2,060</td>
</tr>
</tbody>
</table>
Regressing $Y$ on $X$, one obtains the equation (standard errors in parentheses):

$$\hat{Y} = 369 + 116.8X$$

S.E.E (190) (17.1)

Why is the constant in this equation not equal to 500 and the coefficient of $X$ not equal to 100?

What is the meaning of the standard errors?

2.) The experiment is repeated with nine other new players (the disturbance term being generated by 400 times a different set of 20 random numbers in each case), and the regression results for all ten players are shown in the following table. Why do the constant, the coefficient of $X$, and the standard errors vary from sample to sample?

<table>
<thead>
<tr>
<th>Player</th>
<th>Constant</th>
<th>Standard error of constant</th>
<th>Coefficient of $X$</th>
<th>Standard error of coefficient of $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>369</td>
<td>190</td>
<td>116.8</td>
<td>17.1</td>
</tr>
<tr>
<td>2</td>
<td>699</td>
<td>184</td>
<td>90.1</td>
<td>16.5</td>
</tr>
<tr>
<td>3</td>
<td>531</td>
<td>169</td>
<td>78.5</td>
<td>15.2</td>
</tr>
<tr>
<td>4</td>
<td>555</td>
<td>158</td>
<td>99.5</td>
<td>14.2</td>
</tr>
<tr>
<td>5</td>
<td>407</td>
<td>120</td>
<td>122.6</td>
<td>10.8</td>
</tr>
<tr>
<td>6</td>
<td>427</td>
<td>194</td>
<td>104.3</td>
<td>17.5</td>
</tr>
<tr>
<td>7</td>
<td>412</td>
<td>175</td>
<td>123.8</td>
<td>15.8</td>
</tr>
<tr>
<td>8</td>
<td>613</td>
<td>192</td>
<td>95.8</td>
<td>17.3</td>
</tr>
</tbody>
</table>
The variance of $X$ is equal to 33.25, and the population variance of $\mu$ is equal to 160,000. Using appropriate equation, show that the standard deviation of the probability density function of the coefficient of $X$ is equal to 15.5. Are the standard errors in the table good estimates of this standard deviation?

### 2.2.7.0 REFERENCES /FURTHER READING


**UNIT 3 MULTIPLE REGRESSION ANALYSIS AND MULTICOLLINEARITY**

### CONTENTS

2.3.1.0 Introduction
2.3.2.0 Objectives
2.3.3.0 Main Content
   2.3.3.1 Multiple Regression Coefficients Interpretation
   2.3.3.2 Properties of the Multiple Regression Coefficients
   2.3.3.3 $t$ Tests and Confidence Intervals
   2.3.3.4 Consistency
2.3.4.0 Multicollinearity
   2.3.4.1 Multicollinearity in Models with More Than Two Explanatory Variables
   2.3.4.2 Ways to alleviate multicollinearity problems
2.3.5.0 Summary
2.3.6.0 Conclusion
2.3.7.0 Tutor-Marked Assignment
2.3.8.0 References/Further Reading

2.3.1.0 INTRODUCTION

The multiple regression analysis is an extension of simple regression analysis. It covers cases in which the dependent variable is hypothesized to depend on more than one descriptive variable. Most of the multiple regression analysis is a direct extension of the simple regression model but has only two new dimensions. First, when evaluating the influence of a given descriptive variable on the dependent variable, we would now have to face the problem of discriminating between its effects and the effects of the other descriptive variables. Second, we shall have to tackle the problem of model specification. Often some variables might be thought to influence the behaviour of the dependent variable; though, they might be unconnected. We shall have to decide which should be included in the regression equation and which should be omitted. However, the arrangement of flow for the multiple regression analysis is to firstly, carry out derivation of formula, then estimation procedures using values, followed by presentation of results and lastly interpretations. As an extension of unit 2, we shall discuss multicollinearity being a problem associated with CLRM.

2.3.2.0 OBJECTIVE

The main objective of this unit is to provide broad understanding of the topic; multiple regression analysis and appropriate alleviation measures associated with multicollinearity problems. This understanding includes the knowledge of the properties, principles behind the derivation of and how to interpret multiple regression coefficients.

2.3.3.0 MAIN CONTENTS

2.3.3.1 The Multiple Regression Coefficients Derivation

In the simple regression case, the values of the regression coefficients were chosen to make the fit as good as possible in the hope of obtaining most satisfactory estimates of
the true unknown parameters. Our earlier stated definition of goodness of fit; is the minimization of $RSS$, which is the sum of squares of the residuals:

$$RSS = \sum_{i=1}^{n} e_i^2 \quad \text{...[2.50]}$$

Where $e_i$ is again, the residual in observation $i$, the difference between the actual value $Y_i$ in that observation and the value $\hat{Y}_i$ predicted by the regression equation:

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + e_i \quad \text{...[2.51]}$$

$$e_i = Y_i - \hat{Y}_i - b_2 X_{2i} - b_3 X_{3i} \quad \text{...[2.52]}$$

It could be observed that the $X$ variables now have two subscripts. The first identifies the $X$ variable and the second identifies the observation.

Applying [2.52] into [2.50];

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \quad \text{...[2.53]}$$

From first-order conditions for a minimum;

$$\frac{\partial RSS}{\partial b_1} = 0, \quad \frac{\partial RSS}{\partial b_2} = 0 \quad \text{and} \quad \frac{\partial RSS}{\partial b_3} = 0$$

[2.52] will give the following equations:

$$\frac{\partial RSS}{\partial b_1} = -2 \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \quad \text{...[2.54]}$$

$$\frac{\partial RSS}{\partial b_2} = -2 \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \quad \text{...[2.55]}$$

$$\frac{\partial RSS}{\partial b_3} = -2 \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \quad \text{...[2.56]}$$

Resulting in three equations from the three unknowns, $b_1$, $b_2$, and $b_3$.

The first can easily be rearranged to express $b_1$, regarding $b_2$, $b_3$, and the data on $Y$, $X_2$, and $X_3$:

$$b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3 \quad \text{...[2.57]}$$
From [3.57] and working through [3.55] to [3.56], the following expression for \( b_2 \) is obtained:

\[
b_2 = \frac{Cov(X_2, Y)Var(X_3) - Cov(X_3, Y)Cov(X_3, X_2)}{Var(X_2)Var(X_3) - [Cov(X_3, X_2)]^2}
\]

\[
= \frac{\left[ \sum (X_2, Y)(X_3) \right] - \left[ \sum (X_3, Y)(X_3, X_2) \right]}{\left[ \sum (X_2)\sum (X_3) - \sum [(X_3, X_2)]^2 \right]}
\]

...[2.58]

Similarly, the expression of \( b_3 \) can be obtained by switching \( X_2 \) and \( X_1 \) in [2.58].

Clearly, the principles behind the derivation of the regression coefficients have been shown to be the same for multiple regression as that of the simple regression. But, it should also be observed that the expressions are however different and so should not try to use expressions derived for simple regression in a multiple regression situation.

A generalized framework for the multiple regression model is

\[
Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \mu_i
\]

...[2.59]

We may write [2.59] for three variables as,

\[
Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i
\]

...[2.60]

where \( Y \) is the dependent variable, \( X_2 \) and \( X_k \) (kth term) the regressors, \( \mu \) the stochastic disturbance term and \( i \) the \( i \)th (\( t \)th, if in time series) observation. Also \( \beta_1, \beta_2 \) and \( \beta_k \) are the partial regression coefficients but \( \beta \) is the intercept term which gives the mean effect on \( Y \) of all the variables excluded from the model. That is, in the case of [2.50], when \( X_2 \) and \( X_k \) are set equal to zero.

Zero mean value of \( \mu \) in [2.60] is;

\[
E(\mu \mid X_{2i}, X_{3i}) = 0
\]

...[2.61]

2.3.3.1 Multiple Regression Coefficients Interpretation

Discriminate between the effects of the explanatory variables and making allowance for the fact that they may be correlated is enabled in multiple regression analysis. The regression coefficient of each \( X \) variable provides an estimate of its influence on \( Y \).
There are two ways in which this can be demonstrated. First is the case where there are only two explanatory variables; to demonstrate that the estimators are unbiased if the model is correctly specified and the Gauss–Markov conditions are fulfilled.

The second method is to run a simple regression of $Y$ on one of the $X$ variables, having first purged both $Y$ and the $X$ variable of the components that could be accounted for by the other explanatory variables. The estimate of the slope coefficient and its standard error thus obtained are the same as in the multiple regression. It follows that a scatter diagram plotting the purged $Y$ against the purged $X$ variable will provide a valid graphical representation of their relationship that can be obtained in no other way.

\[ \text{EARNINGS} = \beta_1 + \beta_2S + \beta_3\text{ASVABC} + u \quad \text{...[2.62]} \]

If the graphical illustration is particularly interested in, in the relationship between earnings and schooling; a direct plot of EARNINGS on $S$ would give a distorted view of the relationship. This is because ASVABC is positively correlated with $S$ and having some consequences as $S$ increases. These are [1] EARNINGS will likely increase, because $\beta_2$ is positive; [2] ASVABC will tend to increase, because $S$ and ASVABC are positively correlated; and [3] EARNINGS will receive a lift due to the increase in ASVABC and the fact that $\beta_3$ is positive. That is, the variations in EARNINGS will overstate the apparent influence of $S$ because in part they will be due to associated variations in ASVABC. And the outcome of this is that in a simple regression the estimator of $\beta_2$ will be biased. The graphical illustration is shown in Figure 3.1.
Concerning simple regression analysis, the regression coefficients should be thought of as different categories of random variables whose random components are related to the existence of the disturbance term in the model. Each regression coefficient is calculated as a function of the values of $Y$ and the explanatory variables in the sample. $Y$, in turn, is determined by the explanatory variables and the disturbance term. It follows that the regression coefficients are indeed determined by the values of the explanatory variables and the disturbance term, in which their properties depend critically upon the properties of the disturbance term.

In continuation of the assumption that the Gauss–Markov conditions are satisfied, which are:

(i) that the expected value of $u$ in any observation is 0
(ii) that the population variance of its distribution is the same for all observations
(iii) that the population covariance of its values in any two observations is 0, and
(iv) that it is distributed independently of any explanatory variable.

The first three conditions are the same as for simple regression analysis but (iv) is a generalization of (i) to (iii).

Furthermore, there are two practical requirements to be met.
(i) There must be enough data to fit the regression line. That is, there must be at least as many (independent) observations as there are parameters to be estimated.
(ii) There must not be an exact linear relationship among the explanatory variables.

2.3.3.3 $t$ Tests and Confidence Intervals

The $t$ tests on the regression coefficients are performed in the same way as for simple regression analysis. Particular attention should, however, be taken when looking up the critical level of $t$ at any given significance level. It depends on the number of degrees of freedom $(n - k)$; the number of observations $n$ minus the number of parameters estimated $k$.

The confidence intervals are also obtained in the same manner as in simple regression analysis and equally based on the number of degrees of freedom $(n - k)$.

2.3.3.4 Consistency

Once the fourth Gauss–Markov condition is satisfied, OLS yields consistent estimates in the multiple regression models, as is the case in the simple regression model. One condition for consistency is that when $n$ becomes large, the population variance of the estimator of each regression coefficient tends to 0, and the distribution falls to a spike. The other condition for consistency is since the estimator is unbiased, the spike would be located at the true value.

2.3.4.0 MULTICOLLINEARITY

In most situations, the available data for use in multiple regression analysis would not provide significant solutions to problems at hand. The reason being that the standard errors are very high, or the $t$ test ratios are very low. Which means the confidence intervals for such parameters are very wide. A situation of this nature occurs when the explanatory variables show little variation and high intercorrelations. Multicollinearity is the aspect of the situation where the explanatory variables are highly intercorrelated.
Let’s look at multicollinearity in a model with two explanatory variables. It would be observed that the higher the correlation between the explanatory variables, the larger the population variances of the distributions of their coefficients and the greater the possibility of attaining irregular estimates of the coefficients.

You should, however, bear in mind that a high correlation does not necessarily lead to poor estimates. If all the other elements determining the variances of the regression coefficients are properly in the number of observations and the sample variances of the explanatory variables are large and the variance of the disturbance term small, good estimates could still be obtained. Multicollinearity, therefore, must be caused by a mixture of a high correlation and one or more of the other elements being inappropriate. This is a matter of degree and not kind of element of which any regression will suffer from it to some extent unless all the explanatory variables are uncorrelated. But the consequence is only taken into consideration when it is obviously going to have a serious effect on the regression results.

It is a common problem in time series regressions, particularly where the data consists of a series of observations on the variables over a number of time periods. Which may give rise to multicollinearity if two or more of the explanatory variables are highly correlated in a strong time trend.

Using Table 3.1 as an example let’s consider first the case of exact multicollinearity where the explanatory variables are perfectly correlated.

Table 3.1

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOUN
Let \([2.40]\) be the true relationship, that is;

\[ Y = 2 + 3X_2 + X_3 + u \quad \ldots[2.63] \]

Suppose that there is a linear relationship between \(X_2\) and \(X_3\):

\[ X_3 = 2X_2 - 1 \quad \ldots[2.64] \]

and suppose that \(X_2\) increases by one unit in each observation. \(X_3\) will increase by two units, and \(Y\) by approximately five units as indicated in Table 3.1. Applying the linear relationship between \(X_2\) and \(X_3\) in manipulating [2.40] will result in different conclusions for \(Y\).

In such a situation it is impossible for regression analysis, or any other technique for that matter, to distinguish between these possibilities. You would not even be able to calculate the regression coefficients because both the numerator and the denominator of the regression coefficients would collapse to 0. This will be demonstrated with the general two-variable case. Suppose

\[ Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + u \quad \ldots[2.65] \]

And

\[ X_3 = \lambda + \mu X_2 \quad \ldots[2.66] \]

Substituting for \(X_3\) in [3.58] gives

\[
\frac{\text{Cov}(X_2,Y)\text{Var}(\lambda + \mu X_2) - \text{Cov}([\lambda + \mu X_2],Y)\text{Cov}(X_2,[\lambda + \mu X_2])}{\text{Var}(X_2)\text{Var}(\lambda + \mu X_2) - [\text{Cov}(X_2,[\lambda + \mu X_2])]^2} = \frac{\text{Cov}(X_2,Y)\text{Var}(\lambda+\mu X_2)-\text{Cov}(\mu X_2,Y)\text{Cov}(X_2,\mu X_2)}{\text{Var}(X_2)\text{Var}(\mu X_2)-[\text{Cov}(X_2,\mu X_2)]^2} \ldots[2.67]
\]

From Variance Rule 4, the additive \(\lambda\) in the variances can be dropped. A similar rule could be developed for covariances, since an additive \(\lambda\) does not affect them either.
Therefore,

\[ b_2 = \frac{\text{Cov}(X_2, Y)u^2\text{Var}(X_2) - u\text{Cov}(\mu X_2, Y)u\text{Cov}(X_2, \mu X_2)}{\text{Var}(X_2)u^2\text{Var}(X_2) - [u\text{Cov}(X_2, \mu X_2)]^2} \]  

\[ = \frac{u^2\text{Cov}(X_2, Y)\text{Var}(X_2) - u^2\text{Cov}(X_2, Y)\text{Var}(X_2)}{u^2\text{Var}(X_2)\text{Var}(X_2) - [u\text{Var}(X_2)]^2} = \frac{0}{0} \]  

...[2.68]

Which is unusual for there to be an exact relationship among the explanatory variables in a regression. So, when this occurs, it is typical because there is a logical error in the specification.

2.3.4.1 Multicollinearity in Models with More Than Two Explanatory Variables

The previous discussion of multicollinearity was restricted to the case where there are two explanatory variables. In models with a greater number of explanatory variables, multicollinearity may be caused by an approximately linear relationship among them. It may be difficult to discriminate between the effects of one variable and those of a linear combination of the remainder. In the model with two explanatory variables, an approximately linear relationship automatically means a high correlation, but when there are three or more, this is not necessarily the case. A linear relationship does not inevitably imply high pairwise correlations between any of the variables. The effects of multicollinearity are the same as in the case with two explanatory variables and as in that case, the problem may not be serious if the population variance of the disturbance term is small, the number of observations large and the variances of the explanatory variables are equally large.

2.3.4.2 Ways to alleviate multicollinearity problems

Two categories exist to alleviate multicollinearity problems:

i. The direct attempts to improve the four conditions responsible for the reliability of the regression estimates, and
The indirect methods.

First, you may try to reduce $\sigma_u^2$. The disturbance term is the joint effect of all the variables influencing $Y$ that you have not included explicitly in the regression equation. If you can think of an important variable that you have omitted, and is therefore contributing to $u$, you will reduce the population variance of the disturbance term if you add it to the regression equation.

Second, consider $n$, the number of observations. If you are working with cross-section data (individuals, households, enterprises, etc.) and you are undertaking a survey, you could increase the size of the sample by negotiating a bigger budget. Alternatively, you could make a fixed budget go further by using a technique known as clustering.

A further way of dealing with the problem of multicollinearity is to use minor information, if available, concerning the coefficient of one of the variables.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u \quad \quad \quad \ldots[2.70]$$

For example, suppose that $Y$ in equation is the aggregate demand for a category of consumer expenditure, $X$ is aggregate disposable personal income, and $P$ is a price index for the category. To fit a model of this type, you would use time series data. If $X$ and $P$ possess strong time trends and are therefore highly correlated, which is often the case with time series variables, multicollinearity is likely to be a problem. Suppose, however, that you also have cross-section data on $Y$ and $X$ derived from a separate household survey. These variables will be denoted $Y'$ and $X'$ to indicate that the data are household data, not aggregate data. Assuming that all the households in the survey were paying roughly the same price for the commodity, one would fit the simple regression

$$\hat{Y} = b_1 + b_2 X \quad \quad \quad \ldots[2.71]$$

Now substitute $b_2$ for $\beta_2$ in the time series model

$$Y = \beta_1 + b_2 X + \beta_3 P + u \quad \quad \quad \ldots[2.72]$$

Subtract $b_2 X$ from both sides,
\[ Y - b_2X = \beta_1 + \beta_3P + u \quad \text{...[2.73]} \]

And regress \( Z = Y - b_2X \) on price. This is a simple regression, so multicollinearity has been eliminated.

There are, however, two possible problems with this technique. First, the estimate of \( \beta_3 \) depends on the accuracy of the estimate of \( b_2' \), and this of course is subject to sampling error. Second, you are assuming that the income coefficient has the same meaning in time series and cross-section contexts, and this may not be the case.

For many commodities, the short-run and long-run effects of changes in income may differ because expenditure patterns are subject to inertia. A change in income can affect expenditure both directly, by altering the budget constraint, and indirectly, through causing a change in lifestyle, and the indirect effect is much slower than the direct one. As a first approximation, it is commonly argued that time series regressions, particularly those using short sample periods, estimate short-run effects while cross-section regressions estimate long-run ones.

For the indirect methods to alleviate multicollinearity problems. If the correlated variables are similar conceptually, it may be reasonable to combine them into some overall index.

2.3.7.0 SUMMARY

In this unit, we discussed the multiple regression model, a model in which there is more than one descriptive variable but a direct extension of the simple regression model having two new dimensions. We introduced the arrangement of flow for the multiple regression analysis and started with the derivation of formula by showing the linkage of the simple regression model and the multiple regression. However, we briefly discussed on interpretation and properties of multiple regression coefficients. For more understanding, the student may use the reference materials to look up estimation procedures using values and presentation of results, aspects of arrangement of flow for the multiple regression analysis not discussed. Finally, the concept of multicollinearity as an existence of linear relationship in the midst of regressors was equally discussed in this unit. Students are shown two ways to alleviate multicollinearity being a problem associated with CLRM.
First, when evaluating the influence of a given descriptive variable on the dependent variable, we would now have to face the problem of discriminating between its effects and the effects of the other descriptive variables. Second, we shall have to tackle the problem of model specification. Often some variables might be thought to influence the behaviour of the dependent variable; though, they might be unconnected. We shall have to decide which should be included in the regression equation and which should be omitted. However, the arrangement of flow for the multiple regression analysis is to firstly, carry out derivation of formula, then estimation procedures using values, followed by presentation of results and lastly interpretations.

### 2.3.6.0 CONCLUSION

The features of multiple regression analyses and multicollinearity introduced in this unit are extension of unit 2. Here, we pointed out some of the complications arising from the introduction of several descriptive variables. In the discussions, we explained that when we go beyond the two-variable model and consider multiple regression models we add the assumption that there is no perfect multicollinearity (assumption 10 of CLRM). That is, there are no perfect linear relationships among the descriptive variables when two or more of these variables move together and difficult to determine their separate influences.

### 2.3.7.0 TUTOR-MARKED ASSIGNMENT

1.) The following earnings functions were fitted separately for males and females (standard errors in parentheses):

**Males**

\[
\text{EARNINGS} = -3.6121 + 0.7499S + 0.1558ASVABC
\]

<table>
<thead>
<tr>
<th>S.E.E</th>
<th>(2.8420)</th>
<th>(0.2434)</th>
<th>(0.0600)</th>
</tr>
</thead>
</table>

**Females**

\[
\text{EARNINGS} = -5.9010 + 0.8803S + 0.1088ASVABC
\]

<table>
<thead>
<tr>
<th>S.E.E</th>
<th>(2.6315)</th>
<th>(0.1910)</th>
<th>(0.0577)</th>
</tr>
</thead>
</table>

3.) Explain why the standard errors of the coefficients of \( S \) and \( ASVABC \) are greater for the male subsample than for the female subsample, and why the difference in the standard errors are relatively large for \( S \).
2.3.8.0 REFERENCES /FURTHER READING


UNIT 4: TRANSFORMATIONS OF VARIABLES

CONTENTS
2.4.1.0 Introduction
2.4.2.0 Objectives
2.4.3.0 Main Content
2.4.4.0 Summary
2.4.5.0 Conclusion
2.4.6.0 References/Further Reading

2.4.1.0 INTRODUCTION

In model transformation, the functional form of an equation or model determines the estimation techniques and interpretation of results obtained from it. Transforming a variable involves using mathematical procedure to modify its measured values. Single equation (or any other form of equation) may be in different forms. There are two kinds of transformations and generally, models can be of the form;

i. Linear transformation; this preserves the linear relationships between variables (parameters and variables are linear). That is the correlation between $x$ and $y$ (say) would be unchanged after a linear transformation.

Examples of a linear transformation to variable $x$ would be multiplying $x$ by a constant, dividing $x$ by a constant, or adding a constant to $x$.

ii. Nonlinear transformation; A nonlinear transformation changes (increases or decreases) linear relationships between variables and, thus, changes the correlation between variables.

Examples of a nonlinear transformation of variable $x$ would be taking the square root of $x$ or the reciprocal of $x$. By extension, nonlinear transformation is a non-linear model that can be made linear. For example;

$$ Y_t = AX_t^\beta e^u $$ is an example of production function that can be made linear by taking logarithms, that is; $ \ln Y_t = \ln A + \beta \ln X_t + u_t $.
In regression, however, a transformation to achieve linearity is a special kind of nonlinear transformation. It is a nonlinear transformation that increases the linear relationship between two variables.

2.4.2.0 OBJECTIVE
The main objective of this unit is to show that regression analysis can be extended to fit nonlinear models through transformation of nonlinear model that can be made linear.

2.4.3.0 MAIN CONTENT
A limitation out of other limitations of linear regression analysis is that it is contained in its very name, in that it can be used to fit only linear equations where every explanatory term, except the constant, is written in the form of a coefficient multiplied by variable:

\[ y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \]  

...[2.74]

Y equations such as the two below are non-linear

\[ Y = \beta + \beta_2 \frac{1}{X} \]  

...[2.75]

And

\[ Y = \beta_1 X^{\beta_2} \]  

...[2.76]

Nevertheless, both [2.75] and [2.76] have been suggested as suitable forms for Engel curves, (the relationship between the demand for a particular commodity, Y and income, X). As an illustration, given data on Y and X, how could one estimate the parameters \( \beta_1 \) and \( \beta_2 \) in these equations? Actually, in both cases, with a little preparation one can actually use linear regression analysis.

Here, first, note that [2.74] is linear in two ways. The right side is linear in variables because the variables are included exactly as defined, rather than as functions. It, therefore, consists of a weighted sum of the variables, the parameters being the weights. The right side is also linear in the parameters since it consists of a weighted sum of these as well, the X variables being the weights in this respect.
For the purpose of linear regression analysis, only the second type of linearity is important.
Nonlinearity in the variables can always be sidestepped by using appropriate definitions.

For example, suppose that the relationship was of the form

\[ Y = \beta_1 + \beta_2 X_2^2 + \beta_3 \sqrt{X_3} + \beta_4 \log X_4 + \ldots \] \[ \text{...}[2.77] \]

By defining \( Z_2 = X_2^2, Z_3 = \sqrt{X_3}, Z_4 = \log X_4 \) etc, the relationship can be rewritten

\[ Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \ldots \] \[ \text{...}[2.78] \]

and it is now linear in variables as well as in parameters. This type of transformation is only beautifying, and you will usually see the regression equation presented with the variables written in their nonlinear form. This avoids the need for explanation and extra notation.

But \[2.76\] is nonlinear in both parameters and variables and cannot be handled by a mere redefinition. That is, even if attempted, the equation cannot be made linear by defining \( Z = X^{\beta_2} \) and replacing \( X^{\beta_2} \) with \( Z \); since you do not know \( \beta_2 \), you have no way of calculating sample data for \( Z \).

However, you could define \( Z = \frac{1}{X} \), the equation now becomes

\[ Y = \beta_1 + \beta_2 Z \] \[ \text{...}[2.79] \]

and this is linear, which is the regress of \( Y \) on \( Z \). The constant term in the regression will be an estimate of \( \beta_1 \) and the coefficient of \( Z \) will be an estimate of \( \beta_2 \).

2.4.4.0 SUMMARY

This unit discussed transformation of variables. Two kinds of transformations were discussed, the linear transformation in which the linear relationships between the parameters and variables are preserved after transformation. As well as the nonlinear
transformation in which there is increase or decrease in the linear relationship of the variables involved.

2.4.5.0 CONCLUSION

In this unit the concept of transformation of variables is discussed to show that regression analysis can be extended to fit nonlinear models through transformation of non-linear models. That nonlinearity in the variables can always be sidestepped by using appropriate definitions. Example of these definitions is taking logarithm of the nonlinear model and application of the least squares principle when the model cannot be linearised.

2.4.6.0 REFERENCES /FURTHER READING


UNIT 5: DUMMY VARIABLES
2.5.1.0 INTRODUCTION

It sometimes happens that some descriptive variables do exist in our regression equation, and/or the factors that you would like to introduce into a regression model are qualitative (racial, sex or age differences) in nature and therefore not measurable in numerical terms. In such circumstances, dummy variables are utilised.

2.5.2.0 OBJECTIVE

The main objective of this unit is to provide basic understanding of the topic ‘Dummy Variable’ through the use of imitation variables existing or being introduced into a regression equation to solve some variables that are qualitative or immeasurable in numerical terms.

2.5.3.0 MAIN CONTENTS

The inherent assumption for the application of dummy variables is that the regression lines for the different groups differ only in the intercept term but have the same slope coefficients. For example; (1). You are investigating the relationship between schooling $x$ and earnings $y$, and you have both males and females in your sample. You would like to see if the sex of the respondent makes a difference.
(2). You are investigating the relationship between income and expenditure in Cameroon, and your sample includes both English-speaking and French-speaking households. You would like to find out whether the ethnic difference is relevant.

(3). You have data on the growth rate of GDP per capita and foreign aid per capital for a sample of developing countries, of which some are democracies and some are not. You would like to investigate whether the impact of foreign aid on growth is affected by the type of government.

A solution to these examples would be to run separate regressions for the two categories and see if the coefficients are different. Alternatively, you could run a single regression using all the observations together, measuring the effect of the qualitative factor with what is known as a dummy variable. This effect has the two important advantages of providing a simple way of testing whether the effect of the qualitative factor is significant.

The qualitative variable has four categories, and we need to develop a more elaborate set of dummy variables. The standard procedure is to choose one category as the reference category to which the basic equation applies, and then to define dummy variables for each of the other categories. In general, it is good practice to select the dominant or most normal category, if there is one, as the reference category.

Accordingly, we will define dummy variables for the other three types. TECH will be the dummy variable for the technical schools: TECH is equal to 1 if the observation relates to a technical school, 0 otherwise. Similarly, we will define dummy variables WORKER and VOC for the skilled workers’ schools and the vocational schools. The regression model is now

\[ \text{COST} = \beta_1 + \delta_T \text{TECH} + \delta_W \text{WORKER} + \delta_V \text{VOC} + \beta_2 N + u \]  

[2.80]

Where \(\delta_T, \delta_W, \text{and} \delta_V\) are coefficients that represent the extra overhead costs of the technical, skilled workers’, and vocational schools, relative to the cost of a general school. Note that you do not include a dummy variable for the reference category, and that is the reason that the reference category is usually described as the omitted
category. Note that we do not make any prior assumption about the size, or even the
sign, of the $\delta$ coefficients.

2.5.3.1 The Dummy Variable Trap

What would happen if you included a dummy variable for the reference category? There would be two consequences.

$i.$ Were it is possible to compute regression coefficients, you would not be able to give them an interpretation. The coefficient $b_1$ is a basic estimate of the intercept, and the coefficients of the dummies are the estimates of the increase in the intercept from this basic level, but now there is no definition of what is basic, so the interpretation collapses.

$ii.$ The other consequence is that the numerical procedure for calculating the regression coefficients will break down, and the computer will simply send you an error message (or possibly, in sophisticated applications, drop one of the dummies for you). Suppose that there are $m$ dummy categories, and you define dummy variables $D1...Dm$.

Then, in observation $i$, $\sum_{j=1}^{m} D_{ji} = 1$ because one of the dummy variables will be equal to 1 and all the others will be equal to 0. But the intercept $\beta_1$ is really the product of the parameter $\beta_1$ and a special variable whose value is 1 in all observations. Hence, for all observations, the sum of the dummy variables is equal to this special variable, and one has an exact linear relationship among the variables in the regression model. As a consequence the model is subject to a special case of exact multicollinearity, making it impossible to compute regression coefficients.

2.5.3.2 Change of Reference Category

The skilled workers' schools are considerably less academic than the others, even the technical schools. Suppose that we wish to investigate whether their costs are significantly different from the others. The easiest way to do this is to make them the omitted category (reference category). Then the coefficients of the dummy variables become estimates of the differences between the overhead costs of the other types of school and those of the skilled workers' schools. Since skilled workers' schools are
now the reference category, we need a dummy variable, which will be called $GEN$, for the general academic schools. The model becomes

$$COST = \beta_1 + \delta{TECH} + \delta{VOC} + \delta{GEN} + \beta_2N + u \quad \text{...}[2.81]$$

where $\delta{T}, \delta{V}, \text{and } \delta{G}$ are the extra costs of technical, vocational, and general schools relative to skilled workers’ schools.

### 2.5.3.3 Slope Dummy Variables

We have so far assumed that the qualitative variables we have introduced into the regression model are responsible only for shifts in the intercept of the regression line. We have implicitly assumed that the slope of the regression line is the same for each category of the qualitative variables.

This is not necessarily a plausible assumption, and we will now see how to relax it, and test it, using the device known as a slope dummy variable (also sometimes known as an interactive dummy variable).

The assumption that the marginal cost per student is the same for occupational and regular schools is unrealistic. Because occupational schools incur expenditure on training materials related to the number of students, and the staff-student ratio has to be higher in occupational schools because workshop groups cannot be, or at least should not be, as large as academic classes. We can relax the assumption by introducing the slope dummy variable, $NOCC$, defined as the product of $N$ and $OCC$:

$$COST = \beta_1 + \delta{OCC} + \beta_2N + \lambda{NOCC} + u \quad \text{...}[2.82]$$

If this is rewritten

$$COST = \beta_1 + \delta{OCC} + (\beta_2 + \lambda{OCC})N + u, \quad \text{...}[2.83]$$

it can be seen that the effect of the slope dummy variable is to allow the coefficient of $N$ for occupational schools to be $\lambda$ greater than that for regular schools. If $OCC$ is 0, so is $NOCC$ and the equation becomes

$$COST = \beta_1 + \beta_2N + u \quad \text{...}[2.84]$$

If $OCC$ is 1, $NOCC$ is equal to $N$ and the equation becomes
\[ \text{COST} = \beta_1 + \delta + (\beta_2 + \lambda)N + u \] \quad \ldots [2.85]

\( \lambda \) is thus the incremental marginal cost associated with occupational schools, in the same way that \( \lambda \) is the incremental overhead cost associated with them.

2.5.5.0 SUMMARY

In this unit, the essentials and applications of the concept of dummy variable estimation was introduced. By way of illustrative example to the students, an investigation is made which showed that in the application of dummy variables the regression lines for the different groups differ only in the intercept term but have the same slope coefficients. We also discussed two traps that may occur in the application of dummy variable. First, change in reference category, in which a reference variable is created out of the variables in consideration for analysis. Second, the slope dummy variables; which considered the slope of the regression line not being the same for each category of the qualitative variables.

2.5.4.0 CONCLUSION

This unit discussed the use of dummy variable estimation to solve some variables that are existing or being introduced into a regression equation which are qualitative or immeasurable in numerical terms. In this discussion, students are made aware that the inherent assumption for the application of dummy variables is that the regression lines for the different groups differ only in the intercept term but have the same slope coefficients.

2.5.6.0 REFERENCES /FURTHER READING


UNIT 6: SPECIFICATION OF REGRESSION VARIABLES: A PRELIMINARY SKIRMISH

CONTENTS
2.6.1.0 Introduction
2.6.2.0 Objectives
2.6.3.0 Main Content
   2.6.3.1 Model Specification of Regression Variables
2.6.4.0 Summary
2.6.5.0 Conclusion
2.6.6.0 References/Further Reading

2.6.1.0 INTRODUCTION

The construction of an economic model involves the specification of the relationships that constitute it, the specification of the variables that participate in each relationship and the mathematical function representing each relationship.

2.6.2.0 OBJECTIVE

The main objective of this unit is to provide a general understanding of the topic ‘Specification of Regression Variable’. This include the creating an opportunity for the students to know that model specification denotes the determination of which independent variables may be included in or omitted from a regression equation.

2.6.3.0 MAIN CONTENTS

2.6.3.1 Model Specification

The knowledge of exactly which descriptive variables ought to be included in the equation helps when we undertake regression analysis, our task would equally be limited to calculating estimates of their coefficients, confidence intervals for these estimates and so on. In practice, however, we can never be sure that we have specified the equation properly. Economic theory ought to provide a guide, but the theory is never flawless. Unaware, we might be including some variables that ought not to be in the model and we might be leaving out others that ought to be incorporated.
Existing properties of the regression estimates of the coefficients depend significantly on the validity of the specification of the model. The consequences of misspecification of the variables in a relationship are stated below.

1. When a variable that ought to be included is left out, the regression estimates are in general (but not always) biased. The standard errors of the coefficients and the corresponding \( t \) tests are in general invalid. Another serious consequence of omitting a variable that ought to be included in the regression is that the standard errors of the coefficients and the test statistics are in general invalidated. This means of course that you are not in principle able to test any hypotheses with your regression results.

2. On the other hand, if you include a variable that ought not to be in the equation, the regression coefficients are in general (but not always) inefficient but not biased. The standard errors are in general valid but, because the regression estimation is inefficient, they will be needlessly large.

2.6.4.0 SUMMARY

In this unit, the specification of regression variables at a preliminary skirmish is discussed. In general, an introductory opportunity is created for the students to know that the specification of a regression model should be based primarily on theoretical considerations rather than what may be obtainable in practice.

2.6.5.0 CONCLUSION

The specification of regression variables at a preliminary skirmish was explained in this unit. And briefly made the students aware that this is one of the foundational econometrics topic that prepares the readers for intermediate econometrics. That is, specification of model is the first and most critical of the stages in regression analysis which students may use to identify functions and solve problems associated with all other topics discussed in this module.

2.6.6.0 REFERENCES /FURTHER READING


MODULE 3: HETEROSCEDASTICITY

CONTENTS
3.1.1.0 Introduction
3.1.2.0 Objectives
3.1.3.0 Main Content
3.1.3.1 Heteroscedasticity and Its Effects
3.1.3.2 Likely Sources of Heteroscedasticity
3.1.3.3 Detection of Heteroscedasticity
3.1.3.4 The Spearman Rank Correlation Test
3.1.3.5 The Goldfeld–Quandt Test
3.1.3.6 The Glejser Test
3.1.3.6 Solution to Heteroscedasticity
3.1.3.7 Consequences of Heteroscedasticity
3.1.4.0 Summary
3.1.5.0 Conclusion
3.1.6.0 Tutor-Marked Assignment
3.1.7.0 References/Further Reading

3.1.1.0 INTRODUCTION

The general aim of this module is to provide you with a thorough understanding of the violation of one of the classical assumptions, equal variances (homoscedastic). The properties of the estimators of the regression coefficients depend on the properties of the disturbance term in the regression model. In this module, we shall be looking at some of the problems that arise when violations of the Gauss–Markov conditions, the assumptions relating to the disturbance term, are not satisfied. Basic understanding of heteroscedasticity (unequal-variances) will be likewise explained.

3.1.2.0 OBJECTIVE

The main objective of this unit is to provide a platform for the students to understand that in statistic, heteroscedasticity is a collection of random variables and the absence of it is homoscedasticity.
3.1.3.0 MAIN CONTENTS

3.1.3.1 Heteroscedasticity and Its Effects

Gauss–Markov second conditions listed in the previous module states; that the variance of the disturbance term in each observation should be constant. This sounds peculiar and needs a bit of explanation. The disturbance term in each observation has only one value, so what can be meant by its "variance"?

The focus point of discussion here is, its potential behaviour before the sample is generated. So when the model is written as:

\[ Y = \beta_1 + \beta_2 X + u \] ...

Figure 1.1 Homoscedasticity

[3.01] has in it the first two Gauss–Markov conditions stating that the disturbance terms \( u_1, \ldots, u_n \) in the \( n \) observations are drawn from probability distributions that have 0 mean and the same variance. Their actual values in the sample will sometimes be positive, sometimes negative, sometimes relatively far from 0, sometimes relatively close, but there will be no a priori reason to anticipate a particularly erratic value in any given observation. To put it another way, the probability of \( u \) reaching a given positive or negative value will be the same in all observations. This condition is known as homoscedasticity, which means "same dispersion".
Figure 1.1 is a depiction of homoscedasticity. For a simple illustration, the sample in Figure 1.1 contains only five observations. Let us start with the first observation, where $X$ has the value $X_1$. If there were no disturbance term in the model, the observation would be represented by the circle vertically above $X_1$ on the line $Y = \beta_1 + \beta_2 X$. The effect of the disturbance term is to shift the observation upwards or downwards vertically. The potential distribution of the disturbance term, before the observation has been generated, is shown by the normal distribution centered on the circle. The actual value of the disturbance term for this observation turned out to be negative, the observation being represented by the darkened indicator. The potential distribution of the disturbance term, and the actual outcome, are shown in a similar way for the other four observations. Although homoscedasticity is often taken for granted in regression analysis, in some contexts it may be more reasonable to suppose that the potential distribution of the disturbance term is different for different observations in the sample. This is illustrated in Figure 1.2 where the variance of the potential distribution of the disturbance term is increasing as $X$ increases. This does not mean that the disturbance term will necessarily have a particularly large (positive or negative) value in an observation where $X$ is large, but it does mean that the a priori probability of having an erratic value will be relatively high. This is an example of heteroscedasticity, which means "differing dispersion".

Mathematically, homoscedasticity and heteroscedasticity may be defined:

Homoscedasticity: $\sigma^2_{ui} = \sigma^2_u$ same for all observations
Heteroscedasticity: $\sigma^2_{ui}$ not the same for all observations
Figure 1.2 Heteroscedasticity

Figure 1.3 Model with a heteroscedastic disturbance term

Figure 1.3 shows how a typical scatter diagram would look if \( Y \) were an increasing function of \( X \) and the heteroscedasticity were of the type shown in Figure 1.2. It could be seen that, although the observations are not necessarily further away from the non-stochastic component of the relationship, represented by the line \( Y = \beta_1 + \beta_2X \), there is a tendency for their dispersion to increase as \( X \) increases. Thus this particular Gauss–Markov condition does not seem to have been used anywhere in the analysis so
far, so it might look almost irrelevant. In particular, the proofs of the unbiasedness of the OLS regression coefficients did not use this condition. There are however two explanations for the presence of heteroscedasticity.

The first explanation has to do with making the variances of the regression coefficients as small as possible, so that in a probabilistic sense, maximum precision is achieved. If there is no heteroscedasticity and if the other Gauss–Markov conditions are satisfied, the OLS regression coefficients have the lowest variances of all the unbiased estimators that are linear functions of the observations of $Y$. If heteroscedasticity is present, the OLS estimators are inefficient because there are still other estimators that have smaller variances and are still unbiased.

The other reason is that the estimators of the standard errors of the regression coefficients will be wrong. This is because their computation is based on the assumption that the distribution of the disturbance term is homoscedastic. Otherwise, they are biased. As a consequence, the $t$-tests and also the usual $F$-tests will be invalid. It is therefore quite likely that the standard errors will be underestimated, so the $t$-statistics will be overestimated which will have a misleading impression of the precision of the regression coefficients. The coefficient may appear significantly different from 0, at a given significance level, when in fact, it is not. The inefficiency property can be explained quite easily assuming that heteroscedasticity of the type displayed in Figures 1.2 and 1.3 is present.

Which is an observation where the potential distribution of the disturbance term has a small standard deviation, similar to that of Figure 1.1.

### 3.1.3.2 Likely Sources of Heteroscedasticity

For heteroscedasticity, it is likely to be a problem when the values of the variables in the sample vary substantially in different observations. Given that $Y = \beta_1 + \beta_2X + u$, the variations in the omitted variables and the measurement errors that are jointly responsible for the disturbance term ($u$) would be somewhat small when $Y$ and $X$ are small and large when they are large. This is simply because economic variables in such a true relationship tend to move in size together.

### 3.1.3.3 Detection of Heteroscedasticity
There seems to be no limit to the different possible types of heteroscedasticity, and consequently, a large number of different tests appropriate for different conditions have been suggested. The attention here would, however, be focused on three tests that hypothesize a relationship between the variance of the disturbance term and the size of the explanatory variable(s). These would be the Spearman rank correlation, Goldfeld–Quandt, and Glejser tests.

3.1.3.4 The Spearman Rank Correlation Test

This test assumes that the variance of the disturbance term is either increasing or decreasing as \( X \) increases and that there will be a correlation between the absolute size of the residuals and the size of \( X \) in an OLS regression. The data on \( X \) and the absolute values of the residuals are both ranked, and the rank correlation coefficient is defined as

\[
r_{x,e} = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \quad \ldots[3.02]\]

where \( D_i \) is the difference between the rank of \( X \) and the rank of \( e \) in observation \( i \). Under the assumption that the population correlation coefficient is 0, the rank correlation coefficient has a normal distribution with 0 mean and variance \( \frac{1}{n-1} \) in large samples. The appropriate test statistic is therefore \( r_{x,e} \sqrt{n-1} \) and the null hypothesis of homoscedasticity will be rejected at the 5 percent level if its absolute value is greater than 1.96 and at the 1 percent level if its absolute value is greater than 2.58, using two-tailed tests. If there is more than one explanatory variable in the model, the test may be performed with any one of them.
Example

Table 1.1

<table>
<thead>
<tr>
<th>Country</th>
<th>MANU</th>
<th>GDP</th>
<th>POP</th>
<th>MANU/POP</th>
<th>GDP/POP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>44517</td>
<td>232006</td>
<td>10.093</td>
<td>4411</td>
<td>22987</td>
</tr>
<tr>
<td>Canada</td>
<td>112617</td>
<td>547203</td>
<td>29.109</td>
<td>3869</td>
<td>18798</td>
</tr>
<tr>
<td>Chile</td>
<td>13096</td>
<td>50919</td>
<td>13.994</td>
<td>936</td>
<td>3639</td>
</tr>
<tr>
<td>Denmark</td>
<td>25927</td>
<td>151266</td>
<td>5.207</td>
<td>4979</td>
<td>29050</td>
</tr>
<tr>
<td>Finland</td>
<td>21581</td>
<td>97624</td>
<td>5.085</td>
<td>4244</td>
<td>19199</td>
</tr>
<tr>
<td>France</td>
<td>256316</td>
<td>1330998</td>
<td>57.856</td>
<td>4430</td>
<td>23005</td>
</tr>
<tr>
<td>Greece</td>
<td>9392</td>
<td>98861</td>
<td>10.413</td>
<td>902</td>
<td>9494</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>11758</td>
<td>130823</td>
<td>6.044</td>
<td>1945</td>
<td>21645</td>
</tr>
<tr>
<td>Hungary</td>
<td>7227</td>
<td>41506</td>
<td>10.162</td>
<td>711</td>
<td>4084</td>
</tr>
<tr>
<td>Ireland</td>
<td>17572</td>
<td>52662</td>
<td>3.536</td>
<td>4970</td>
<td>14893</td>
</tr>
<tr>
<td>Israel</td>
<td>11349</td>
<td>74121</td>
<td>5.362</td>
<td>2117</td>
<td>13823</td>
</tr>
<tr>
<td>Italy</td>
<td>145013</td>
<td>1016286</td>
<td>57.177</td>
<td>2536</td>
<td>17774</td>
</tr>
<tr>
<td>Korea, S.</td>
<td>161318</td>
<td>380820</td>
<td>44.501</td>
<td>3625</td>
<td>8558</td>
</tr>
<tr>
<td>Kuwait</td>
<td>2797</td>
<td>24848</td>
<td>1.754</td>
<td>1595</td>
<td>14167</td>
</tr>
<tr>
<td>Malaysia</td>
<td>18874</td>
<td>72505</td>
<td>19.695</td>
<td>958</td>
<td>3681</td>
</tr>
<tr>
<td>Mexico</td>
<td>55073</td>
<td>420788</td>
<td>89.564</td>
<td>615</td>
<td>4968</td>
</tr>
<tr>
<td>Netherlands</td>
<td>48595</td>
<td>334286</td>
<td>15.382</td>
<td>3159</td>
<td>21732</td>
</tr>
<tr>
<td>Norway</td>
<td>13484</td>
<td>122926</td>
<td>4.314</td>
<td>3126</td>
<td>28495</td>
</tr>
<tr>
<td>Portugal</td>
<td>17025</td>
<td>87352</td>
<td>9.824</td>
<td>1733</td>
<td>8892</td>
</tr>
<tr>
<td>Singapore</td>
<td>20648</td>
<td>71039</td>
<td>3.268</td>
<td>6318</td>
<td>21738</td>
</tr>
<tr>
<td>Slovakia</td>
<td>2720</td>
<td>13746</td>
<td>5.325</td>
<td>511</td>
<td>2581</td>
</tr>
<tr>
<td>Slovenia</td>
<td>4520</td>
<td>14386</td>
<td>1.925</td>
<td>2348</td>
<td>7473</td>
</tr>
<tr>
<td>Spain</td>
<td>80104</td>
<td>483652</td>
<td>39.577</td>
<td>2024</td>
<td>12221</td>
</tr>
<tr>
<td>Sweden</td>
<td>34806</td>
<td>198432</td>
<td>8.751</td>
<td>3977</td>
<td>22675</td>
</tr>
<tr>
<td>Switzerland</td>
<td>57503</td>
<td>261388</td>
<td>7.104</td>
<td>8094</td>
<td>36794</td>
</tr>
<tr>
<td>Syria</td>
<td>3317</td>
<td>44753</td>
<td>13.840</td>
<td>240</td>
<td>3234</td>
</tr>
<tr>
<td>Turkey</td>
<td>31115</td>
<td>135961</td>
<td>59.903</td>
<td>519</td>
<td>2270</td>
</tr>
<tr>
<td>UK</td>
<td>244397</td>
<td>1024609</td>
<td>58.005</td>
<td>4213</td>
<td>17664</td>
</tr>
</tbody>
</table>

*Source:* UNIDO Yearbook 1997

*Note:* MANU and GDP are measured in U.S. $ million. POP is measured in million. MANU/POP and GDP/POP are measured in U.S. $.

Using the data in Table 1.1 above, an OLS regression of manufacturing output on GDP yields the following result (standard errors in parentheses):

\[
MANU = 604 + 0.194 \times GDP^2 = 0.8
\]

S.E.E. (5700) (0.013)

This implies that manufacturing accounts for $194,000 out of every $1 million increase in GDP in the cross-section. The residuals from the regression and GDP are both ranked in Table 1.2 and Di and D_i^2 are computed.
The sum of the latter came to 1608. The rank correlation coefficient is thus

\[
1 - \frac{6 \times 1608}{28 \times 783} = 0.56
\]

and the test statistic is \(0.56 \sqrt{27} = 2.91\). This is above 2.58 and hence the null hypothesis of homoscedasticity is rejected at the 1 percent level.

### 3.1.3.5 The Goldfeld–Quandt Test

Goldfeld and Quandt (1965) are so far attributed with the most common formal test for heteroscedasticity. The test assumes that \(\sigma_i\), the standard deviation of the probability distribution of the disturbance term in observation \(i\), is about the size of \(X_i\). It also assumes that the disturbance term is distributed and satisfies the other Gauss–Markov conditions. The size of \(X\) order the \(n\) observations in the sample and separate regressions are carried out for the first \(n\) and the last \(n'\) observations, the middle \((n-2n')\) observations being dropped completely. If heteroscedasticity is present, and if the assumption regarding its nature is correct, the variance of \(u\) in the last \(n'\) observations will be more than that in the first \(n'\) and this will be reflected in the RSS in the two
sub-regressions. Representing these by $RSS_1$ and $RSS_2$ for the sub-regressions with the first $n'$ and the last $n'$ observations, respectively. The ratio $RSS_2/RSS_1$ will be distributed as an $F$-statistic with $(n' - k)$ and $(n' - k)$ degrees of freedom, where $k$ is the number of parameters in the equation, under the null hypothesis of homoscedasticity. The power of the test depends on the choice of $n'$ about $n$. As a result of some experiments undertaken by Goldfeld and Quandt, they recommend that in general, $n'$ should be about 11 when $n$ is 30 and about 22 when $n$ is 60. Which clearly shows that $n'$ should be about $\frac{3}{8}$ of $n$.

If there is more than one explanatory variable in the model, the observations should be ordered by that which is hypothesized to be associated with the null hypothesis for the test is that $RSS_2$ is not significantly greater than $RSS_1$, and the alternative hypothesis is that it is significantly greater. If $RSS_2$ turns out to be smaller than $RSS_1$, the null hypothesis should not be rejected; it only means that there would not be any point in computing the test statistic $RSS_2/RSS_1$. However, the Goldfeld–Quandt test can also be used for the case where the standard deviation of the disturbance term is hypothesized to be inversely proportional to $X_i$. The procedure is the same as before, but the test statistic is now $RSS_1/RSS_2$, and it will again be distributed as an $F$-statistic with $(n' - k)$ and $(n' - k)$ degrees of freedom under the null hypothesis of homoscedasticity.

### 3.1.3.6 The Glejser Test

This test permits you to search the nature of the heteroscedasticity a little more closely. Here, the assumption that $\sigma_u$ is a relative quantity to $X_i$ is relaxed, and you can then investigate whether some other efficient form may be more suitable, for example

$$\sigma_u = \beta_1 + \beta_2 X_i^\gamma$$  \[3.03\]

To use the procedure, you regress $Y$ on $X$ using OLS and then fit the absolute values of the residuals, $|e|$ to the function for a given value of $\gamma$. You may fit several such functions, varying the choice of $\gamma$. In each case the null hypothesis of homoscedasticity will be rejected if the estimate of $\beta_2$ is significantly different from 0.
If more than one function gives rise to a significant estimate of $\beta_2$, that with the best fit may be a guide to the nature of the heteroscedasticity.

### 3.1.3.6 Solution to Heteroscedasticity Problem

Suppose that the true relationship is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots[3.04]$$

Let the standard deviation of the disturbance term in observation $i$ be $\sigma_u$. If you happened to now $\sigma_u$ for each observation, you could eliminate the heteroscedasticity by dividing each observation by its value of $\sigma$. The model becomes

$$\frac{Y_i}{\sigma_u} = \beta_1 \frac{1}{\sigma_u} + \beta_2 \frac{X_i}{\sigma_u} + \frac{u_i}{\sigma_u} \ldots[3.05]$$

The disturbance term $\frac{u_i}{\sigma_u}$ becomes homoscedastic because the population variance of $\frac{u_i}{\sigma_u}$ is

$$E\left(\left(\frac{u_i}{\sigma_u}\right)^2\right) = \frac{1}{\sigma_u^2} E(u_i^2) = \frac{1}{\sigma_u^2} \sigma_u^2 1 \quad \ldots[3.06]$$

That is, every observation will have a disturbance term drawn from a distribution with population variance 1, and the model will be homoscedastic. The revised model may be rewritten as;

$$Y_i = \beta_1 h_i + \beta_2 X_i + u_i' \quad \ldots[3.07]$$

where $Y_i' = \frac{Y_i}{\sigma_u}, \ X_i' = \frac{X_i}{\sigma_u}$, $h_i$ is a new variable whose value in observation $i$ is $\frac{1}{\sigma_u}$, and $u_i' = \frac{u_i}{\sigma_u}$.
Note that there should not be a constant term in the equation. By regressing \( Y' \) on \( h \) and \( X' \), you will obtain efficient estimates of \( \beta_1 \) and \( \beta_2 \) with unbiased standard errors.

### 3.1.3.7 Consequences of Heteroscedasticity

The seriousness of the consequences of heteroscedasticity will depend on the nature of the occurred heteroscedasticity, and there are no general rules. In the case of the heteroscedasticity, where the standard deviation of the disturbance term is proportional to \( X \) and the values of \( X \) are integers from 5 to 44. Here, the population variance of the OLS estimator of the slope coefficient is approximately double that of the estimator, where the heteroscedasticity has been eliminated by dividing through by \( X \). Further, the standard errors of the OLS estimators are underestimated, giving a misleading impression of the precision of the OLS coefficients.

### 3.1.4.0 SUMMARY

This unit begins with a general discussion of heteroscedasticity and its meaning. Also discussed are the reasons why the distribution of a disturbance term may be subject to heteroscedasticity, and the consequences of the heteroscedasticity problem for OLS estimators. We continued with presentation of several tests for heteroscedasticity and methods of alleviating the problem.

### 3.1.5.0 CONCLUSION

The discussion in this unit concludes with an awareness to the students that the concept of heteroscedasticity and associated problems are indications of how an apparent case of heteroscedasticity may be caused by model misspecification.

### 3.1.6.0 TUTOR-MARKED ASSIGNMENT

A researcher investigating whether government expenditure tends to crowd out investment fits the regression (standard errors in parentheses):

\[
\hat{I} = 18.10 - 1.07G + 0.3Y \quad R^2 = 0.99
\]

S.E.E. (7.79) \quad (0.14) \quad (0.02)
She sorts the observations by increasing size of $Y$ and runs the regression again for the 11 countries with smallest $Y$ and the 11 countries with largest $Y$. $RSS$ for these regressions is 321 and 28101, respectively. Perform a Goldfeld–Quandt test for heteroscedasticity.

### 3.1.7.0 REFERENCES /FURTHER READING


### MODULE 4: ECONOMETRIC MODELLING AND AUTOCORRELATION

The general aim of this module is to provide you with a thorough understanding of the basic rudiments of econometric modelling. Stochastic Regression and Measurement Errors, autocorrelation, econometric modelling and models using time series data are explained. By the end of this module, you would have been able to understand the components of the module stated below. The units to be studied are:

**Unit 1: Stochastic Regression and Measurement Errors**

**Unit 2: Autocorrelation**

**Unit 3: Econometric Modelling and Models Using Time Series Data**

### UNIT 1: STOCHASTIC REGRESSORS AND MEASUREMENT ERRORS

**CONTENTS**

4.1.1.0 Introduction  
4.1.2.0 Objectives  
4.1.3.0 Main Content  
4.1.3.1 Stochastic Regressors  
4.1.3.2 Unbiasedness
The least squares regression model assumed that the explanatory variables are nonstochastic, that is, that they do not have random components. Although relaxing this assumption does not in itself undermine the OLS regression technique, it is typically an unrealistic assumption, so it is important you know the consequences of relaxing it. We shall see that in some contexts we can continue to use OLS, but in others, for example when one or more explanatory variables are subject to measurement error, it is a biased and inconsistent estimator.

The main objective of this unit is to provide a general understanding of the topic 'stochastic regressors and measurement errors and point out that random element in a regression model is not the only disturbance term but that the variables themselves do have random components.

Based on the adopted assumption that the regressors, which is the explanatory variables in the regression model are nonstochastic, their values in the sample are therefore fixed and unaffected by the way the sample is generated. Perhaps the best example of a nonstochastic variable is time, which, as we will see when we come to time series analysis, is sometimes included in the regression model as a proxy for
variables that are difficult to measure, such as technical progress or changes in tastes. Nonstochastic explanatory variables are unusual in regression analysis.

A rationale for making the nonstochastic assumption has been one of simplifying the analysis of the properties of the regression estimators. For example, we saw that in the regression model

\[ Y = \beta_1 + \beta_2 X + u \]  \[\text{...}[4.01]\]

the OLS estimator of the slope coefficient may be decomposed as follows:

\[ b_2 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta_2 + \frac{\text{Cov}(X, u)}{\text{Var}(X)} \]  \[\text{...}[4.02]\]

Here, if \( X \) is nonstochastic, so is \( \text{Var}(X) \), and the expected value of the error term can be written \( E[\text{Cov}(X, u)]/\text{Var}(X) \).

Also if \( X \) is nonstochastic, \( E[\text{Cov}(X, u)] \) is 0.

Which easily helps us to prove that \( b_2 \) is an unbiased estimator of \( \beta_2 \).

The desirable properties of the OLS estimators remain unchanged even if the descriptive variables have stochastic components, provided that these components are distributed independently of the disturbance term, and provided that their distributions do not depend on the parameters \( \beta_1, \beta_2 \) or \( \sigma_u \). Let us demonstrate the unbiasedness and consistency properties and as typical, taking an efficient approach.

### 4.1.3.2 Unbiasedness

Once \( X \) is stochastic, \( \text{Var}(X) \) cannot be treated as a scalar, so we cannot rewrite \( E[\text{Cov}(X, u)/\text{Var}(X)] \) as \( E[\text{Cov}(X, u)]/\text{Var}(X) \). Hence the previous proof of unbiasedness is blocked. However, we can find another route by decomposing the error term:

\[ \frac{\text{Cov}(X, u)}{\text{Var}(X)} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{(X_i - \bar{X})}{\text{Var}(X)} \right) (u_i - \bar{u}) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)(u_i - \bar{u}) \]  \[\text{...}[4.03]\]
where \( f(X_i) = \frac{(X_i - \bar{X})}{\text{Var}(X)} \). Now, if \( X \) and \( u \) are independently distributed, 

\[
E[f(X_i)(u_i - \bar{u})]
\]

may be decomposed as the product of \( E[f(x_i)] \) and \( E[(u_i - \bar{u})] \). Hence

\[
E[f(X_i)(u_i - \bar{u})] = E[f(X_i)E(u_i - \bar{u})] = E[f(X_i)] \times 0 \quad \text{...[4.04]}
\]

since by assumption \( E(u_i) \) is 0 in each observation. This implies, of course, that \( E(\bar{u}) \) is also 0.

Hence, when we take the expectation of \( \frac{1}{n} \sum_{i=1}^{n} f(X_i)(u_i - \bar{u}) \), each term within the summation has expected value 0. Thus the error term as a whole has expected value 0 and \( b_2 \) is an unbiased estimator of \( \beta_2 \).

### 4.1.3.3 Consistency

Generally stated, \( \text{plim} (A/B) \) is equal to \( \text{plim}(A)/\text{plim}(B) \), where \( A \) and \( B \) are any two stochastic quantities, on condition that both \( \text{plim}(A) \) and \( \text{plim}(B) \) exist and that \( \text{plim}(B) \) is nonzero ("\( \text{plim} \)" is the limiting value as the sample size becomes large). As also stated, sample expressions tend to their population counterparts as the sample size becomes large, so \( \text{plimCov}(X,u) \) is the population covariance of \( X \) and \( u \) and \( \text{plimVar}(X) \) is \( \sigma^2 \), the population variance of \( X \). If \( X \) and \( u \) are independent, the population covariance of \( X \) and \( u \) is 0 and we can write that:

\[
\text{plim}b_2 = \beta_2 + \frac{\text{plimCov}(X,u)}{\text{plimVar}(X)} = \beta_2 + \frac{0}{\sigma^2} = \beta_2 \quad \text{...[4.05]}
\]

### 4.1.3.4 The Consequences of Measurement Errors

As it is in other human activities, it habitually happens in economics that, when investigating a relationship, the variables involved could be measured defectively. For example, surveys often contain errors caused by the person being interviewed not remembering properly or not understanding the question correctly. However, misreporting is not the only source of inaccuracy. It sometimes happens that you have defined a variable in your model in a certain way, but the available data correspond to a slightly different definition.
4.1.3.5 Measurement Errors in the Descriptive Variable(s)

To keep the analysis simple, we will confine it to the simple regression model. Let us suppose that a variable $Y$ depends on a variable $Z$ according to the relationship

$$ Y_i = \beta_1 + \beta_2 Z_i + v_i \quad \ldots [4.06] $$

where $v$ is a disturbance term with mean 0 and variance $\sigma^2_v$, distributed independently of $Z$. We shall suppose that $Z$ cannot be measured absolutely accurately, and we shall use $X$ to denote its measured value. In observation $i$, $X_i$ is equal to the true value, $Z_i$, plus the measurement error, $w_i$:

$$ X_i = Z_i + w_i \quad \ldots [4.07] $$

We shall suppose that $w$ has mean 0 and variance $\sigma^2_w$, that $Z$ has population variance $\sigma^2_Z$, and that $w$ is distributed independently of $Z$ and $v$.

[4.07] into [4.06], will yield

$$ Y_i = \beta_1 + \beta_2 (X_i - w_i) + v_i = \beta_1 + \beta_2 X_i + v_i - \beta_2 w_i \quad \ldots [4.08] $$

Two random components are present in [4.08], the original disturbance term $v$ and the measurement error (multiplied by $-\beta_2$). Together they form a composite disturbance term, which we shall call $u$:

$$ u_i = v_i - \beta_2 w_i \quad \ldots [4.09] $$

Therefore, [4.08] becomes

$$ Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots [4.10] $$

You have your data on $Y$ (which, for the time being, we shall assume has been measured accurately) and $X$, and you unsuspectingly regress $Y$ on $X$.

As usual, the regression coefficient $b$ is given by

$$ b_2 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \beta_2 + \frac{\text{Cov}(X,u)}{\text{Var}(X)} \quad \ldots [4.11] $$
Looking at the error term, we can see that it is going to behave badly. By [4.07] and [4.09], both $X_i$ and $u_i$ depend on $w_i$. The population covariance between $X$ and $u$ is nonzero and, so $b_2$ is an inconsistent estimator of $\beta_2$. Even if you had a very large sample, your estimate would be inaccurate. In the limit it would underestimate $\beta_2$ by an amount

$$\frac{\sigma_u^2}{\sigma_\varepsilon^2 + \sigma_u^2} \beta_2 \quad \ldots[4.12]$$

4.1.3.6 Measurement Errors in the Dependent Variable

These measurement errors in the dependent variable do not matter as much. In practice, they can be thought of as contributing to the disturbance term. They are undesirable, because anything that increases the noise in the model will tend to make the regression estimates less accurate, but they will not cause the regression estimates to be biased.

By assumption, let the true value of the dependent variable be $Q_i$, and the true relationship be

$$Q_i = \beta_1 + \beta_2 X_i + v_i \quad \ldots[4.13]$$

where $v$ is a disturbance term. If $Y_i$ is the measured value of the dependent variable in observation $i$, and $r_i$ is the measurement error,

$$Y_i = Q_i + r_i \quad \ldots[4.14]$$

which may be rewritten

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots[4.15]$$

where $u$ is the composite disturbance term ($v + r$)

The only difference from the usual model is that the disturbance term in [4.15] has two components: the original disturbance term and the error in measuring $Y$. The important thing is that the explanatory variable $X$ has not been affected. Hence OLS still yields unbiased estimates provided that $X$ is nonstochastic or that it is distributed...
independently of \( v \) and \( r \). The population variance of the slope coefficient will be given by

\[
\sigma^2_{b_2} = \frac{\sigma^2_u}{n\sigma^2_\varepsilon} = \frac{\sigma^2_v + \sigma^2_\varepsilon}{n\sigma^2_\varepsilon} \quad \text{...[4.16]}
\]

and so will be greater than it would have been in the absence of measurement error, reducing the precision of the estimator. The standard errors remain valid but will be larger than they would have been in the absence of the measurement error, reflecting the loss of precision.

4.1.4.0 SUMMARY

In this unit, in other for the students to have understanding of the topic stochastic regressors and measurement errors, we explained conditions under which OLS estimator remain unbiased when the variable in a regression model possessing random components. A demonstration of the unbiasedness and consistency properties was also approached. Equally, the consequences of measurement errors, errors in descriptive and dependent variables were discussed.

4.1.5.0 CONCLUSION

The unit concludes that under general conditions the regression model remain unchanged even if the descriptive variables have stochastic components. Provided that these components are distributed independently of the disturbance term and considering measurement errors in the descriptive and dependent variables.

4.1.6.0 TUTOR-MARKED ASSIGNMENT

In a certain industry, firms relate their stocks of finished goods, \( Y \), to their expected annual sales, \( X^e \), according to a linear relationship

\[
Y = \beta_1 + \beta_2 X^e
\]

Actual sales, \( X \), differ from expected sales by a random quantity \( u \) that is distributed with mean 0 and constant variance:
\[ X = X^e + u \]

uis distributed independently of \(X^e\). An investigator has data on \(Y\) and \(X\) (but not on \(X^e\)) for a cross-section of firms in the industry. Describe the problems that would be encountered if OLS were used to estimate \(\beta_1\) and \(\beta_2\), regressing \(Y\) on \(X\).

### 4.1.7.0 REFERENCES /FURTHER READING


UNIT 2: AUTOCORRELATION

CONTENTS
4.2.1.0 Introduction
4.2.2.0 Objectives
4.2.3.0 Main Content
4.2.3.1 Possible Causes of Autocorrelation
4.2.3.2 Detection of First-Order Autocorrelation: the Durbin–Watson Test
4.2.4.0 Summary
4.2.5.0 Conclusion
4.2.6.0 Tutor-Marked Assignment
4.2.7.0 References/Further Reading

4.2.1.0 INTRODUCTION

Autocorrelation is the correlation between the error terms arising in time series data. Such correlation in the error terms often arises from the correlation of the omitted variables that the error term captures. Furthermore, the assumption in the third Gauss–Markov condition is that the value taken by the disturbance term in any observation and determined independently of its values in all the other observations, is satisfied, and hence that the population covariance of $u_i$ and $u_j$ is 0 for $i \neq j$. When the condition is not satisfied, the disturbance term is said to be subject to autocorrelation, often called serial correlation or cross-autocorrelation.

4.2.2.0 OBJECTIVE

The main objective of this unit is to provide a basic understanding that autocorrelation may arise as a consequence of the exclusion of a significant variable or the mathematical misspecification of regression model.
4.2.3.0 MAIN CONTENTS

The significances of autocorrelation for OLS are to some extent comparable to those of heteroscedasticity. The regression coefficients remain unbiased, but OLS is inefficient because one can find an alternative unbiased estimator with smaller variance. The other main concern, which should not be mixed up with the first, is that the standard errors are estimated wrongly, probably being biased downwards. Finally, although in general autocorrelation does not cause OLS estimates to be biased, there is an important special case where it does.

4.2.3.1 Possible Causes of Autocorrelation

There is two forms autocorrelation occurrence, which could either be positive and negative. Persistent effects of excluded variables are probably the most frequent cause of positive autocorrelation, the usual type of economic analysis. In Figure 4.1, $Y$ depends on $X$ and some minor variables not included explicitly in the specification. The disturbance term in the model is generated by the combined effects of these excluded variables. In the first observation, the excluded variables have a net positive effect and the disturbance term is positive. If the excluded variables change slowly, their positive effect will persist, and the disturbance term will remain positive. In time the balance will change, and the net effect of the excluded variables becomes negative. Here, the persistence effect works the other way, and the disturbance term remains negative for a few observations. The duration and amplitude of each positive and negative sequence are essentially random, but overall there will be a tendency for positive values of the disturbance term to be followed by positive ones and for negative values to be followed by negative ones. However, a factor to note is that autocorrelation is on the whole more likely to be a problem for shorter intervals between observations.
Negative autocorrelation means that the correlation between successive values of the disturbance term is negative. A positive value in one observation is more likely to be followed by a negative value than a positive value in the next, and vice versa; this is shown by an illustrative scatter diagram in Figure 4.2. A line joining successive observations to one another would cross the line relating $Y$ to $X$ with greater frequency than one would expect if the values of the disturbance term were independent of each other. Economic examples of negative autocorrelation are relatively uncommon, but sometimes it is induced by manipulations used to transform the original specification of a model into a form suitable for regression analysis.
When an error term $U_t$ at time period $t$ is correlated with error terms in time series, the correlation between $U_t$ and $U_{t-k}$ is called an autocorrelation of order $k$. The correlation between $U_t$ and $U_{t-1}$ is the first-order autocorrelation and is usually denoted by $\rho_1$. The correlation between $U_t$ and $U_{t-2}$ is called the second order autocorrelation and is denoted by $\rho_2$, and so on. There are $(n-1)$ such autocorrelations if we have $n$ observations. However, we cannot hope to estimate all of these from our data. Hence we often assume that these $(n-1)$ autocorrelations can be represented in terms of one or two parameters.

### 4.2.3.2 Detection of First-Order Autocorrelation: the Durbin–Watson Test
We will mostly be concerned with first-order autoregressive autocorrelation, often denoted AR (1). AR (1) appears to be the most common type of autocorrelation approximation. It is described as positive or negative according to the sign of $\rho$. Note that if $\rho$ is 0, there is no autocorrelation occurrence.

There are two major things that will be discussed in this unit, which are:

1. Test for the presence of serial correlation.
2. Estimate the regression equation when the errors are serially correlated.

**Durbin-Watson Test (DW)**

The simplest and most commonly used model is one where the errors $U_t$ and $U_{t-1}$ have a correlation $\rho$. For this model one can think of testing hypotheses about $\rho$ on the basis of $\rho$, the correlation between the least squares residuals $U_t$ and $U_{t-1}$. A commonly used statistic for this purpose which is related to $\rho$ is the DW statistic, which will be denote by $d$. It is defined as

$$d = \frac{\sum_{2}^{n} (U_t - U_{t-1})^2}{\sum_{2}^{n} u_t^2} \quad \ldots[4.17]$$

Where $U_t$ is the estimated residual for period $t$. DW can be re-written as

$$d = \frac{\sum u_t^2 + \sum u_{t-1}^2 - 2\sum u_t u_{t-1}}{\sum u_t^2} \quad \ldots[4.18]$$

Since $\sum u_t^2$ and $\sum u_{t-1}^2$ are approximately equal if the sample is large, we have $d = 2(1 - \rho)$. If $\rho = +1, then \, d = 0 \, and \, if \, \rho = -1, then \, d = 4. We \, have \, d = 2 \, if \, \rho = 0. \, if \, d \, is \, close \, to \, 0 \, or \, 4, \, the \, residuals \, are \, highly \, correlated.$

The sampling distribution of $d$ depends on the values of the explanatory variables and hence DW derived upper $d(u)$ limits and lower $d(l)$ limits for the significance levels for $d$. There are tables to test the hypothesis of zero autocorrelation against the hypothesis of first-order positive autocorrelation. (For negative autocorrelation we interchange ($l$) and ($d(u)$), hence;
If $d < d(l)$, we reject the null hypothesis of no autocorrelation.
If $d > d(u)$, we do not reject the null hypothesis.
If $d(l) < d < d(u)$ the test is inconclusive.

The upper bound of the DW statistic is a good approximation to its distribution when the regressors are slowly changing. DW argue that economic time series are slowly changing, and hence one can use $d(u)$ as the correct significance point.

The significance points in the DW tables are tabulated for testing $\rho = 0$ against $\rho > 0$. If $d > 2$ and we wish to test the hypothesis $\rho = 0$ against $\rho < 0$, we consider $4 \ldots d$ and refer to the DW tables as if we are testing for positive autocorrelation. Although we have said that $d \rightarrow 2(1 - \rho)$ this approximation is valid only in large samples. The mean of $d$ when $\rho = 0$ has been shown to be given approximately by

$$E(d) = 2 + \frac{2(k-1)}{n-k}$$

...[4.19]

where $k$ is the number of regression parameters estimated (including the constant term), and $n$ is the sample size. Thus, even for zero serial correlation, the statistic is biased upward from 2. If $k = 5$ and $n = 15$, the bias is as large as 0.8.

### 4.2.5.0 SUMMARY

The unit explained the concept of autocorrelation at first order, its possible causes and detection (with particular interest on the first-order autoregressive autocorrelation, denoted by AR (1)) using Durbin-Watson test.

### 4.2.4.0 CONCLUSION

In this unit, autocorrelation is statistically explained as a random process that measures the linear correlation between values of the process at different times, as a function of time or of the time lag. The significances of autocorrelation for OLS are shown to be comparable to those of heteroscedasticity and have two forms of occurrences, which could either be positive or negative. Students are advised to use the further reading materials to look at more autocorrelation techniques and study more on the correlation between the error terms arising in time series data as indicated in the introduction section of this unit.
4.2.6.0 REFERENCES /FURTHER READING

UNIT 3: ECONOMETRIC MODELLING AND MODEL USING TIME-SERIES DATA

CONTENTS
4.3.1.0 Introduction
4.3.2.0 Objectives
4.3.3.0 Main Content
4.3.3.1 The Adaptive Expectations Model
4.3.4.0 Summary
4.3.5.0 Conclusion
4.3.6.0 Tutor-Marked Assignment
4.3.7.0 References/Further Reading

4.3.1.0 INTRODUCTION

The modelling of expectations using time series data is often an important and difficult task of the applied economist. This is especially true in macroeconomics, in that investment, saving, and the demand for assets are all sensitive to expectations about the future. Unfortunately, there is no satisfactory way of measuring expectations directly for macroeconomic purposes. Consequently, macroeconomic models tend not to give particularly accurate forecasts, and this makes economic management difficult.

4.3.2.0 OBJECTIVE

The main objective of this unit is to introduce the application of regression analysis to time series data, starting with static models and then continuing to dynamic models with lagged variables used as descriptive variables.

4.3.3.0 MAIN CONTENTS

4.3.3.1 The Adaptive Expectations Model

As a makeshift solution, some models use an indirect technique known as the adaptive expectations process. This involves a simple learning process in which, in each period, the actual value of the variable is compared with the value that had been expected. If
the actual value is greater, The expected value is adjusted upwards for the next period. If it is lower, the expected value is adjusted downwards. The size of the adjustment is hypothesized to be proportional to the discrepancy between the actual and expected value.

If $X$ is the variable in question, and $X^e_t$ is the value expected in time period $t$ given the information available at time period $t-1$,

$$X^e_{t+1} - X^e_t = \lambda(X_t - X^e_t)(0 \leq \lambda \leq 1) \quad \text{...}[4.20]$$

This can be rewritten

$$X^e_{t+1} = \lambda X_t + (1 - \lambda)X^e_t(0 \leq \lambda \leq 1) \quad \text{...}[4.21]$$

Which states that the expected value of $X$ in the next period is a weighted average of the actual value of $X$ in the current period and the value that had been expected. The larger the value of $\lambda$, the quicker the expected value adjusts to previous actual outcomes.

For example, suppose that you hypothesize that a dependent variable, $Y_t$, is related to the expected value of the descriptive variable, $X$, in year $t+1$, $X^e_{t+1}$:

$$Y_t = \beta_1 + \beta_2 X^e_{t+1} + u_t \quad \text{...}[4.22]$$

expresses $Y_t$ in terms of $X^e_{t+1}$, which is unobservable and must somehow be replaced by observable variables, that is, by actual current and lagged values of $X$, and perhaps lagged values of $Y$. We start by substituting for $X^e_{t+1}$,

$$Y_t = \beta_1 + \beta_2(\lambda X^e_{t+1} + (1 - \lambda)X^e_t ) + u_t = \beta_1 + \beta_2\lambda X^e_{t+1} + \beta_2(1 - \lambda)X^e_t + u_t \quad \text{...}[4.23]$$

Of course, we still have unobservable variable $X^e_t$ as a descriptive variable, but if it is true for time period $t$, it is also true for time period $t-1$:

$$X^e_t = \lambda X_t + (1 - \lambda)X^e_t \quad \text{...}[4.24]$$

Substituting for $X^e_t$, in [4.23] we now have
\[ Y_t = \beta_1 + \beta_2 \lambda X_t + \beta_2 (1 - \lambda) X_{t-1} + \beta_2 \lambda (1 - \lambda)^2 X_{t-2} + \cdots + \beta_2 \lambda (1 - \lambda)^s X_{t-s+1} + \beta_2 (1 - \lambda)^s X_{t-s+1} + \epsilon_t \]

Now it is reasonable to suppose that \( \lambda \) lies between 0 and 1, in which case \((1 - \lambda)^s\) will also lie between 0 and 1. Thus \((1 - \lambda)^s\) becomes progressively smaller as \(s\) increases. Eventually, there will be a point where the term \(\beta_2 (1 - \lambda)^s X_{t-s+1}\) is so small that it can be neglected and we have a model in which all the variables are observable.

A lag structure with geometrically declining weights, such as this one, is described as having a Koyck distribution. It is highly sparing regarding its constraint, requiring only one parameter more than the static version. Since it is nonlinear in the parameters, OLS should not be used to fit it, for two reasons. First, multicollinearity would almost certainly make the estimates of the coefficients so erratic that they would be worthless – it is precisely this problem that caused us to search for another way of specifying a lag structure. Second, the point estimates of the coefficients would yield conflicting estimates of the parameters.

**4.3.4.0 SUMMARY**

In this unit, an indirect technique known as the adaptive expectations process is explained as a makeshift solution used in some models. This involves simple learning process for which, in each period, the size of adjustment is proportional to the discrepancy between the actual and expected value.

**4.3.5.0 CONCLUSION**

The unit introduces the application of regression analysis to time series data, starting with static models and then proceeding to dynamic models with lagged variables used as descriptive variables. In this unit, the adaptive expectations process was mainly used for explanation but the multicollinearity problem in time series models, especially dynamic ones with lagged descriptive variables was also explained. The students may use the reference materials for more understanding and further readings.

**4.3.6.0 TUTOR-MARKED ASSIGNMENT**
1.) The results of linear and logarithmic regressions of consumer expenditure on food, \( FOOD \), on \( DPI \) and a relative price index series for food, \( PRELFOOD \), using the Demand Functions data set, are summarized below. Provide an economic interpretation of the coefficients and perform appropriate statistical tests.

\[
\hat{FOOD} = 232.6 + 0.089DPI + 0.534PRELFOOD \quad R^2 = 0.989
\]
S.E.E.(31.9) (0.002) (0.332)

\[
\hat{LGF\hat{O}OD} = 2.66 + 0.61LGDPI - 0.30LGPRELFOOD \quad R^2 = 0.993
\]
S.E.E(0.28) (0.01) (0.07)

2.) Sometimes a time trend is included in a regression as an explanatory variable, acting as a proxy for some gradual change not associated with income or price. Changing tastes might be an example. However, in the present case, the addition of a time trend might give rise to a problem of multicollinearity because it will be highly correlated with the income series and perhaps also the price series. Calculate the correlations between the \( TIME \) variable in the data set, \( LGDPI \), and the logarithm of expenditure on your category. Regress the logarithm of expenditure on your category on \( LGDPI \), the logarithm of the relative price series and \( TIME \) (not the logarithm of \( TIME \)). Provide an interpretation of the regression coefficients, perform appropriate statistical tests, and compare the regression results with those of the same regression without \( TIME \).

4.3.7.0 REFERENCES /FURTHER READING


MODULE 5: SIMULTANEOUS EQUATION, BINARY CHOICE, AND MAXIMUM LIKELIHOOD ESTIMATION

The general aim of this module is to provide you with a thorough understanding of the basic rudiments of Simultaneous Equation, Binary Choice, and Maximum Likelihood Estimation. By the end of this module, you should be able to understand the components of the module stated below. The units to be studied are;

Unit 1: Simultaneous Equation

Unit 2: Binary Choice and Limited Dependent Models with Maximum Likelihood Estimation

UNIT 1: SIMULTANEOUS EQUATION ESTIMATION

CONTENTS
5.1.1.0 Introduction
5.1.2.0 Objectives
5.1.3.0 Main Content
5.1.3.1 Simultaneous Equations Models: Structural and Reduced Form Equations
5.1.3.2 Simultaneous Equations Bias
5.1.4.0 Summary
5.1.5.0 Conclusion
5.1.6.0 Tutor-Marked Assignment
5.1.7.0 References/Further Reading

5.1.1.0 INTRODUCTION

In the engagement of OLS to estimate the factors of an equation that is set in a simultaneous equations model, it is likely that the estimates will be biased and erratic which would invariably make the statistical tests invalid and inconsistent.
5.1.2.0 OBJECTIVE

The main objective of this unit is to demonstrate to the students that in practice most economic relationships interact with others in a system of simultaneous equations and when this is the case the application of OLS to a single relationship in isolation yields biased estimates.

5.1.3.0 MAIN CONTENTS

5.1.3.1 Simultaneous Equations Models: Structural and Reduced Form Equations

As explained earlier in other modules, measurement error is not the only probable cause why the fourth Gauss–Markov condition may not be satisfied. Simultaneous equations bias is another. To illustrate this; suppose there is an investigation on the determinants of price inflation and wage inflation. For ease, it would be better to start with a very simple model that supposes that \( p \), the annual rate of growth of prices, is related to \( w \), the annual rate of growth of wages, it being assumed that increases in wage costs force prices upwards:

That is;

\[
p = \beta_1 + \beta_2 w + u_p \quad \text{...}[5.01]
\]

Here, \( w \) is related to \( p \) and \( U \), the rate of unemployment, workers protecting their real wages by demanding increases in wages as prices rise, but their ability to do so being the weaker, the higher the rate of unemployment \( (\alpha_3 < 0) \). Which is stated as:

\[
w = \alpha_1 + \alpha_2 p + \alpha_3 U + u_w \quad \text{...}[5.02]
\]

where, \( u_p \) and \( u_w \) are disturbance terms

Clearly, this simultaneous equations model involves a certain amount of complexity: \( w \) determines \( p \) in the first equation \([5.01]\), and in turn, \( p \) helps to determine \( w \) in the second \([5.02]\). For better clarity in resolving this complexity, we need to make a distinction between endogenous and exogenous variables. Endogenous variables are variables whose values are determined by the interaction of the relationships in the model. Exogenous ones are those whose values are determined...
externally. Thus in the present case, \( p \) and \( w \) are both endogenous, and \( U \) is exogenous. The exogenous variables and the disturbance terms ultimately determine the values of the endogenous variables, once the complexity is cleared. The mathematical relationships expressing the endogenous variables regarding the exogenous variables and disturbance terms are known as the reduced form equations. The original equations that we wrote down when specifying the model are described as the structural equations. We will derive the reduced form equations for \( p \) and \( w \). To obtain that for \( p \), we take the structural equation for \( p \) and substitute for \( w \) from the second equation:

\[
p = \beta_1 + \beta_2 w + u_p = \beta_1 + \beta_2 (\alpha_1 + \alpha_2 p + \alpha_3 U + u_w) + u_p \quad \text{...}[5.03]
\]

Hence,

\[
(1 - \alpha_2 \beta_2)p = \beta_1 + \alpha_1 \beta_2 + \alpha_3 \beta_2 U + u_p + \beta_2 u_w \quad \text{...}[5.04]
\]

and so we have the reduced form equation for \( p \):

\[
p = \frac{\beta_1 + \alpha_1 \beta_2 + \alpha_3 \beta_2 U + u_p + \beta_2 u_w}{(1 - \alpha_2 \beta_2)} \quad \text{...}[5.05]
\]

Similarly we obtain the reduced form equation for \( w \):

\[
w = \alpha_1 + \alpha_2 p + \alpha_3 U + u_w = \alpha_1 + \alpha_2 (\beta_1 + \beta_2 w + u_p) + \alpha_3 U + u_w \quad \text{...}[5.06]
\]

Hence

\[
(1 - \alpha_2 \beta_2)w = \alpha_1 + \alpha_2 \beta_1 + \alpha_3 U + u_w + \alpha_2 u_p \quad \text{...}[5.07]
\]

and so

\[
w = \frac{\alpha_1 + \alpha_2 \beta_1 + \alpha_3 U + u_w + \alpha_2 u_p}{1 - \alpha_2 \beta_2} \quad \text{...}[5.08]
\]

5.1.3.2 Simultaneous Equations Bias

In almost all simultaneous equations models, the reduced form equations express the endogenous variables regarding all of the exogenous variables and all of the
disturbance terms. You can see that this is the case with the price inflation/wage inflation model. In this model, there is only one exogenous variable, \( U \).

\( w \) depends on it directly; \( p \) does not depend on it directly but does so indirectly because \( w \) determines it. Similarly, both \( p \) and \( w \) depend on \( u_p \), \( p \) directly and \( w \) indirectly. And both depend on \( u_w \), \( w \) directly and \( p \) indirectly.

The dependence of \( w \) on \( u_p \) means that OLS would yield inconsistent estimates if used to fit equation [5.01], the structural equation for \( p \). \( w \) is a stochastic regressor and its random component is not distributed independently of the disturbance term \( u_p \). Similarly the dependence of \( p \) on \( u_w \) means that OLS would yield inconsistent estimates if used to fit [5.02]. Since [5.01] is a simple regression equation, it is easy to analyze the large-sample bias in the OLS estimator of \( \beta_2 \).

5.1.5.0 SUMMARY

In this unit, we started by explaining structural and reduced form of equations which was illustrated by showing that measurement error is not the only probable cause why the fourth Gauss–Markov condition may not be satisfied for which the biasness of simultaneous equations is an example. We then went further to explain the simultaneous equations bias.

5.1.4.0 CONCLUSION

In this unit, Simultaneous equations estimation is discussed. The structural and reduced form of equations as it relates to Simultaneous equations model bias is also explained. Clearly, for deeper understanding of the equation models, the students should make a distinction between endogenous and exogenous variables in Simultaneous equations estimation.

5.1.6.0 TUTOR-MARKED ASSIGNMENT
1.) Simple macroeconomic model consists of a consumption function and an income identity:

\[ C = \beta_1 + \beta_2 Y + u \]
\[ Y = C + I \]

where \( C \) is aggregate consumption, \( I \) is aggregate investment, \( Y \) is aggregate income, and \( u \) is a disturbance term. On the assumption that \( I \) is exogenous, derive the reduced form equations for \( C \) and \( Y \).

2.) From the model above, demonstrate that OLS would yield inconsistent results if used to fit the consumption function, and investigate the direction of the bias in the slope coefficient.

5.1.7.0 REFERENCES /FURTHER READING


Unit 2: Binary Choice and Limited Dependent Models and Maximum Likelihood Estimation

CONTENTS
5.2.1.0 Introduction
5.2.2.0 Objectives
5.2.3.0 Main Content
5.2.3.1 The Linear Probability Model
5.2.3.2 Goodness of Fit and Statistical Tests
5.2.4.0 Summary
5.2.5.0 Conclusion
5.2.6.0 Tutor-Marked Assignment
5.2.7.0 References/Further Reading

5.2.1.0 INTRODUCTION

Most times economists are known to be interested in the factors behind the decision-making of individuals or enterprises. Examples are:

- Why do some people go to college while others do not?
- Why do some women enter the labour force while others do not?
- Why do some people buy houses while others rent?
- Why do some people migrate while others stay put?

Models have been developed to proffer solutions to these interest, and they are known as a binary choice or qualitative response models. The outcome will be denoted by $Y$, and assigned a value of 1 if the event occurs and 0 otherwise. Models with more than two possible outcomes have also been developed, but let us restrict our scope to a binary choice. The linear probability model apart, binary choice models are fitted using maximum likelihood estimation.

5.2.2.0 OBJECTIVE

The main objective of this unit is to provide the students with a clear understanding that apart from the linear probability model, binary choice models are fitted using maximum likelihood estimation.
5.2.3.0 MAIN CONTENTS

5.2.3.1 The Linear Probability Model

The simplest binary choice model is the linear probability model where, as the name implies, the probability of the event occurring, \( p \), is assumed to be a linear function of a set of descriptive variable(s). That is:

\[
p_i = p(Y_i = 1) = \beta_1 + \beta_2 x_i \quad \text{...}[5.09]
\]

For one descriptive variable, the relationship is as shown in Figure 5.1. Of course, \( p \) is unobservable, and as expected there is only one data \( Y \), on the outcome. In the linear probability model, this is used as a dummy variable for the dependent variable.

![Figure 5.1. Linear Probability Model](image)

Regrettably, the linear probability model though simple still has some serious defects. First, there are problems with the disturbance term. As usual, the value of the dependent variable \( Y_i \) in observation \( i \), has a nonstochastic component and a random component. The nonstochastic component depends on \( X_i \) and the parameters and is the expected value of \( Y_i \) given \( X_i \), \( E(Y_i | X_i) \). The random component is the disturbance term.

\[
Y_i = E(Y_i | X_i) + u_i \quad \text{...}[5.10]
\]
It is simple to compute the nonstochastic component in observation \( i \) because \( Y \) can take only two values. It is 1 with probability \( p_i \) and 0 with probability \((1 - p_i)\):

\[
E(Y_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i = \beta_1 + \beta_2 X_i \quad \text{...[5.11]}
\]

The expected value in observation \( i \) is therefore \( \beta_1 + \beta_2 X_i \). This means that we can rewrite the model as:

\[
Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{...[5.12]}
\]

Probability function is thus also the nonstochastic component of the relationship between \( Y \) and \( X \). It follows that, for the outcome variable \( Y_1 \) to be equal to 1, as represented by the point \( A \) in Figure 5.2, the disturbance term must be equal to \((1 - \beta_1 - \beta_2 X_i)\). For the outcome to be 0, as represented by the point \( B \), the disturbance term must be \((-\beta_1 - \beta_2 X_i)\). Thus the distribution of the disturbance term consists of just two specific values.

![Figure 5.2. Linear Probability Model](image)

Which means that the standard errors and the usual test statistics are invalidated. For good measure, the two possible values of the disturbance term change with \( X \), so the distribution is heteroscedastic as well. It can be shown that the population variance of \( u_i \) is \((\beta_1 + \beta_2 X_i) (1 - \beta_1 - \beta_2 X_i)\), and this varies with \( X_i \).
The other problem is that the predicted probability may be greater than 1 or less than 0 for extreme values of $X$. The first problem is dealt with by fitting the model with a technique known as maximum likelihood estimation.

The second problem involves elaborating the model as follows. Define a variable $Z$ that is a linear function of the descriptive variables. In the present case, since we have only one descriptive variable, this function is:

$$Z_i = \beta_1 + \beta_2 X_i$$  \[5.13\]

### 5.2.3.2 Goodness of Fit and Statistical Tests

Even though numerous measures have been proposed for comparing alternative model specifications, there is still no measure of goodness of fit equivalent to $R^2$ in maximum likelihood estimation. Denoting the actual outcome in observation $i$ as $Y_i$, with $Y_i = 1$ if the event occurs and 0 if it does not, and denoting the predicted probability of the event occurring $\hat{P}_i$, the measures include the following:

- **i.** the number of outcomes correctly predicted, taking the prediction in observation $i$ as 1 if $\hat{P}_i$ is greater than 0.5 and 0 if it is less;
- **ii.** the sum of the squared residuals $\sum_{i=1}^{n}(Y_i - \hat{P}_i)^2$;
- **iii.** the correlation between the outcomes and predicted probabilities, $r_{\hat{P}_i Y_i}$;
- **iv.** the pseudo-$R^2$ in the logit output.

Every of these measures has its shortcomings, and it is recommended to consider more than one and compare their results. Nevertheless, the standard significance tests are similar to those for the standard regression model. The significance of an individual coefficient can be evaluated via its $t$ statistic. However, since the standard error is valid only asymptotically (in large samples), the same goes for the $t$ statistic, and since the $t$ distribution converges to the normal distribution in large samples, the critical values of the latter should be used. The counterpart of the $F$ test of the explanatory power of the model ($H_0$: all the slope coefficients are 0, $H_1$: at least one is nonzero) is a chi-squared test with the chi-squared statistic in the logit output distributed under $H_0$ with degrees of freedom equal to the number of explanatory variables.
5.2.4.0 SUMMARY

In this unit, we started with the linear probability model being the simplest binary choice model where the probability of the event occurring is assumed to be a linear function of a set of descriptive variables. We then proceeded to goodness of fit and statistical tests using maximum likelihood estimation as a method of estimating the parameters of a model given observations, by finding the parameter values that maximise the likelihood of making the observations given the parameters.

5.2.5.0 CONCLUSION

Although numerous measures have been proposed for comparing alternative model specifications, there is still no measure of goodness of fit equivalent to maximum likelihood estimation. The students should be of the opinion that every of the estimation measure has its shortcomings and it is recommended to consider more than one and compare their results.

5.2.6.0 TUTOR-MARKED ASSIGNMENT

A researcher, using a sample of 2,868 individuals from the NLSY (National Longitudinal Survey of Young Men), is investigating how the probability of a respondent obtaining a bachelor’s degree from a four-year college is related to the respondent’s score on ASVABC. 26.7 percent of the respondents earned bachelor’s degrees. ASVABC ranged from 22 to 65, with mean value 50.2, and most scores were in the range 40 to 60. Defining a variable BACH to be equal to 1 if the respondent has a bachelor’s degree (or higher degree) and 0 otherwise, the researcher fitted the OLS regression (standard errors in parentheses):

\[
\hat{\text{BACH}} = -0.864 + 0.023 \text{ASVABC} \quad R^2 = 0.21
\]

S.E.E. (0.042) (0.001)

The researcher also fitted the following logit regression:

\[
z = -11.103 + 0.189 \text{ASVABC}
\]

S.E.E. (0.487) (0.009)

where \(Z\) is the variable in the logit function. Using this regression, the researcher plotted the probability and marginal effect functions shown in the diagram below.
a.) Give an interpretation of the OLS regression and explain why OLS is not a satisfactory estimation method for this kind of model.

b.) With reference to the diagram below, discuss the variation of the marginal effect of the ASVABC score implicit in the logit regression and compare it with that in the OLS regression.

c.) Sketch the probability and marginal effect diagrams for the OLS regression and compare them with those for the logit regression. (In your discussion, make use of the information in the first paragraph of this question.)

**5.2.7.0 REFERENCES /FURTHER READING**

