NATIONAL OPEN UNIVERSITY OF NIGERIA

RESEARCH METHODOLOGY
COURSE CODE: ECO 715

FACULTY OF SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

COURSE CONTENT DEVELOPERS
Dr Aminu Muhammad Mustapha
Department of Economics and Development Studies,
Faculty of Arts and Social Sciences
Federal University Dutse, Jigawa State - Nigeria
&
Dr *(Mrs)* Fa'izah Adhama Mukhtar
Department of Economics and Development Studies,
Faculty of Arts and Social Sciences
Federal University Dutse, Jigawa State - Nigeria

Course Content Editor
Prof. Dimis I. Mailafia
Department of Economics
University of Jos
Plateau State

© 2020 by NOUN Press
National Open University of Nigeria,
Headquarters,
University Village,
Plot 91, Cadastral Zone,
Nnamdi Azikiwe Expressway,
Jabi, Abuja.

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

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

Printed: 2020 ISBN: 978-058-023-X.

CONTENT

Introduction

Welcome to ECO: 715 RESEARCH METHODOLOGY

ECO 715: Research Methodology is a three-credit and one-semester postgraduate course for Development Economics students. The course is made up of twelve units spread across twelve lectures weeks. This course guide gives you an insight to research and its methods in social sciences and Economics in particular in a broader way and how to study make use of the methods in applied research. It tells you about the course materials and how you can work your way through these materials. It suggests some general guidelines for the amount of time required of you on each unit in order to achieve the course aims and objectives successfully. Answers to your tutor marked assignments (TMAs) are therein already.

Course Content

This course is basically on Research Methodology because as you are aspiring to become a development economist, you must be able to apply the knowledge of research to find answers to various economic questions and problems. The topics covered include Research methodology and the philosophy of economic research; identification of researchable problems and the development of hypotheses or research questions. The course will also involve a detailed treatment of the methods and problems of collecting relevant research data, the format for presenting research results (i.e. from designing the table of contents to referencing or bibliography). Also to be covered are the various methods of establishing relationships between economic variables; basic elements of model building in economics;

application of multivariate analysis, correlation and discriminant analysis; tests of causality Chi-square tests, etc. students will be required to write a seminar paper in this course.

Course Aims

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

- Research methodology
- Conceptual and Philosophical foundations of Economic research
- Identification of researchable problems and the development of hypotheses or research questions.
- Data and its types and manipulation
- Methods of collecting relevant research data,
- Variable types and their measurement
- Sampling
- Basic elements of model building in economics;
- Bivariate and multivariate methods of establishing relationships between economic variables;
- Common statistical tests in economics
- The format for presenting research results
- Referencing

Course Objectives

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

- be acquainted with all salient aspects of research methodology in accordance with the current body of scientific literature on this indispensable area of social science

- apply the theoretical knowledge acquired in this course with the appropriate context-related modifications - be applied to numerous real-life situations in Economics

- stimulate interest in the field as a prospective career field

Demands on Course Participants

- Acquisition and careful application of knowledge

- Analytical and critical thinking, innovation, inquisitiveness
- All-inclusive viewpoint

Assumptions about the students. You:

- have basic knowledge of economic theories
- know basic statistics and social sciences analytical techniques
- are able to think abstractly
- think critically (but not in extreme form – cynicism, which is a barrier to understanding)
- have the ability to synthesize from the facts and information in front of you
- have the ability to discern privately held beliefs from concepts supported by evidence – i.e. need for objectivity
- You are currently initiating a research project

Working Through the Course

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

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

Course Material

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

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

Study Unit

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

MODULE ONE: INTRODUCTION TO RESEARCH AND RESEARCH METHODS

UNIT 1 Introduction to Research
UNIT 2 Philosophical and Conceptual Foundations of Research

MODULE TWO: DATA: TYPES AND COLLECTION METHODS

Unit 1 Data: Types and manipulation
Unit 2 Methods of Collecting Research Data

MODULE THREE: BASICS OF MODEL BUILDING IN ECONOMICS

UNIT 1 Variables Types
UNIT 2 Measurement of Variables
UNIT 3 Sampling
UNIT 4 Model Building in Economics

MODULE FOUR: QUANTITATIVE DATA ANALYSIS

UNIT 1 Bivariate Analysis
UNIT 2 Multivariate Analysis

MODULE FIVE: RESEARCH ORGANISATION AND REPORTING

UNIT 1 Manuscript Structure and Contents
UNIT 2 Basics of Citing Sources

Each study unit will take at least two hours, and it include the introduction, objective, main content, self-assessment exercise, conclusion, summary and reference. Other areas border on the Tutor-Marked Assessment (TMA) questions. Some of the self-assessment exercise will necessitate discussion, brainstorming and argument with some of your colleges. You are advised to do so in order to understand and get acquainted with fundamentals of economics research methods and their applications.

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

Textbooks and References

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

Abu-Rizaiza, O. (2009). *How to write scientific Articles: A handbook for Non-native Speakers of English* (1st ed.). Jeddah, KSA: King Abdulaziz University Press.

Allchin, D. (2001). Error types. *Perspectives on Science*, *9*(1), 38-58.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Asika, N. (1991). *Research Methodology in the Behavioural Sciences.* Lagos, Nigeria: Longman Nigeria.

Bajpai, S. R., & Bajpai, R. C. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health*, *3*(2), 112-115.

Begg, D., Fischer, S., & Dornbusch, R. (2000). *Economics* (6th ed). London: McGraw-Hill.

Bhakar, S. S., & Nathani, N. (2015). *A Handbook on writing Research Paper in Social Sciences*. New Delhi, India: Bharti Publications.

Bhattarai, K. (2015). *Research Methods for Economics*. UK: University of Hull Business School.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education, 3*(1), 1-8.

Collis, J. & Hussey, R. (2003) *Business Research: A Practical Guide for Undergraduate and Postgraduate Students* (2nd ed.). Basingstoke: Palgrave Macmillan.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Muffin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

DeCoster, J. (2001). Transforming and Restructuring Data. Retrieved <September, 12, 2017 downloaded this file> from http://www.stat-help.com/notes.html

Devillis, R. E. (2006). Scale development: Theory and application. *Applied Social Science Research Method Series, Vol. 26*, Newbury Park: SAGE.

Dimitrios, A., & Stephen G. H. (2007). *Applied Econometrics: A Modern Approach*. NY: Palgrave Macmillan

Drost, E. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1).

Garba, T. (n. d.) Research Method. ECO 903 Lecture notes, *Department of Economics, Usmanu Danfodiyo University Sokoto.*

Greener, S. (2008). *Business Research Methods*. London: Ventus.

Gujarati, D. (2004). Basic Econometrics (4th ed.). NY: McGraw-Hill.

Jones, P., Evans, M., & Lipson, K. (2008). *Essential Further Mathematics – Core* (4th ed.). Cambridge University.

Joshua, O. (2013). *The Essentials of Research Methodology and Statistics in Education.* Jigbik.

Kabir, S.M.S. (2016). Writing Research Report. In S. M. S. Kabir (Ed,), *Basic Guidelines for Research: An Introductory Approach for all Disciplines* (1st ed., pp 500-518). Chittagong, Bangladesh: Book Zone.

Kenny, D. A. (1979). *Correlation and Causality*. NY: John Wiley

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

Kumar, R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners* (3rd ed.). New Delhi: SAGE.

Madan, C. R., & Kiesinger, E. A. (2017). Test–retest reliability of brain morphology estimates. *Brain Informatics, 4,* 107–121.

Malhotra, N. K. (2004). *Marketing Research: An Applied Orientation* (4th ed.). New Jersey: Pearson Education, Inc.

Mankiw, N. G. (2001). *Principles of Economics* (2nd ed.). Orlando: Harcourt College.

Mohajan (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University, 17*(3): 58-82

Nicholson, W., & Snyder, C. (2012). *Microeconomic Theory Basic Principles and Extensions* (10th ed.). OH, USA: South-Western CENGAGE Learning.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). NY: Mcgraw-Hill.

Prentice, C. (2013). *Introduction to the APA* (6th ed.). Minnesota: Saint Mary's University.

Ramayah, T., Ahmad, N.H., Abdul Halim, H., Zainal, S.R.M., & Lo, M. (2010). Discriminant analysis: An illustrated example. *African Journal of Business Management, 4*(9).

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.). NY: John Wiley & Sons.

Samuelson, P. A., & Nordhaus, W. D. (1998). *Economics* (16th ed.). Irwin: McGraw-Hill.

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business: A Skill Building Approach* (5th ed.). West Sussex, UK: John Wiley and Sons.

Shiker, M. A.K. (2012). Multivariate statistical analysis. *British Journal of Science, 6* (1).

Studenmund, A. H. (2000). *Using Econometrics: A Practical Guide* (4th ed.). Addison-Wesley.

Trochim, W. M. K. (2006). Introduction to validity. *Social Research Methods*, retrieved from www.socialresearchmethods.net/kb/introval.php, September 9, 2010.

UCOL Student Success Team. (2015). *A Guide to APA 6th ed. Referencing Style.* Matauranga: Author.

Walliman, N. (2011). *Research Methods: The Basics*. London: Routledge.

Weiner, J. (2007). *Measurement: Reliability and Validity Measures*. Bloomberg School of Public Health: Johns Hopkins University

Yarnold, P. R. (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Emergency severity index (version 3) and Canadian triage acuity scale. *Optimal Data Analysis, 3*(4), 50-54.

Assignment File

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

There are five assignments in this course. The five course assignments will cover:
Assignment 1 - All TMAs’ question in Units 1 – 2 (Module 1)
Assignment 2 - All TMAs' question in Units 3 – 4 (Module 2)
Assignment 3 - All TMAs' question in Units 5 – 8 (Module 3)
Assignment 4 - All TMAs' question in Units 9 – 10 (Module 4)
Assignment 5 - All TMAs' question in Units 11 – 12 (Module 5)

Presentation Schedule

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

Assessment

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

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

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

Tutor-Marked Assignments (TMAs)

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

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

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

Final Examination and Grading

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

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

Course Marking Scheme

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

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

Course Overview

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Research Methodology (ECO 715).

Units	**Title of Work**	**Week's Activities**	**Assessment (end of unit)**
	Course Guide		
MODULE 1: INTRODUCTION TO RESEARCH AND RESEARCH METHODS			
1	Introduction to Research	Week 1	Assignment 1
2	Philosophical and Conceptual Foundations of Research	Week 2	Assignment 1
MODULE 2: DATA: TYPES AND COLLECTION METHODS			
3	Data: Types and Manipulation	Week 3	Assignment 2
4	Methods Collecting Research Data	Week 4	Assignment 2
MODULE 3: BASICS OF MODEL BUILDING IN ECONOMICS			
5	Variable Types	Week 5	Assignment 3
6	Measurement of Variables	Week 6	Assignment 3
7	Sampling	Week 7	Assignment 3
8	Model Building in Economics	Week 8	Assignment 3
MODULE FOUR: QUANTITATIVE DATA ANALYSIS			
9	Bivariate Analysis	Week 9	Assignment 4
10	Multivariate Analysis	Week 10	Assignment 4
MODULE FIVE: RESEARCH ORGANISATION AND REPORTING			
11	Manuscript Structure and Contents	Week 11	Assignment 5

12	Basics of Citing Sources	Week 12	Assignment 5
	Revision	*Week 13*	
	Examination	*Weeks 14 & 15*	

How to Get the Most from This Course

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

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

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

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

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

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

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

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

1. Read this Course Guide thoroughly.

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

Tutors and Tutorials

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

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

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

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

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

Summary

The course, Research Methodology (ECO 715), expose you to the analysis of Research Methodology and you will also be introduced to its philosophical and conceptual foundation. This course also gives you an insight into the identification and formulation of research problem and hypothesis and research organization as well as report writing. Thereafter it shall enlighten you about data types and its different transformations; methods of collecting research data and variables as well as their measurement. Finally, the course will expose you to the quantitative data analysis covering the basic elements of model building and the different statistical tests in economics

On successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective discussion on Research Methodology. However, to gain a lot from the course please try to apply anything you learn in the course to term papers writing in other economics courses. We wish you success with the course and hope that you will find it fascinating and handy.

MODULE ONE: INTRODUCTION TO RESEARCH AND RESEARCH METHODS

UNIT 1 Introduction to research

UNIT 2 Philosophical and conceptual foundations of research

UNIT ONE: OVERVIEW OF RESEARCH

CONTENTS

1.0 INTRODUCTION

Research evolves from human quest for knowledge. It is therefore tailored along the theoretical perspectives or orientation to guide the logic of enquiry; tools and techniques of data collection, analysis and the systematic format for organizing the research and reporting its findings. The hallmarks or main distinguishing characteristics of scientific research are among others its Purposiveness, Rigor, Testability, Replicability, Precision and confidence, Objectivity, Generalizability and Parsimony. There are different 'types of research' depending on their nature and field of specialization. It is, therefore, useful to first of all take a look at the underlying basic features so as to be able to identify the different types of research within their respective connotations and usage. The type of research would also vary depending on the objectives of the study.

2.0 OBJECTIVES

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

- Describe what research is
- Differentiate between research methods and methodology
- Explain the main characteristics of research
- Distinguish between the different types of research

3.0 MAIN CONTENT

3.1 Overview of research

At the outset, knowledge of philosophical, theoretical and conceptual perspectives lays the foundation of research and exposes a researcher to characteristics underlying the variables to be studied. The application of theories in solving real world problems involves systematic collection of information on variables identified by the relevant theory. Before analyzing the variables and empirically testing the claims made by theories, we must need data that describes them. However, no data can be systematically collected without adequate knowledge of techniques of data collection. Likewise, data analysis will not be possible without knowledge of statistical techniques. These statistical procedures are now handy in computer packages and help in great deal in making data analysis. Findings from this process will be aimed at improving knowledge. To effectively communicate the new knowledge such scholarly manuscript must be structured and presented clearly and logically. Scholars use standardised style to acknowledge the source of information used in the process of acquiring the new knowledge. Therefore, research methodology as a broader term consists of five important elements:

i. Theoretical perspectives or orientation to guide research and logic of enquiry.

ii. Tools and techniques of data collection.

iii. Methods of data analysis.

iv. Research organization and reporting.

v. Referencing.

3.2 Research and research methodology

Research is about finding new things and making original contribution to the literature (Bhattarai, 2015). Research as a process of enquiry and investigation; it is systematic, methodological and ethical. It can help solve practical problems and increase knowledge (Neville, 2007). It is the systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analyzing the facts and reaching certain conclusions either in the form of solutions(s) towards the concerned problem or in certain generalizations for some theoretical formulation (Kothari, 2004).

We can infer from the foregoing that research involves a step by step process of inquiry. We have to find new things in order to make contribution to knowledge. In our quest for the new knowledge we must follow a systematic and scientific procedure. Therefore, the methods of research are interdependent on each other in solving or talking theoretical or practical problem.

3.2.1 Distinction between Research Method and Methodology

The term methodology refers to the overall approaches & perspectives to the research process as a whole and is concerned with the following main issues:

- **Why** you collected certain data
- **What** data you collected
- **Where** you collected it
- **How** you collected it
- **How** you analysed it

(*Collis & Hussey*, 2003, p.55).

Methodology is therefore, the manner in which we approach and execute systematic investigative activities. Within a discipline, there are accepted rules of evidence and reasoning. Research methodology provides the principles for organizing, planning, designing and conducting research. (It does not tell you how to do specific research).

A **research method** on the other hand refers only to the various specific tools or ways data can be collected and analysed, e.g. a questionnaire; interview checklist; data analysis software etc.). It consists of specific details of how we do the task.

We need to further differentiate research methodology from research methods:

- ➢ Methodology – general approaches or guidelines
- ➢ Methods – specific details and/or procedures to accomplish a task
- ➢ *One course cannot teach all methods in Agricultural Economics!*
- ➢ **Examples of methods?**
- ➢ (Regression analysis, optimization models, surveys, matching, simulation etc.)

3.2.2 Research Methodology in Economics

Research methodology in Economics is a study which integrates the various components of economics to accomplish a defined, goal-directed research. It aims at:

- ○ expanding our knowledge and make it useful to the study of world economic problems.
- ○ helping us to learn by doing under the supervision of an advisor (shown to be an effective model)
- ○ pulling together various aspects of economic theories, methods, and analysis to present in a coherent, logical, reliable and useful manner.
- ○ providing a time-tested, proven means of producing new, reliable theories and their applications in solving real world economic problems

SELF ASSESSMENT EXERCISE

Describe in your own words what you understand by research

3.3 The main characteristics of scientific research

The hallmarks or main distinguishing characteristics of scientific research are among others its Purposiveness, Rigor, Testability, Replicability, Precision and confidence, Objectivity, Generalizability and Parsimony. Explanation of each of these characteristics thus follows:

3.3.1 Purposiveness

Research must have an aim, reason or definite purpose. For instance, a business firm may commission a research with the aim of establishing level of acceptance of its products in the market which will translate into more turnover and increased profitability. The research thus has a **purposive** focus.

3.3.2 Rigor

A good theoretical base and a sound methodological design add **rigor** to a purposive study. Rigor connotes carefulness, scrupulousness, and the degree of exactitude in research investigations. In the case of our example in 1 above, let us say the manager asks 10 to 12 of its potential and actual customers to indicate what would increase their level of acceptance of the firm’s product. If, solely on the basis of their responses, the manager reaches several conclusions on how the product acceptance in the market can be increased, the whole approach to the investigation is unscientific. It lacks rigor for the following reasons:

i. The conclusions are incorrectly drawn because they are based on the responses of just a few employees whose opinions may not be representative of those of the entire workforce.

ii. The manner of framing and addressing the questions could have introduced bias or incorrectness in the responses.

iii. There might be many other important factors such as price competitiveness of the product, promotional activities etc. that this small sample of respondents did not or could not verbalize during the interviews, and the researcher has therefore failed to include them.

Therefore, conclusions drawn from an investigation that lacks a good theoretical foundation, as evidenced by reason III, and methodological sophistication, as evident from I and II above, are unscientific. Rigorous research involves a good theoretical base and a carefully thought-out methodology. These factors enable the researcher to collect the right kind of information from an appropriate sample with the minimum degree of bias, and facilitate suitable analysis of the data gathered.

3.3.3 Testability

If, after talking to a random selection of respondents and study of the previous research done in the area, the researcher develops certain hypotheses on how product acceptance can be enhanced, then these can be tested by applying certain statistical tests to the data collected for the purpose. For instance, the researcher might hypothesize that price has significant effect on product acceptance. This is a hypothesis that can be tested when the data are collected. The use of several statistical tests, confirm or reject such tentative statement. Scientific research thus lends itself to testing logically developed hypotheses to see whether or not the data support the educated conjectures or hypotheses that are developed after a careful study of the problem situation. **Testability** thus becomes another hallmark of scientific research.

3.3.4 Replicability

This has to do with how findings can be simulated on the basis of data collected by a different research employing the same methods. To put it differently, the results of the tests of hypotheses should be supported again and yet again when the same type of research is repeated in other similar circumstances. To the extent that this does happen (i.e., the results are *replicated* or repeated), we will gain confidence in the scientific nature of our research. In other words, our hypotheses have not been supported merely by chance, but are reflective of the true state of affairs in the population. **Replicability** is thus another hallmark of scientific research.

3.3.5 Precision

In social science research, we seldom have the luxury of being able to draw “definitive” conclusions on the basis of the results of data analysis. This is because we are unable to study the universe of items, events, or population we are interested in, and have to base our findings on a sample that we draw from the universe. In all probability, the sample in question may not reflect the exact characteristics of the phenomenon we are trying to study. Measurement errors and other problems are also bound to introduce an element of bias or error in our findings. However, we would like to design the research in a manner that ensures that our findings are as close to reality (i.e., the true state of affairs in the universe) as possible, so that we can place reliance or confidence in the results.

Precision refers to the closeness of the findings to “reality” based on a sample. In other words, precision reflects the degree of accuracy or exactitude of the results on the basis of the sample, to what really exists in the universe. You may use the term *confidence interval* in statistics, which is what is referred to here as precision. **Confidence** refers to the probability that our estimations are correct. That is, it is not merely enough to be precise, but it is also important that we can confidently claim that 95% of the time our result will be true and there is only a 5% chance of our being wrong. This is also known as the *confidence level*. The narrower the limits within which we can estimate the range of our predictions (i.e., the more precise our findings) and the greater the confidence we have in our research results, the more useful and scientific the findings become. In social science research, a 95% confidence level – which implies that there is only a 5% probability that the findings may *not* be correct – is accepted as conventional, and is usually referred to as a significance level of 0.05 ($p = 0.05$). Thus, precision and confidence are important aspects of research, which are attained through appropriate scientific sampling design. The greater the precision and confidence we aim at in our research, the more scientific is the investigation and the more useful are the results.

3.3.6 Objectivity

The conclusions drawn through the interpretation of the results of data analysis should be objective; that is, they should be based on the facts of the findings derived from actual data, and not on our own subjective or emotional values. If we are to go by the researcher's conviction all along, then there was no need to do the research in the first place! The more objective the interpretation of the data, the more scientific the research investigation becomes. Though researchers might start with some initial subjective values and beliefs, their interpretation of the data should be stripped of personal values and bias. They should be particularly sensitive to this aspect. **Objectivity** is thus another hallmark of scientific investigation.

3.3.7 Generalizability

Generalizability refers to the scope of applicability of the research findings in different contexts or settings. Obviously, the wider the range of applicability of the solutions generated by research, the more useful the research is to the users. The more generalizable the research, the greater its usefulness and value. However, not many research findings can be generalized to all other settings, situations, or organizations. For wider generalizability, the research sampling design has to be logically developed and a number of other details in the data-collection methods need to be meticulously followed. However, a more elaborate sampling design, which would doubtless increase the generalizability of the results, would also increase the costs of research. Most applied research is generally confined to research within the particular organization where the problem arises, and the results, at best, are generalizable only to other identical situations and settings. Though such limited applicability does not necessarily decrease its scientific value (subject to proper research), its generalizability is restricted.

3.3.8 Parsimony

Simplicity in explaining the phenomena or problems that occur, and in generating solutions for the problems, is always preferred to complex research frameworks that consider an unmanageable number of factors. Therefore, the achievement of a meaningful and parsimonious, rather than an elaborate and cumbersome, model for problem solution becomes a critical issue in research. Economy in research models is achieved when we can build into our research framework a lesser number of variables that explain the variance far more efficiently than a complex set of variables that only marginally add to the variance explained. **Parsimony** can be introduced with a good understanding of the problem and the important factors that influence it. Such a good conceptual theoretical model can be realized through unstructured and structured interviews with the concerned people, and a thorough literature review of the previous research work in the particular problem area.

In sum, scientific research encompasses the eight criteria just discussed. These are discussed in more detail later in this course. At this point, a question that might be asked is why a scientific approach is necessary for investigations when systematic research by simply collecting and analysing data would produce results that could be applied to solve the problem. The reason for following a scientific method is that the results will be less prone to error and more confidence can be placed in the findings because of the greater rigor in application of the design details. This also increases the replicability and generalizability of the findings.

SELF ASSESSMENT EXERCISE

What are the main characteristics of research?

3.4 Types of Research

There are different 'types of research' depending on their nature and field of specialization. While a classification based on a dichotomous distinction (like, theoretical/applied, descriptive/analytical, conceptual/empirical, etc.) is possible, it is necessary to recognize that there may be overlapping features rendering such a classification less perfect to that extent. It is, therefore, useful to first of all take a look at the underlying basic features so as to be able to identify the different types of research within their respective connotations and usage. The type of research would also vary depending on the objectives of the study. Research design varies with the type of research one likes to pursue. With this background, we can now proceed to know the distinctions between different types of research.

The general classification of research is divided into Qualitative and quantitative. Its classification according the nature of the study is descriptive and analytical while according to the research design we have exploratory and conclusive. Apart from that, there other types of research that may be distinguished from the mainstream.

3.4.1 General classification

Broadly two types of approaches are used in conducting research in social sciences: quantitative and qualitative. However, mixed methods research by combining quantitative methods and qualitative methods has also emerged as an approach to social enquiry.

a. **Quantitative research:** the studies conducted within the perspective (framework) of positivism/post-positivism/realism generally resort the quantitative approach and are termed as 'quantitative research'. **Quantitative Research** is the systematic and scientific investigation of quantitative properties and phenomena and their relationships. The objective of quantitative research is to develop mathematical models, and to test hypotheses. It thus integrates purposes and procedures that are deductive, objective and generalized. Emphasis is laid on the construction of general theories which are applied universally. Well controlled procedures with large number of cases are followed in conducting the studies.

Quantitative approach can be further sub-classified into: inferential approach; experimental approach; and simulation approach. The purpose of **inferential approach research** is to form a data base from which to infer characteristics or relationships of population. This usually means survey research where a sample of population is studied (questioned or observed) to determine its characteristics, and it is then inferred that the population has the same characteristics.

Experimental approach is characterised by much greater control over the research environment and in this case some variables are manipulated to observe their effect on other variables. Simulation approach involves the construction of an artificial environment within which relevant information and data can be generated. This permits an observation of the dynamic behaviour of a system (or its sub-system) under controlled conditions. The term 'simulation' in the context of business and social sciences applications refers to "the operation of a numerical model that represents the structure of a dynamic process. Given the values of initial conditions, parameters and exogenous variables, a simulation is run to represent the behaviour of the process over time." Simulation approach can also be useful in building models for understanding future conditions.

b. **Qualitative research:** on other hand, are the studies conducted within the perspective of critical theory and interpretive paradigms. By using induction as a research strategy, qualitative research creates the theory and discovery through flexible, emergent research designs. It tries to evolve meaning and interpretation based on closer contacts between researchers and the people they study. Thus qualitative research consists of purposes and procedures that integrate inductive, subjective and contextual approaches. It deals with the subjective assessment of attitudes, opinions and behaviour of respondents in the field. Results are generated either in non-quantitative form or in a form which are subjected to relatively less rigorous quantitative treatment. Various techniques like group discussions, projective techniques, in-depth interviews etc., are used.

3.4.2 Classification of research based on its nature (Descriptive and Explanatory Research)

a. **Descriptive research: describes** a situation, events or social systems. It aims to describe the state of affairs as it exists. Surveys and fact-finding enquiries of different kinds are part of descriptive research. Survey methods of all kinds including comparative and correlational methods are used in descriptive research studies. A survey of socio-economic conditions of rural/urban labour is an area of descriptive research. In descriptive research studies, the researchers have no control over variable. They can report only what has happened or is happening. In social science and business research we quite often use the term *Ex post facto* research for descriptive research studies. Most ex post facto research projects are used for descriptive studies in which the researcher seeks to measure such items as, for example, frequency of shopping, preferences of people, or similar data. Ex post facto studies also include attempts by researchers to discover causes even when they cannot control the variables.

b. **Explanatory research**: aims at establishing the cause and effect relationship. The researcher uses the facts or information already available to analyse and make a critical evaluation of the data/information. An example of explanatory research is: 'whether increase in agricultural productivity is explained by improved rural roads?' Analytical research often extends the descriptive approach to suggest or explain why or how something is happening. An important feature of this type of research is in locating and identifying the different factors (or variables) involved.

3.4.3 **Other types of research**

a. **Exploratory research** aims at developing the hypothesis rather than testing a pre-conceived hypothetical contention or notion. Exploratory research is undertaken when few or no previous studies exist. The aim is to look for patterns, hypotheses or ideas that can be tested and will form the basis for further research. Typical research techniques would include case studies, observation and reviews of previous related studies and data.

b. formalized or conclusive research studies are those with substantial structure and with specific hypotheses to be tested. Formalised research studies deal with a definitive structure within which specific hypotheses are tested.

c. **Theoretical research** can also be considered as '*fundamental or basic research*' as its outcome serves as a foundation for all subsequent development in the field. Fundamental (or basic) research mainly concerns with formulation of theory with knowledge perceived as an end in itself. It, thus, aims at obtaining knowledge of the logical processes involved in a phenomenon. It pertains to the quest for knowledge about a phenomenon without concern for its practical use. Such a research may either verify the old theory or establish a new one. For example, fundamental research in economics may consist of research to develop and improve economic theories or evolve quantitative techniques to measure parameters such as multiplier effect, elasticity of demand and supply, etc. Fundamental research is essentially positive in nature. Basic research, therefore, can be treated as building blocks for applied research.

d. Applied research, on the other hand, aims at finding a solution for a problem facing society or industry. It is, thus, applied to practical situations or contexts. While pure or basic research discovers principles and laws, applied research discovers ways of applying them to solve specific problems. It is useful to test the theories developed empirically and can as a result also contribute to improving the tools and techniques of measurement. Illustrations of applied research in economics can be measurement of poverty, employment, rural development, agriculture, environment, etc. Applied research thus aims at finding a solution for an immediate problem facing a society or an industrial/business organisation. Research studies, concerning human behaviour carried on with a view to make generalisations about human behaviour, are also examples of fundamental research, but research aimed at certain conclusions (say, a solution) facing a concrete social or business problem is an example of applied research. Research to identify social, economic or political trends that may affect a particular institution or the copy research (research to find out whether certain communications will be read and understood) or the marketing research or evaluation research are examples of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problem, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.

e. Conceptual research is related to abstract ideas or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret the existing theories. It is thus related to some abstract ideas or theory.

f. Empirical research relies on experience or observation. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. This type of research is particularly useful when validation or verification of an aspect is required. We can also call it as experimental type of research. In such a research it is necessary to get at facts first hand, at their source, and actively to go about doing certain things to stimulate the production of desired information. In such a research, the researcher must first provide himself with a working hypothesis or guess as to the probable results. He then works to get enough facts (data) to prove or disprove his hypothesis. He then sets up experimental designs which he thinks will manipulate the persons or the materials concerned so as to bring forth the desired information. Such research is thus characterised by the experimenter's control over the variables under study and his deliberate manipulation of one of them to study its effects. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis.

g. Experimental research aims at identifying the causal factors by means of experiments.

h. In evaluative research, the cost effectiveness of a programme is examined. Such research addresses the question of the efficiency of a programme and are useful in taking policy decisions on issues like whether the programme is effective and/or needs modification or re-orientation? Whether it should be continued?

i. Predictive Research The aim of Predictive research is to speculate intelligently on future possibilities, based on close analysis of available evidence of cause and effect, e.g. predicting when and where future economic recession might take place

j. Historical research utilizes existing documents to study events or ideas of the past including the philosophy of persons and groups at any remote point of time.

k. Action research is another type of research that deals with real world problems aimed at finding out practical solutions or answers to them. It gathers feedback which is then used to generate ideas for improvement.

l. One-time and Longitudinal

Form the point of view of time, we can think of research either as *one-time research or longitudinal research*. The former is confined to a single time-period, whereas in the latter case the research is carried on over several time-periods.

SELF ASSESSMENT EXERCISE

With the use of relevant examples, explain the different types of research

4.0 CONCLUSION

This unit concludes that research is a systematic scientific process of inquiry aimed at solving a problem. A thin line can be drawn from research methods and methodology. The latter is more general while the former a specific. The position of research in Economics is indispensable. Research has distinct characteristics and various classifications.

5.0 SUMMARY

In this unit, we have discussed an overview of research, its meaning, distinguishing characteristics and different types. A thin line was drawn between a research and research methodology.

6.0 TUTORED MARKED ASSIGNMENTS

1. Briefly give an overview of research.
2. How is a research different from a research methodology?
3. Describe the characteristics of research.

7.0 REFERENCES/FURTHER READINGS

Bhattarai, K. (2015). *Research Methods for Economics*. UK: University of Hull Business School.

Collis, J. & Hussey, R. (2003) *Business Research: A Practical Guide for Undergraduate and Postgraduate Students* (2nd ed.). Basingstoke: Palgrave Macmillan.

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business: A Skill Building Approach* (5th ed.). West Sussex, UK: John Wiley and Sons.

UNIT 2 PHILOSOPHICAL AND CONCEPTUAL FOUNDATIONS OF RESEARCH

CONTENTS

1.0 INTRODUCTION

Research is a systematic way for finding things that are not known, which are called research problems, it is a process consisting of the identifying and defining research problem, formulating and testing the hypothesis through data collection, organization and analysis, making deductions and reaching of conclusion from the test results of the hypotheses, and reporting and evaluating the research (Pardede, 2018). Prior to the creation of a research proposal, there is the need to identify a problem to address and then questions to ask regarding the targeted problem. This module first discusses the identification of research problem and research hypothesis

2.0 OBJECTIVE

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

- describe philosophical foundation of economic research
- identify and explain research problem
- explain the steps involved in formulating research problem
- Develop research questions and hypotheses

3.0 MAIN CONTENT

3.1 Philosophical foundation of research

All philosophical positions and their attendant methodologies, explicitly or implicitly, hold a view about reality. This view, in turn, will determine what can be regarded as legitimate knowledge. Philosophy works by making arguments explicit. You need to develop sensitivity towards philosophical issues so that you can evaluate research critically. It will help you to discern the underlying, and perhaps contentious, assumptions upon which research reports are based even when these are not explicit, and thus enable you to judge the appropriateness of the methods that have been employed and the validity of the conclusions reached. Obviously, you will also have to consider these aspects in regard to your own research work. Your research, and how you carry it out, is deeply influenced by the theory or philosophy that underpins it. There are different ways of going about doing research depending on your assumptions about what actually exists in reality and what we can know and how we can acquire knowledge.

Theoretical perspectives relate to theories of knowledge which lies within the domain of philosophy of social science. The key concept associated with the perspectives is the paradigm. Some authors have classified these paradigms into three categories by re-naming them as:

3.1.1 Realism

Realism assumes that there is a real world that is external to the experience of any particular person and the goal of research is to understand that world.

3.1.2 Constructivism

Constructivism assumes that everyone has unique experience and beliefs and it posits that no reality exists outside of those perceptions.

3.1.3 Pragmatism

Pragmatism considers realism and constructivism as two alternate ways to understand the world. However, the questions about the nature of reality are less important than questions about what is meant to act and experience the consequence of those actions.

The knowledge of all these perspectives enable a researcher to make a meaningful choice about the research problem;

i. the research questions to investigate this problem;
ii. the research strategies to answer these questions;
iii. the approaches to social enquiry that accompany these strategies;
iv. the concept and theory that direct the investigation;
v. the sources, forms and types of data;
vi. the methods for collecting and analyzing the data to answer these questions.

SELF ASSESSMENT EXERISE

Briefly explain the three paradigms that underpin research

3.2 Conceptual Foundations of Research

Research is deeply influenced by the theory or philosophy that underpins it. There are different approaches to research depending on your assumptions about what actually exists in reality and what we can know and how we can acquire knowledge. In our reasoning process, we are faced with two separate worlds – the world of reality and the imaginary world (the world of theory or abstraction). Reality is the phenomenon which exist somewhere and can be verified empirically. Scientist are interested in the world of reality. In science, we get close to reality through experience, logical reasoning, induction and deduction. Whenever, we develop a process that led to eventual attainment of reality, we have contributed to knowledge.

The researcher works through a thought process that requires some basic scientific concepts or language in order to understand the reality. These are facts, concepts, constructs, rules, hypotheses, theories and laws. He may also need models, definitions and generalisations.

3.2 1 Fact

Fact is a reality which exists. It is a truth that can be known only by observation or experience. Most scientific discoveries are facts or realities which are knowable but at one time or the other were not known experienced or observed. With time and through gradual scientific process, the unknown becomes known. Thus, the unknown reality becomes a known fact. Example: sometimes in the past electricity existed but unknown, so also solar radiation and other forms of electricity

3.2.2 Concepts

A concept is our perception of reality to which we have attached some word labels for the purpose of identification. To conceptualise, mean the mental image of reality. Thus, concepts are very important in research and science. A concept therefore expresses an abstraction formed from our generalisation of different forms of reality.

3.2.3 Constructs

A construct is simply as a concept that is deliberately defined for a particular scientific purpose which becomes a concept when formalised. Thus there is a thin line between a concept and construct.

3.2.4 Variables

A variable is a construct or concept to which numerical values can be attached. It is a concept which can take on different values. For example, height, weight, income. There is the need to identify the different variables so as to k now how to manipulate them and obtain desired result in a research.

3.2.5 Definitions

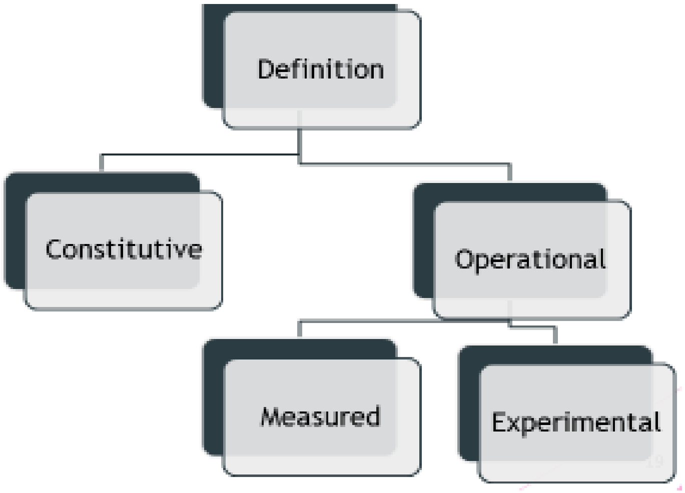
The desired meaning assigned to concept or variable in order to define it. There two kinds: Constitutive and operational

i. In constitutive definition, we substitute the concept or construct we are defining with other concepts or constructs. For Example, we define trade as an act of buying and selling. Thus, we have substituted the concept of trade with the concepts of act, buying, selling. It gives dictionary definition of variables, concepts or constructs.

ii. In operational definition, the concept or construct is assigned a type of meaning which we want it to carry throughout the study. Here, the researcher manipulates or measures the concept to get desired result.

    Operational definition can either be ***measured or experimental***. Example, performance / productivity as the quantity of goods produced, no of students supervised and graduated etc. An experimental operational definition shows how the researcher can practically manipulate a given variable to get the result. Example, Motivation: Give a worker incentive or reward to observe whether he /she is motivated.

Figure 3.1: Definitions



Source: Authors

3.2.6 Hypothesis

Tentative statement about relationships that exist between two or among many variables. It is a statement about relationships that need to be tested and subsequently accepted or rejected. It puts together all the concepts, constructs and the variables to give the researcher a clearer view of the problem(s) under study.

3.2.7 Laws

If hypothesis is true, it states a law. It is a statement of invariant relationship among observable or measurable properties. We can distinguish laws of nature from the laws of Science. Laws of nature hold independently anyone is aware of it or not. On the other hand, laws of science are hypotheses or postulates which are the objects of rational belief based on evidence and which states the laws of nature.

3.2.8 Theories

Invariant relationship among measurable phenomenon with the purpose of explaining and predicting phenomenon. A theory consists of constructs, concepts, definitions, propositions (hypotheses) and all these put together to present a systematic view of phenomenon and possibly predict the phenomenon.

SELF ASSESSMENT EXERISE

Briefly explain the concepts that underpin research

3.3 Research Process

Research process refers to the different steps involved in a desired sequence in carrying out research. However, this does not mean that these steps are always in a given sequence. The following are the series of actions or steps necessary to effectively carry out research:

3.3.1 Formulating the research problem: A research problem is the issue being addressed in a study. The issue can be a difficulty or conflict to be eliminated, a condition to be improved, a concern to handle, a troubling question, a theoretical or practical controversy (or a gap) that exists in scholarly literature. It is the focus or reason for engaging in a research, it is the obstacles which hinder the researcher's path.

3.3.2 Literature review: A literature review is a step-by-step process that involves the identification of published and unpublished work from secondary data sources on the topic of interest, the evaluation of this work in relation to the problem, and the documentation of this work (Sekaran & Bougie, 2009). In order to gain knowledge on your area of research you need to read and assess the previous studies so as to know where your will fit into that body.

3.3.3 **Developing the hypothesis:** A research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable.

3.3.4 **Preparing the research design:** It constitutes the blueprint for the collection, measurement and analysis of data. It includes an outline of what the researcher will do from writing the hypothesis and its operational implications to the final analysis of data. More explicitly, the design decisions happen to be in respect of The sampling design which deals with the method of selecting items to be observed for the given study; The observational design which relates to the conditions under which the observations are to be made; The statistical design which concerns with the question of how many items are to be observed and how the information and data gathered are to be analysed; and The operational design which deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

3.3.5 **Collecting the data:** Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The goal for data collection is to capture quality evidence that then translates to rich data analysis and allows the building of a convincing and credible answer to questions that have been posed. It is one of the most important stages in conducting a research. The task of data collection begins after a research problem has been defined and research design/plan drawn out. While deciding about the method of data collection to be used in a study, two types of data should be kept in mind, primary and secondary.

3.3.6 **Analysis of data:** The analysis of data requires a number of related operations on the raw data such as coding, tabulation and then drawing statistical inferences. Hypotheses earlier stated will be subjected to tests of significance to determine with what validity data can be said to indicate any conclusion(s). If the researcher had no hypothesis to start with, he might seek to explain his findings on the basis of some theory.

3.3.7 **Preparation of the research report:** At the final stage, writing of report must keep in view (i) the preliminary pages; (ii) the main text, and (iii) the end matter.

i. **Preliminary pages** contain the title, acknowledgements, certification, a table of contents a list of tables and list of graphs and charts, if any, given in the report.

ii. **The main text** of the report should have Introduction: It should contain a clear statement of the objective of the research and the scope of the study along with various limitations. Conceptual, empirical and the theoretical literature review An explanation of the methodology adopted in accomplishing the research A statement of findings, and recommendations in non-technical language. If the findings are extensive, they should be summarized.

iii. **At the end of the report,** bibliography, i.e., list of books, journals, reports, etc., consulted, should be given; appendices should be enlisted in respect of all technical data.

In the subsequent units each process will be discussed in much detail.

SELF ASSESSMENT EXERCISE

List the different steps involved in carrying out research.

3.4 Identification of Research Problem

Before moving into the identification of research problem, it is important to know what is a research problem. A research problem is the issue being addressed in a research study. The issue can be a difficulty or conflict to be eliminated, a condition to be improved, a concern to handle, a troubling question, a theoretical or practical controversy (or a gap) that exists in scholarly literature. It is the focus or reason for engaging in a research, it is the obstacles which hinder the researcher's path.

A research problem helps in narrowing the topic down to something that is reasonable for conducting a study. Creswell (2012) defined research problem as "a general educational issue, concern, or controversy addressed in research that narrows the topic. Ogbonna (2006) defines research problem as a felt difficult, a puzzle, a vague feeling or a guest in the researcher's mind to complete a blank or fill in a gap in the researcher's experience. Awotunde and Ugodulunwa (2004) defines a research problem as an unanswered question. Identification of research problem is the first step in a research process. It serves as a foundation upon which other activities in the research process are build, if it is well formulated, you can expect a good study to follow. Student researchers finds it difficult to identify research problems. Identifying the research problem could be accomplished by asking ourselves the following questions; what is the issue, problem, or controversy that needs to be addressed? What controversy leads to a need for this study? What were the concerns being addressed prior to this study? The following steps are to be followed in identifying a research problem:

a. Determining the field of research in which the research is to be conducted.
b. Develop the mastery on the area.
c. Review previous researches conducted in the area to know the recent trend and studies conducted in the area, this will aid in identifying the problem.
d. Draw an analogy and insight in identifying a problem or employ personal experience of the field in locating the problem or seek for assistance from an expert in the field.
e. Pin point specific aspect of the problem which is to be investigated.

3.4.1 The Sources of Research Problem:

a. The classroom, school, home, community and other agencies of education are obvious sources.
b. Social developments and technological changes are constantly bringing forth new problems and opportunities for research.
c. Record of previous research, such specialized sources as the encyclopedias of educational, research abstracts, research bulletins, research reports, journals of researches, dissertations and many similar publications are rich sources of research problems.
d. Text book assignments, special assignments, reports and term papers will suggest additional areas of needed research.
e. Discussions-Classroom discussions, seminars and exchange of ideas with faculty members and fellow scholars and students will suggest many stimulating problems to be solved, close Professional relationships, academic discussions and constructive academic climate are especially advantageous opportunities.
f. Questioning attitude: A questioning attitude towards prevailing practices and research oriented academic experience will effectively promote problem awareness.

g. The most practical source of problem is to consult supervisor, experts of the field and most experienced persons of the field. They may suggest most significant problems of the area.

3.4.2 Steps in formulating a research problem

The formulation of a research problem is the most crucial part of the research as the quality and relevance of the research entirely depends upon it. After the identification of research problem, the next step is to formulate research problem. Every step that constitutes the how part of the research excursion depends upon the way the researcher formulates his/her research problem. The process of formulating a research problem consists of a number of steps as follows:

a. Identify a broad field or subject area of interest: A research should think of a research area that he/she is interested in. If a researcher is not interested in a research area, he/she will not be keen on the research. Also finding an interested area will help him/her to have the zeal to conduct the research effectively.in an interesting topic, for example, if you are an economic researcher you might be interested in researching consumer behavior, energy pricing etc. As far as the research journey goes, these are the broad research areas. It is imperative that he/she identifies one of interest before undertaking the research excursion.

b. Divide the broad area into subareas: At the beginning, the researcher will realize that the broad areas mentioned (consumer behaviour and energy pricing) have many aspects. For example, there are many aspects and issues in the area of energy, such as oil, gas, wind, hydro, thermal. Make a list of these areas. In preparing this list of subareas there is the need to consult others who have some knowledge in the area and review literature in the subject area.

c. The researcher should select what is of most interest him/her: The researcher should select subarea which is most suitable and adoring to him/her. This is because the researchers' interest should be the most important determinant for selection, even though there are other considerations. One way to decide what interests the researcher most is to start with the process of elimination. Go through the list and delete all those subareas in which he/she is not very interested in.

d. Raise research question:. At this step the researcher should ask his/her self, ‘What is it that I want to find out about in this subarea?’ He/she should make a list of whatever questions come to his/her mind involving the chosen subarea and if he/she thinks they are too many to be handled, chose the ones that are more important and discard the rest.

e. Formulate objectives:Both the main objectives and sub objectives should be formulated, the objectives should be drawn from the research questions: The main difference between objectives and research questions is the way in which they are written. Research questions are obviously those – questions. Objectives transform these questions into behavioural aims by using action oriented words such as ‘to find out’, ‘to determine’, ‘to ascertain’ and ‘to examine’ ‘to analyse’.

f. Assess objectives: The researcher should examine the objectives to find out the feasibility of achieving them through the research process. Consider them in the light of the time, resources (financial and human) and technical expertise at disposal.

g. Double-check: The researcher should go back and give final consideration to whether or not he/she is sufficiently interested in the study, and have adequate resources to undertake it.

SELF ASSESSMENT EXERCISE

1. List the steps involved in identifying a research problem.
2. List and explain the steps involved in formulating a research problem.
3. List at least 5 sources of identifying research problem.

3.5 Developing Research questions and hypotheses

3.5.1 Research questions

Research questions help researchers to focus on their research by providing a path through the research and writing process.

Steps involved in developing a research question:

a. Choose an interesting general topic: Most professional researchers focus on topics they are genuinely interested in studying. A researcher should always choose a broad topic about which he/she would like to know more. An example of a general topic might be “greenhouse gases emission and global warming.”

b. Do some preliminary research on your general topic and narrow it down: Search current periodicals and journals on the topic to see what’s already been done, this will assist in narrowing the focus. What issues are scholars and researchers discussing, when it comes to topic? What questions occur to the researcher as he/she read these articles?

c. Consider the audience: For most college papers, your audience will be academic, but always keep your audience in mind when narrowing your topic and developing your question. Would that particular audience be interested in the question you are developing?

d. Ask questions relating to the research topic: Taking into consideration all of the above, the researcher should ask “how” and “why” questions about your general topic. For example, “how does greenhouse gases emission contributes to global warming?” or “why does greenhouse gases emission contributes to global warming?”

e. Evaluate the question asked: After writing down the questions on paper, evaluate these questions to determine whether they would be effective research questions or whether they need more revising and refining.

   - Is the research question clear? With so much research available on any given topic, research questions must be as clear as possible in order to be effective in helping the writer direct his or her research.

- Is the research question focused? Research questions must be specific enough to be well covered in the space available.
- Is the research question complex? Research questions should not be answerable with a simple “yes” or “no” or by easily-found facts. They should, instead, require both research and analysis on the part of the writer. They often begin with “How” or “Why.”

f. Start the research**:** After coming up with a question, think about the possible paths the research could take. What sources should be consult as a researcher seeks answers to the question/questions? What research process will ensure that a variety of perspectives and responses to the question / questions are found?

3.5.2 Development of hypothesis

A hypothesis is a tentative answer to a research problem that is advanced so that it can be tested. The definition of a hypothesis stresses that it can be tested. To meet this criterion, the concepts employed in the hypothesis must be measurable. Developing hypotheses requires that a researcher identifies one variable that causes, affects, or has an influence on, another variable. The variable that affects other variables is called the independent /regressor/explanatory variable, while the variable which is explained by the independent variable is called the dependent/regressand/ response variable.

After extensive literature review, researcher should state in clear terms the research hypothesis or hypotheses. Research hypothesis is tentative assumption made in order to draw out and test its logical or empirical consequences. As such the way in which research hypotheses are developed is particularly important since they provide the central point for research. They also affect the way in which tests must be conducted in the analysis of data and indirectly the quality of data which is required for the analysis. In economic research, the development of research hypothesis plays an important role. Hypothesis should be very specific and limited to the piece of research in hand because it has to be tested. It should also be simple, and conceptually clear. There is no place for ambiguity in the construction of a hypothesis, as ambiguity will make the verification of the hypothesis almost impossible. It should be ‘unidimensional’ – that is, it should test only one relationship at a time.

The role of the hypothesis is to guide the researcher by delimiting the area of research and to keep him/her on the right track. It sharpens his/her thinking and focuses attention on the more important aspects of the problem. It also indicates the type of data required and the type of methods of data analysis to be used.

A research hypothesis being a predictive statement, it is capable of being tested by scientific methods, that relates an independent variable to some dependent variable. For example, “Government expenditure causes economic growth.”

In statistical terms we have alternative hypothesis (H_1) which the research wishes to prove and the null hypothesis (Ho) is the one which the research wishes to disprove. Thus, a null hypothesis represents the hypothesis to be rejected, and alternative hypothesis represents the hypothesis to be accepted, for example;

H_O: Government expenditure does not cause economic growth.

H_1: Government expenditure causes economic growth.

Characteristics of Hypothesis.

Hypothesis should be:

a. Clear and precise.

b. Capable of being tested.

c. Able to state relationship between variables, if it happens to be a relational hypothesis.

d. Limited in scope and must be specific.

e. Stated in simple terms.

f. Consistent with most known facts i.e. one which judges accept as being the most likely.

g. Testable within a reasonable time for one cannot spend a life-time collecting data to test it.

Steps in Developing Research Hypotheses

a. Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution.
b. Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues of the research problem.
c. Review of similar studies in the area or of the studies on similar problems.
d. Exploratory personal investigation which involves original field interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

Therefore, research hypotheses arise as a result of a-priori thinking about the subject, examination of the available data and material including related studies and the counsel of experts and interested parties. Research hypotheses are more useful when stated in precise and clearly defined terms.

Errors in research

When a researcher tests hypothesis in research two types of errors may occur in his/her research procedures. These are Type I error and Type II error.

Type I error: If the null hypothesis of a research is true, but the researcher takes decision to reject it; then an error must occur, it is called Type I error (false positives). It occurs when the researcher concludes that there is a statistically significant difference when in reality it does not exists. This is an analogy of a test that shows a patient to have a disease when in fact the patient does not have the disease. Imagine the consequences of Type I error of a HIV test on a patient promising a 99% accuracy rate.

Type II error: If the null hypothesis of a research is actually false, and the alternative hypothesis is true. The researcher decides not to reject the null hypothesis, then he is said to commit a Type II error (false negatives). For example, a blood test failing to detect the disease it was designed to detect in a patient who really has the disease is a Type II error. Neyman and Pearson, (1928) were the first to both Types I and II errors (Mohajan, 2017). The Type I error is more serious than Type II, because a researcher has wrongly rejected the null hypothesis. Both Type I and Type II errors are factors that every researcher must take into account.

We have observed that a research is error free in the two cases:

i. If the null hypothesis is true and the decision is made to accept it, and

ii. If the null hypothesis is false and the decision is made to reject it.

SELF ASSESSMENT EXERCISE

Briefly discuss the guidelines involved in setting a research question.

4.0 CONCLUSION

In this unit, we conclude that some authors have classified philosophical foundation of research into three categories mainly; realism, constructivism and pragmatism. We also conclude that a research problem is an issue being addressed in a research study, it helps in narrowing the topic down to something that is reasonable for conducting a study; and hypothesis is a tentative answer to a research problem that is advanced so that it can be tested.

5.0 SUMMARY

This unit discussed the philosophical and conceptual foundation of research, identification and formulation of research problem, meanings of research question and hypothesis and the steps involved in their development; and lastly, the various types of errors encountered in research process.

6.0 TUTOR-MARKED ASSIGNMENT

1. Mention and explain the procedure of formulating research problem
2. List and explain the types of errors a researcher may come across in research process.
3. How is a research question developed?

7.0 REFERENCES/FURTHER READINGS

Allchin, D. (2001). Error types. *Perspectives on Science*, *9*(1), 38-58.

Asika, N. (1991). *Research Methodology in the Behavioural Sciences.* Lagos, Nigeria: Longman Nigeria.

Bajpai, S. R., & Bajpai, R. C. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health*, *3*(2), 112-115.

Mohajan (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University*, *17*(3): 58-82

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20*A(1/2), 175-240.

Pandey, P., & Pandy, M.M. (2015). *Research Methodology: Tools and Techniques*. Romania: Bridge Center.

MODULE TWO: DATA: TYPES AND COLLECTION METHODS

Unit 1: Data: Types and manipulation
Unit 2: Methods Collecting Research Data

UNIT ONE: DATA: TYPES AND MANIPULATION

CONTENT

1.0 Introduction

2.0 Objectives

3.0 Main Contents

CONTENTS

1.0 Introduction
2.0 Objectives
3.0 Main Content
3.1 Data Types
3.1.1 Time series data
3.1.2 Cross section data
3.1.3 Panel data
3.1.4 Primary and Secondary data
3.1.5 Quantitative and qualitative data
3.2 Data Manipulation / Transformation
3.2.1 Square transformation
3.2.2 Log transformation
3.2.3 Level and growth rates
3.2.4 Index numbers
3.2.5 Reciprocal transformation
3.2.6 Nominal and real variables
4.0 Conclusion
5.0 Summary
6.0. Tutor-Marked Assignment
7.0 References/Further Readings

3.0 MAIN CONTNT

3.1 Data types

The data used in Economics may be of three types; Time series, Cross-section and Panel.

3.1.1 Time series data

This is data collected at specific points in time. It is a set of observations on the values that a variable takes **at different times**. Such data may be collected at regular time intervals such as daily (like stock prices, whether reports etc.), weekly (like money supply figures), monthly (like consumer price index etc.) quarterly (like GDP), annually (like government budget etc.) or biannually etc. Financial data measures phenomena such as changes in the price of stocks. This type of data is collected more frequently than the above, for instance, daily or even hourly. In all of these examples, the data are ordered by time and are referred to as **time series** data.

In economics, we commonly use the notation Y_t to indicate an observation on variable Y (e.g. real GDP) at time t. A series of data runs from period $t = 1$ to $t = T$. "T" is used to indicate the total number of time periods covered in a data set. To give an example, if we were to use annual real GDP data from 1980–2017 – a period of 38 years – then $t = 1$ would indicate 1980 up to $t = 38$ for 2017 and $T = 38$ the total number of years. Hence, $Y1$ would be real GDP in 1980, $Y2$ real GDP for 1981, etc. Time series data is typically presented in **chronological order**. With the advent of high-speed computers, data can now be collected over an extremely short interval of time, such as the data on stock prices, which can be obtained literally continuously (the so-called *real-time quote*).

Although time series data are used heavily in economics, they present special problems for econometricians. Most empirical work based on time series data assumes that the underlying time series is **stationary** and this is not the case with most of the time series data**.** Loosely speaking a time series is stationary if its mean and variance do not vary systematically over time.

3.1.2 Cross section data

Cross-section data are data on one or more variables collected at the **same point of time**. A common example is census of population, data on the wage of all people in a certain company or industry. We use the notation Y_i to indicate an observation on variable Y for individual i. Observations in a cross-sectional data set run from individual i = 1 to N. By convention, N indicates the number of cross-sectional units (e.g. the number of people surveyed). Microeconomist may ask N = 100 managers from manufacturing companies about their profit figures in the last month. In this case, Y_1 will equal the profit reported by the first company, Y_2 the profit reported by the second company, through to Y100, the profit reported by the 100th company. With **cross-sectional** data, the ordering of the data typically does not matter (unlike time series data). Just as time series data create their own special problems (because of the stationarity issue), cross-sectional data too have their own problems, specifically the problem of *heterogeneity*.

3.1.3 Panel data

Some data sets will have both a time series and a cross-sectional component. This data is referred to as **panel** data. Economists working on issues related to growth often make use of panel data. For instance, GDP for many countries from 1980 to the present is available. A panel data set on Y = GDP for 12 African countries would contain the GDP value for each country in 1980 (N = 12 observations), followed by the GDP for each country in 1951 (another N = 12 observations), and so on. Over a period of T years, there would be $T \setminus N$ observations on Y. We use the notation Y_{it} to indicate an observation on variable Y for unit i at time t. In the economic growth example, Y_{11} will be GDP in country 1, year 1, Y_{12} GDP for country 1 in year 2, etc. **Longitudinal or Micro Panel Data** is a special type of pooled data in which the same cross-sectional (say a family or firm) is surveyed overtime.

3.1.4 Primary and secondary data

Research data is often classified as primary and secondary data or quantitative and qualitative data. Primary data is the original data that are collected for the specific research problem at hand. It is collected with the aim of being foundation for the analyses in your investigation. Researcher collects primary data using experiments and surveys. Secondary data on the other hand is data already collected or produced by others. Collection of secondary data is appropriate for laying the foundation for your own work; in attempt to know how, things are done by others or how others interpreted situations or occurrences; and when it is impossible to get primary data.

3.1.5 Quantitative and qualitative data

Quantitative data involves expressing data *statistically* or *numerically*. Such data is usually expressed in one of three ways:

i. **Numbers** (sometimes called *raw numbers*). For example, the total number of people who live in poverty.

ii. **Percentages:** the number *per 100* in a population. For example, 30% of Nigeria's labour force is unemployed.

iii. **Rates**: the number *per 100, 1000 etc* in a population. For example, birth rate, inflation rate etc. Research data is often expressed as a *rate* or *percentage* because it allows accurate comparisons *between* and *within* groups and societies. For example, comparing unemployment between Britain and America as a *raw* number wouldn't tell us very much, since the population of America is roughly 5 times larger. Expressing unemployment as a *percentage* or *rate* on the other hand allows us to compare "like with like". Quantitative data lends itself to statistical comparisons and gives researcher the ability to express relationships *numerically*. This is useful for things like hypothesis testing or *cross-cultural* comparisons.

Although the ability to quantify the social world can be a significant advantage for social researchers one of the major criticisms of quantification relates to the validity of the data collected. Quantitative methods only capture a relatively narrow range of data about people's behaviour. It tells us very little about the *reasons* for people's behaviour. This is partly a problem of lack of depth: the more complex the behaviour, the more difficult it is to quantify.

Qualitative data tries to capture the quality of people's behaviour and such data says something about how people experience the social world. It can be used to understand the meanings they give to behaviour. Where a research objective is to understand the meaning of people's behaviour, they must be allowed the scope to talk freely. Qualitative data encourages this because a researcher doesn't impose their interpretation on a situation (by asking direct, quantifiable, questions for example). Qualitative data provides greater depth and detail about behaviour since they are likely to involve digging more deeply into people’s beliefs and behaviours.

Where qualitative research generally focuses on the intensive study of relatively small groups, opportunities to generalise research data may be limited. For similar reasons it’s difficult to compare qualitative research across time and space because researchers are unlikely to be comparing "like with like". Qualitative data also tends to be structured in ways that make the research difficult to replicate.

Although we’ve considered quantitative and qualitative data as separate entities, there are occasions when a researcher may want to combine quantitative and qualitative types of data. This mixed method is called methodological triangulation and is one that can also be used to improve both research validity - by creating a more-accurate measurement of something - and

reliability by using the strengths of one type of data (the ability to quantify behaviour, for example) to offset the weaknesses of the other.

SELF ASSESSMENT EXCERCISE

Discuss the types of data used in economic research. Differentiate between quantitative and qualitative data.

3.2 Data manipulation / Transformation

Often times the initial form of your data is not the way you want it for analysis. It is part of the analyst’s role to search out different ways of looking at the data in order to enhance our understanding of that data. One of the most powerful tools available to help achieve this task is data transformation or manipulation. Before we delve into data transformation, it is imperative to differentiate between **data and data sets.**

The information that you collect from an experiment, survey, or archival source is referred to as your data. Most generally, data can be defined as list of numerical and/or categorical values possessing meaningful relationships. For analysts to do anything with a group of data they must first translate it into a *data set*. A dataset is a representation of data, defining a set of “variables” that are measured on a set of “cases.” Like variables, a data set will typically contain multiple cases. Cases are also sometimes referred to as *observations*. The object “type” that defines your cases is called your *unit of analysis*. Sometimes the unit of analysis in a data set will be very small and specific, such as the individual responses on a questionnaire. Sometimes it will be very large, such as companies or nations. When describing a data set you should always provide definitions for your variables and the unit of analysis. You typically would not list the specific cases, although you might describe their general characteristics. Many different data sets can be constructed from the same data. Different data sets could contain different variables and possibly even different cases. *Data manipulation* is the procedure of creating a new data set from an existing data set. In almost every study you will need to alter your initial data set in some way before you can begin analysis. The different ways that you can change your data set can be grouped into two general categories.

i. Changes that involve calculating new variables as a function of one or more old variables in your data set are called *transformations*. The new data set will typically have all of the original variables, with the addition of one or more new variables. Sometimes a transformation will simply involve changing the values of an existing variable. After performing a transformation, the cases of the new data set will be exactly the same as those of the old data set.

ii. If you alter your data set in such a way that you end up changing the unit of analysis you are performing *data restructuring*. The new data set will typically use entirely new variables, with maybe a small number that are the same as in the original data set. Additionally, your new data set will be composed of entirely new cases. Restructuring a data set is typically a more difficult and involved procedure than simply transforming variables.

We now examine the following forms of data manipulation.

3.2.1 Square transformation

The square transformation has the effect of decreasing values less than 1, and increasing values greater than 1. Large values are increased the most. For example, 2^2 = 4, while 20^2 = 400, so that while the values 2 and 20 are 18 units apart, the values 4 and 400 are 396 units apart. That is, the effect of the square transformation is to **stretch** the values.

3.2.2 Log transformation

It reduces all values, and values between 0 and 1 become negative. Large values are reduced much more than small values. For example, log 2 = 0.301, while log 20 = 1.301, so that while the values 2 and 20 are 18 units apart, the values 0.301 and 1.303 are only 1 unit apart. That is, the effect of the log transformation is to **compress** the values. Note that the log function can only be applied to values which are greater than 0. One particularly common transformation is the logarithmic one. The logarithm (to the base *B*) of a number, *A*, is the power to which *B* must be raised to give *A*. The notation for this is: log*B*(*A*). So, for instance, if *B* = 10 and *A* = 100 then the logarithm is 2 and we write log10(100) = 2. This follows since 102 = 100. In economics, it is common to work with the so-called natural logarithm which has *B* = *e* where *e* = 2.71828. The natural logarithm operator is denoted by ln; i.e. ln(*A*) = log*e*(*A*). It is common to use ln(*Y*) as the dependent variable and ln(*X*) as the explanatory variable. First, the expressions will often allow us to interpret results quite easily. Our background in calculus will help us to fully understand this point. Second, data transformed in this way often does appear to satisfy the linearity assumption of the regression model.

Given that lnY = a + blnX b can be interpreted as an **elasticity**. Recall that, in the basic regression without logs, we said that "*Y* tends to change by b **units** for a one **unit** change in *X* ". In the regression containing both logged dependent and explanatory variables, we can now say that "*Y* tends to change by b **percent** for a one **percent** change in *X* ". That is, instead of having to worry about units of measurements, regression results using logged variables are always interpreted as elasticities.

Logs are convenient for other reasons too.

For instance, when we have time series data, the percentage change in a variable is approximately 100 \ [ln(*Yt*) - ln(*Yt* -1)].

The second justification for the log transformation is purely practical: With many data sets, if you take the logs of dependent and explanatory variables and make an *XY*-plot the resulting relationship will look linear. Although there is no simple rule that can be given.

3.2.3 Levels and growth rates

Conversion of data from its raw form mostly at level to growth rate is one of the common transformation that econometricians use with time series data.

Suppose we have annual data on real GDP for 1970–2018 (i.e. 49 years of data) denoted by *Yt* , for *t* = 1 to 49. In many empirical projects, this might be the variable of primary interest. We will refer to such series as the **level** of real GDP. However, researchers are often more interested in how the economy is growing over time, or in real GDP **growth**. A simple way to measure growth is to take the real GDP series and calculate a percentage change for each year. The percentage change in real GDP between period *t* and *t* + 1 is calculated according to the formula %change = (Yt+1 – Yt)/Yt *100. The percentage change in real GDP is often referred to as the **growth** of GDP or the **change** in GDP.

3.2.4 Index numbers

Many variables that economists work with come in the form of index numbers. Suppose you are interested in studying a country’s inflation rate, which is a measure of how prices change over time. The question arises as to how we measure “prices” in a country. The price of an individual good (e.g. milk, oranges, electricity, a particular model of car, a pair of shoes, etc.) can be readily measured, but often interest centers not on individual goods, but on the price level of the country as a whole.

The latter concept is usually defined as the price of a “basket” containing the sorts of goods that a typical consumer might buy. The price of this basket is observed at regular intervals over time in order to determine how prices are changing in the country as a whole. But the price of the basket is usually not directly reported by the government agency that collects such data.

To interpret this latter number, you would have to know what precisely was in the basket and in what quantities. Given the millions of goods bought and sold in a modern economy, far too much information would have to be given.

In light of such issues, data often comes in the form of a price index. Indices may be calculated in many ways, and it would distract from the main focus of this course. However, the following points are worth noting at the outset. Firstly, indices almost invariably come as time series data. Secondly, one-time period is usually chosen as a base year and the price level in the base year is set to 100 (some indices set the base year value to 1.00 instead of 100). Thirdly, price levels in other years are measured in percentages relative to the base year.

A price index is very good for measuring changes in prices over time, but should not be used to talk about the level of prices. For instance, it should not be interpreted as an indicator of whether prices are “high” or “low”. Other indices are the stock index, wage index etc.

3.2.5 Reciprocal transformation

Again reduces all values greater than one. Large values are reduced much more than small values. For example, 1/2 = 0.5, while 1/20 = 0.05, so that while the values 2 and 20 are 18 units apart, the values 0.5 and 0.05 are only 0.45 units apart. That is, the effect of the reciprocal transformation is to **compress** the large values to an even greater extent than the log transformation.

3.2.6 Nominal and real variables

This leads researchers to want to correct for the effect of inflation. The way to do this is to divide the GDP measure by a price index (in the case of GDP, the name given to the price index is the GDP deflator). GDP transformed in this way is called real GDP. The original GDP variable is referred to as nominal GDP. This distinction between real and nominal variables is important in many fields of economics. The key things you should remember are that a real variable is a nominal variable divided by a price variable (usually a price index) and that real variables have the effects of inflation removed from them. The case where you wish to correct a growth rate for inflation is slightly different. In this case, creating the real variable involves subtracting the change in the price index from the nominal variable. So, for instance, real interest rates are nominal interest rates minus inflation (where inflation is defined as the change in the price index).

4.0 CONCLUSION

In Economic analysis, three major types of research data are distinguishable. These the time series, cross sections and the panel data. Other types may fall within the three major types. Such data at times cannot be used in its raw form. It therefore needs to be transformed for easy analysis and attainment of the research objectives. Sometimes new variables can easily be generated from the already existing data.

5.0 SUMMARY

The unit discussed the meaning of data and the various types of data; data analysis and interpretation, discussion of research results; and identifies the various transformation of data.

TUTOR-MARKED ASSIGNMENT

1. List the various types of data.
2. Explain the various forms of data transformation.

7.0 REFERENCES/FURTHER READINGS

DeCoster, J. (2001). Transforming and Restructuring Data. Retrieved <September, 12, 2017 downloaded this file> from http://www.stat-help.com/notes.html

Dimitrios, A., & Stephen G. H. (2007). *Applied Econometrics: A Modern Approach*. NY: Palgrave Macmillan

Jones, P., Evans, M., & Lipson, K. (2008). *Essential Further Mathematics – Core* (4th ed.). Cambridge University.

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

Kumar, R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners* (3rd ed.). New Delhi: SAGE.

Walliman, N. (2011). *Research Methods: The Basics*. London: Routledge.

Gujarati, D. (2004). Basic Econometrics (4th ed.). NY: McGraw-Hill.

UNIT TWO: METHODS OF COLLECTING RESEARCH DATA

CONTENTS

1.0 INTRODUCTION

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The goal for data collection is to capture quality evidence that then translates to rich data analysis and allows the building of a convincing and credible answer to questions that have been posed. It is one of the most important stages in conducting a research. The task of data collection begins after a research problem has been defined and research design/plan drawn out. While deciding about the method of data collection to be used in a study, two types of data should be kept in mind, primary and secondary.

2.0 OBJECTIVES

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

- Know the various sources of primary data; their advantages and disadvantages.
- Know the advantages and disadvantages of secondary data.
- Know the sources of secondary data.

3.0 MAIN CONTENT

3.1 Methods of collecting primary data

The primary data are those which are collected directly from the source, and so are original in nature. This type of data has not been published yet and is more reliable, authentic and objective. Primary data can provide information about virtually any aspect of our life and surroundings. However, collecting primary data is time consuming and not always possible. Although more data usually means more reliability, it is costly to organize large surveys and other studies. Furthermore, it is not always possible to get direct access to the subject of research. For example, many historical events have left no direct evidence. The Important ways of collecting data in surveys and descriptive researches are:

i. observation method (ii) interview method (iii) through questionnaires (i) Interview method, (ii) Focus group method

3.1.1 Observation method

Observation is a purposeful, systematic and selective way of watching and listening to an interaction or phenomenon as it takes place. There are many situations in which observation is the most appropriate method of data collection; for example, when you want to learn about the interaction in a group, ascertain the functions performed by a worker, or study the behaviour or personality traits of an individual. It is also appropriate in situations where full and/or accurate information cannot be elicited by questioning, because respondents either are not co-operative or are unaware of the answers because it is difficult for them to detach themselves from the interaction. In summary, when you are more interested in the behavior than in the perceptions of individuals, or when subjects are so involved in the interaction that they are unable to provide objective information about it, observation is the best approach to collect the required information. There are two types of observation participant and non-participant observation.

i. **Participant observation** is when you, as a researcher, participate in the activities of the group being observed in the same manner as its members, with or without their knowing that they are being observed. For example, you might want to study the life of prisoners and pretend to be a prisoner in order to do this.

ii. **Non-participant observation**, on the other hand, is when you, as a researcher, do not get involved in the activities of the group but remain a passive observer, watching and listening to its activities and drawing conclusions from this. For example, you might want to study the functions carried out by brokers in a stock market. As an observer, you could watch, follow and record their activities as they are performed. After making a number of observations, conclusions could be drawn about those functions. Any occupational group in any setting can be observed in the same manner.

Advantages of Observation Method

a. Subjective bias is eliminated, if observation is done accurately.

b. Information obtained under this method relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes.

c. It is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happens to be the case in the interview or the questionnaire method.

d. It is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other.

Disadvantages of observation method

a. It is an expensive method.

b. Information provided by this method is very limited.

c. Observer’s bias may invalidate the observation.

d. It is time consuming and requires much effort.

3.1.2 Interview Method

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews. In social research there are many types of interview. The most common of these are structured and unstructured interviews. **Structured interviews** involve the use of a set of predetermined questions and of highly standardized techniques of recording. Thus, the interviewer in a structured interview follows a rigid procedure laid down, asking questions in a form and order prescribed. As against it, the unstructured interviews are characterized by a flexibility of approach to questioning. **Unstructured interviews** do not follow a system of pre-determined questions and standardized techniques of recording information. In this type of interview, the researcher attempts to achieve a holistic understanding of the interviewees’ point of view or situation. In a non-structured interview, the interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires. He may even change the sequence of questions. He has relatively greater freedom while recording the responses to include some aspects and exclude others. But this sort of flexibility results in lack of comparability of one interview with another and the analysis of unstructured responses becomes much more difficult and time-consuming than that of the structured responses obtained in case of structured interviews. Unstructured interviews also demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, however, happens to be the central technique of collecting information in case of exploratory research studies. But in case of descriptive studies, we quite often use the technique of structured interview because of its being more economical, providing a safe basis for generalization and requiring relatively lesser skill on the part of the interviewer.

Advantages of Interview Method

Despite the variations in interview-techniques, the major advantages of the interview method are as follows:

a. More information can be obtained.

b. Interviewer by his own skill can overcome the resistance, if any, of the respondents; the interview method can be made to yield an almost perfect sample of the general population.

c. There is greater flexibility under this method as the opportunity to restructure questions is always there, especially in case of unstructured interviews.

d. Observation method can as well be applied to recording verbal answers to various questions.

e. Personal information can as well be obtained easily under this method.

f. The interviewer may catch the informant off-guard and therefore may secure the most spontaneous reactions than would be the case if mailed questionnaire is used.

g. The language of the interview can be adopted to the ability or educational level of the person interviewed and as such misinterpretations concerning questions can be avoided.

h. The interviewer can collect supplementary information about the respondent's personal characteristics and environment which is often of great value in interpreting results.

Disadvantages of interview method

a. It is a very expensive method.

b. There is the possibility of the bias of interviewer as well as that of the respondent; there also remains the headache of supervision and control of interviewers.

c. Data may prove inadequate, certain types of respondents such as important officials or executives or people in high income groups may not be easily approachable.

d. It is relatively more-time-consuming.

e. The presence of the interviewer on the spot may over-stimulate the respondent, sometimes even to the extent that he may give imaginary information just to make the interview interesting.

f. Under the interview method the organisation required for selecting, training and supervising the field-staff is more complex with formidable problems.

g. It at times may also introduce systematic errors.

3.1.3 Collection of data through questionnaires

A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms. In this method a questionnaire is sent to the persons concerned with a request to answer the questions and return the questionnaire. The respondents are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own.

Questionnaire is considered as the heart of a survey operation. Hence it should be very carefully constructed. If it is not properly set up, then the survey is bound to fail. This fact requires us to study the main aspects of a questionnaire viz., the general form, question sequence and question formulation and wording.

1. *General form:* It can either be structured or unstructured questionnaire. Structured questionnaires are those questionnaires in which there are definite, concrete and pre-determined questions. The questions are presented with exactly the same wording and in the same order to all respondents. This is aimed at ensuring that all respondents reply to the same set of questions.

2. *Question sequence:* In order to make the questionnaire effective and to ensure quality to the replies received, a researcher should pay attention to the question-sequence in preparing

the questionnaire. A proper sequence of questions reduces considerably the chances of individual questions being misunderstood. The question-sequence must be clear and smoothly-moving, meaning thereby that the relation of one question to another should be readily apparent to the respondent, with questions that are easiest to answer being put in the beginning. The first few questions are particularly important because they are likely to influence the attitude of the respondent and in seeking his desired cooperation. The opening questions should be such as to arouse human interest. The following type of questions should generally be avoided as opening questions in a questionnaire:

- questions that put a strain on the memory or intellect of the respondent;
- questions of a personal character;
- questions related to personal wealth, etc.

Following the opening questions, we should have questions that are really vital to the research problem and a connecting thread should run through successive questions. Ideally, the question sequence should conform to the respondent’s way of thinking. Relatively difficult questions must be relegated towards the end so that even if the respondent decides not to answer such questions, considerable information would have already been obtained. Thus, question-sequence should usually go from the general to the more specific and the researcher must always remember that the answer to a given question is a function not only of the question itself, but of all previous questions as well.

3. *Question formulation and wording:* With regard to this aspect of questionnaire, the researcher should note that each question must be very clear for any sort of misunderstanding. Question should also be impartial in order not to give a biased picture of the true state of affairs. In general, all questions should meet the following standards

a. should be easily understood;

b. should be simple i.e., should convey only one thought at a time;

c. should be concrete and should conform as much as possible to the respondent’s way of thinking.

Concerning the form of questions, they may be either close-ended or open-ended (i.e., inviting free response). Close-ended questions include fixed response e.g. Yes-No, True-False, Multiple Choice, Rating Scale/Continuum (such as a Likert-type scale), Agree-Disagree, Rank ordering. The researcher must choose measurement scale and scoring that provide the information needed and are appropriate for respondents.In the close-ended, the respondent selects one of the alternative possible answers put to him, whereas in the open-ended he has to supply the answer in his own words.

There are some advantages and disadvantages of each possible form of question. Multiple choice or closed questions have the advantages of easy handling, simple to answer, quick and relatively inexpensive to analyse. They are most amenable to statistical analysis. Sometimes, the provision of alternative replies helps to make clear the meaning of the question. But the main drawback of fixed alternative questions is that of "putting answers in people's mouths" i.e., they may force a statement of opinion on an issue about which the respondent does not in fact have any opinion. They are not appropriate when the issue under consideration happens to be a complex one and also when the interest of the researcher is in the exploration of a process.

Getting the replies in respondent's own words is, thus, the major advantage of open-ended questions. But one should not forget that, from an analytical point of view, open-ended questions are more difficult to handle, raising problems of interpretation, comparability and interviewer bias.

Researcher must pay proper attention to the wordings of questions. Simple words, which are familiar to all respondents should be employed. Words with ambiguous meanings must be avoided. Similarly, danger words, catch-words or words with emotional connotations should be avoided. Caution must also be exercised in the use of phrases which reflect upon the prestige of the respondent. In fact, question wording and formulation is an art and can only be learnt by practice.

Other essentials of a good questionnaire:

- To be successful, questionnaire should be comparatively short and simple i.e., the size of the questionnaire should be kept to the minimum.
- Open-ended questions are often difficult to analyze and hence should be avoided in a questionnaire to the extent possible.
- There should be some control questions in the questionnaire which indicate the reliability of the respondent. For instance, a question designed to determine the consumption of particular material may be asked first in terms of financial expenditure and later in terms of weight. The control questions, thus, introduce a cross-check to see whether the information collected is correct or not.
- Questions affecting the sentiments of respondents should be avoided.
- Adequate space for answers should be provided in the questionnaire to help editing and tabulation.
- There should always be provision for indications of uncertainty, e.g., “do not know,” “no preference” and so on.
- Brief directions with regard to filling up the questionnaire should invariably be given in the questionnaire itself. Finally, the physical appearance of the questionnaire affects the cooperation the researcher receives from the recipients and as such an attractive looking questionnaire, particularly in mail surveys, is a plus point for enlisting cooperation.
- The quality of the paper, along with its colour, must be good so that it may attract the attention of recipients.
- To ensure that the survey instrument you develop is appropriate for your audience, “field test” your questionnaire with other people similar to your respondents before administering the final version. This will allow you to improve unclear questions or procedures and detect errors beforehand.

Advantages of Questionnaire Method

a. Large amounts of information can be collected from a large number of people in a short period of time and in a relatively cost effective way.

b. Can be carried out by the researcher or by any number of people with limited affect to its validity and reliability.

c. The results of the questionnaires can usually be quickly and easily quantified by either a researcher or through the use of a software package.

d. Can be analyzed more scientifically and objectively than other forms of research.

e. When data has been quantified, it can be used to compare and contrast other research and may be used to measure change.

Disadvantages of Questionnaire Method

a. To be inadequate to understand some forms of information - i.e. changes of emotions, behavior, feelings etc.

b. Phenomenologists state that quantitative research is simply an artificial creation by the researcher, as it is asking only a limited amount of information without explanation.

c. There is no way to tell how truthful a respondent is being.

d. There is no way of telling how much thought a respondent has put in.

e. The respondent may be forgetful or not thinking within the full context of the situation.

f. People may read differently into each question and therefore reply based on their own interpretation of the question - i.e. what is ‘good’ to someone may be ‘poor’ to someone else, therefore there is a level of subjectivity that is not acknowledged.

3.1.4 Problems of Primary Data Collection

Common problems associated with primary data collection could result in inaccurate and unreliable information. If this is fed into the analysis system, the final conclusions may be misleading and the recommendations inappropriate.

a. Interviewers lack knowledge or skills.

b. Information is incomplete or inaccurate.

c. Questionnaires or checklists neglect key issues.

d. Interviewers and informants are biased.

e. Interviewers and informants become bored.

f. Informants experience assessment fatigue.

3.2 Methods of Collecting Secondary data

Secondary data, on the other hand, are those which have already been collected by someone else and have already been passed through statistical process, this type of data has already been published. All research studies require secondary data for the background to the study. You will inevitably need to ascertain what the context of your research question/problem is, and also get an idea of the current theories and ideas. No type of project is done in a vacuum, not even a pure work of art. The quality of the data depends on the source and the methods of presentation. Secondary data is often readily available. After the expense of electronic media and internet the availability of secondary data has become much easier. There are varieties of published printed sources. Their credibility depends on many factors. For example, on the writer, publishing company and time and date when published. New sources are preferred and old sources should be avoided as new technology and researches bring new facts into light. The following are some of its sources:

3.2.1 Books

Books are available today on any topic that you want to research. The use of books start before even you have selected the topic. After selection of topics books provide insight on how much work has already been done on the same topic and you can prepare your literature review. Books are secondary source but most authentic one in secondary sources.

3.2.2 Journals/periodicals

Journals and periodicals are becoming more important as far as data collection is concerned. The reason is that journals provide up-to-date information which at times books cannot and secondly, journals can give information on the very specific topic on which you are researching rather talking about more general topics.

Magazines are also effective but not very reliable. Newspapers on the other hand are more reliable and in some cases the information can only be obtained from newspapers as in the case of some political studies.

3.2.3 Electronic Sources

As internet is becoming more advance, fast and reachable to the masses; it has been seen that much information that is not available in printed form is available on internet. In the past the credibility of internet was questionable but today it is not. The reason is that in the past journals and books were seldom published on internet but today almost every journal and book is available online. Some are free and for others you have to pay the price.

e-journals are more commonly available than printed journals. Latest journals are difficult to retrieve without subscription but if your university has an e-library you can view any journal, print it and those that are not available you can make an order for them.

3.2.4 Internet

Generally, websites do not contain very reliable information so their content should be checked for the reliability before quoting from them. Weblogs are also becoming common. They are actually diaries written by different people. These diaries are as reliable to use as personal written diaries.

3.2.5 Unpublished Personal Records: Some unpublished data may also be useful in some cases.

a. **Diaries:** Diaries are personal records and are rarely available but if you are conducting a descriptive research then they might be very useful. The Anne Frank's diary is the most famous example of this. That diary contained the most accurate records of Nazi wars.

b. Letters: Letters like diaries are also a rich source but should be checked for their reliability before using them.

3.2.6 Government Records

Government records are very important for marketing, management, humanities and social science research.

3.3 Scrutiny of secondary data sources

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study. By way of caution, the researcher, before using secondary data, must see that they possess following characteristics:

1. Reliability of data: The reliability can be tested by finding out such things about the said data:

(a) Who collected the data? (b) What were the sources of data? (c) Were they collected by using proper methods (d) At what time were they collected? (e) Was there any bias of the compiler? (t) What level of accuracy was desired? Was it achieved?

2. Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not be used by the researcher. In this context, the researcher must very carefully scrutinize the definition of various terms and units of collection used at the time of collecting the data from the primary source originally. Similarly, the object, scope and nature of the original enquiry must also be studied. If the researcher finds differences in these, the data will remain unsuitable for the present enquiry and should not be used.

3. Adequacy of data: If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher. The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry.

From all this we can say that it is very risky to use the already available data. The already available data should be used by the researcher only when he finds them reliable, suitable and adequate. But he should not blindly discard the use of such data if they are readily available from authentic sources and are also suitable and adequate for in that case it will not be economical to spend time and energy in field surveys for collecting information. Secondary data is used with due precaution.

3.3.1 Problems of Secondary Data Collections

a. Availability of data: because the data were not collected to answer a specific research question, particular information that a researcher would like to have may not have been collected. Or it may not have been collected in the geographic region of study or chosen time frame. In any case, a researcher can only work with the data that is available, not what he/she wishes had been collected and that the researcher will be allowed to access the data required.

b. Reliability and validity of data: not all sources of published data can be relied upon for the fact that some sources publish genuine data. For this reason, a secondary data set should be examined carefully to confirm that it includes the necessary data from a genuine source. The researcher did not participate in the planning and execution of the data collection process; he or she does not know exactly how it was done. More to the point, the researcher does not know how well it was done. Also some sources of data are not officially recognized.

SELF ASSESSMENT EXCERCISE

Discuss the advantages and disadvantages of questionnaire as a source of primary data.

List and explain the disadvantages of secondary data.

4.0 CONCLUSION

The unit concludes that the basic methods of collecting primary data include of interview method, observation method and questionnaire method. While methods of collecting secondary data includes; books, periodicals, electronic sources, internet, unpublished personal records, and government records

5.0 SUMMARY

The unit discussed the methods of collecting primary and secondary data, their advantages and disadvantages.

6.0. TUTOR-MARKED ASSIGNMENT

1. List the advantages and disadvantages of primary data.
2. State the various sources of secondary data.
3. State the advantages and disadvantages of observation method.

7.0 REFERENCES/FURTHER READINGS

Joshua, O. (2013). *The Essentials of Research Methodology and Statistics in Education.* Jigbik.

Kumar, R. (2011). *Research Methodology: A Step-by-Step Guide for Beginners* (3rd ed.). New Delhi: SAGE.

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

MODULE THREE: BASICS OF MODEL BUILDING IN ECONOMICS

UNIT ONE: VARIABLE TYPES

CONTENTS

1.0 INTRODUCTION

The initial step in conducting research is to articulate a testable hypothesis on which a conclusion can be drawn. A testable hypothesis is important because it denotes the direction of the research by explicitly stating the relationship between the variables of the study. A variable is a construct or concept to which numerical values can be attached. It is a concept which can take on different values. For example, height, weight, income. There is the need to identify the different variables so as to know how to manipulate them and obtain desired result in a research.

2.0 OBJECTIVES

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

1. Understand the various types of variables used in economic research.
2. Know the significance of dependent and independent variables in economic research.
3. Identify the role of other variables in economic research.

3.0 MAIN CONTENT

Data are generated when you isolate and manipulate one or more variables that supply the causes (independent variables) and observe the effects of this manipulation on variables that are affected by the causes (dependent variables). Thus, variables can be broadly classified as dependent (DV) and independent variables (IV).

3.1 Dependent variable

The dependent variable is the variable of primary interest to the researcher. The researcher's goal is to understand and describe the dependent variable, or to explain its variability, or predict it. In other words, it is the main variable that lends itself for investigation as a viable factor. Through the analysis of the dependent variable (i.e., finding what variables influence it), it is possible to find answers or solutions to the problem. For this purpose, the researcher will be interested in quantifying and measuring the dependent variable, as well as the other variables that influence this variable.

3.2 Independent variable

Independent Variables are defined as a characteristic that a researcher manipulate to identify a particular factor. Independent variables are also known as factor or prediction variable. It is generally conjectured that an independent variable is one that influences the dependent variable in either a positive or negative way. That is, when the independent variable is present, the dependent variable is also present, and with each unit of increase in the independent variable, there is an increase or decrease in the dependent variable. In other words, the variance in the dependent variable is accounted for by the independent variable.

In a study containing more than one independent variable, variables can start interacting with each other, giving rise to complex behavior. For example, if we have two independent variables, the two variables will interact with each other and produce different results than expected. Hence, if the number of independent variables is large, it gets difficult to reach the conclusion. To establish that a change in the independent variable causes a change in the dependent variable, all four of the following conditions should be met:

i. The independent and the dependent variable should covary: in other words, a change in the dependent variable should be associated with a change in the independent variable.

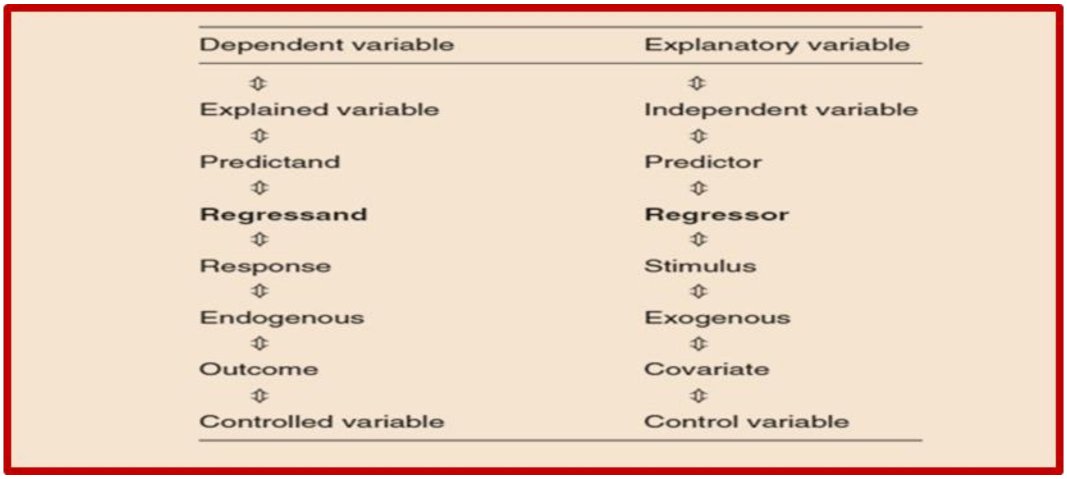
ii. The independent variable (the presumed causal factor) should precede the dependent variable. In other words, there must be a time sequence in which the two occur: the cause must occur before the effect.

iii. No other factor should be a possible cause of the change in the dependent variable. Hence, the researcher should control for the effects of other variables.

iv. A logical explanation (a theory) is needed about why the independent variable affects the dependent variable.

The terms dependent variable and explanatory variable are described variously. A representative as depicted in figure 3.1:

Figure 3.1: Terminologies of dependent and independent variables



Dependent variable	Explanatory variable
⇕	⇕
Explained variable	Independent variable
⇕	⇕
Predictand	Predictor
⇕	⇕
Regressand	Regressor
⇕	⇕
Response	Stimulus
⇕	⇕
Endogenous	Exogenous
⇕	⇕
Outcome	Covariate
⇕	⇕
Controlled variable	Control variable

Source: Adopted from Gujarati (2004)

3.3 Other types of variable

Dependent or independent variable can take any of the following form depending on its scale of measurement:

3.3.1 Binary variable: Observations (i.e., dependent variables) that occur in one of two possible states, often labelled zero and one. E.g., "improved/not improved" and "completed task/failed to complete task,"

3.3.2 Categorical Variable: Usually an independent or predictor variable that contains values indicating membership in one of several possible categories. E.g., gender (male or female), marital status (married, single, divorced, widowed). The categories are often assigned numerical values used as labels, e.g., 0 = male; 1 = female. Synonym for nominal variable.

3.3.3 Continuous variable: A variable that is not restricted to particular values (other than limited by the accuracy of the measuring instrument). E.g., temperature, IQ etc.

3.3.4 Discrete variable: Variable having only integer values. For example, number of students in a class, number of cars in a park etc.

3.3.5 Dummy Variables: Created by recoding categorical variables that have more than two categories into a series of binary variables. E.g., Marital status, if originally labelled 1=married, 2=single, and 3=divorced, widowed, or separated, could be redefined in terms of two variables as follows: var_1: 1=single, 0=otherwise. Var_2: 1=divorced, widowed, or separated, 0=otherwise. For a married person, both var_1 and var_2 would be zero. In general, a categorical variable with k categories would be recoded in terms of k - 1 dummy variables. Dummy variables are used in regression analysis to avoid the unreasonable assumption that the original numerical codes for the categories, i.e., the values 1, 2, ..., k, correspond to an interval scale. Use: to place cases in specific groups.

3.3.6 Intervening variable: A variable that explains a relation or provides a causal link between other variables. Also called by some authors "mediating variable" or "intermediary variable." Example: The statistical association between income and longevity needs to be explained because just having money does not make one live longer. Other variables intervene between money and long life. People with high incomes tend to have better medical care than those with low incomes. Medical care is an intervening variable. It mediates the relation between income and longevity.

3.3.7 Latent variable: An underlying variable that cannot be observed. It is hypothesized to exist in order to explain other variables, such as specific behaviors, that can be observed. Example: openness of economy, trustworthiness, conservatism, liberalism.

3.3.8 Manifest variable: An observed variable assumed to indicate the presence of a latent variable. Also known as an indicator variable. We cannot observe intelligence directly, for it is a latent variable. We can look at indicators such as vocabulary size, success in one's occupation, IQ test score, ability to play complicated games and so on.

3.3.9 Moderating variable: A variable that influences, or moderates, the relation between two other variables and thus produces an interaction effect.

3.3.10 Polychotomous variables: Variables that can have more than two possible values. Strictly speaking, this includes all but binary variables. The usual reference is to categorical variables with more than two categories.

SELF ASSESSMENT EXCERCISE

1. With relevant examples, explain any five types of variable.
2. Distinguish between dependent and dependent variables

4.0 CONCLUSION

Variables can be broadly classified as dependent and independent variables. The two are of primary interest to the researcher. Other types of variables are distinguishable though they can fall into the broad category of dependent and independent variables subject upon their usage.

5.0 SUMMARY

Variables can be broadly classified as dependent and independent variables. Others are binary, categorical, continuous, discrete, dummy, intervening, latent, manifest, moderating and polycotomous.

6.0 TUTOR MARKED ASSIGNMENTS

3. With relevant examples, explain any five types of variable.
4. Distinguish between dependent and dependent variables

7.0 REFERENCES/FURTHER READINGS

Gujarati, D. (2004). Basic Econometrics (4th ed.). NY: McGraw-Hill.

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business: A Skill Building Approach* (5th ed.). West Sussex, UK: John Wiley and Sons.

UNIT TWO: MEASUREMENT OF VARIABLES

CONTENTS

1.0 INTRODUCTION

Measurement is the systematic, replicable process by which variables are quantified and/or classified.

2.0 OBJECTIVES

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

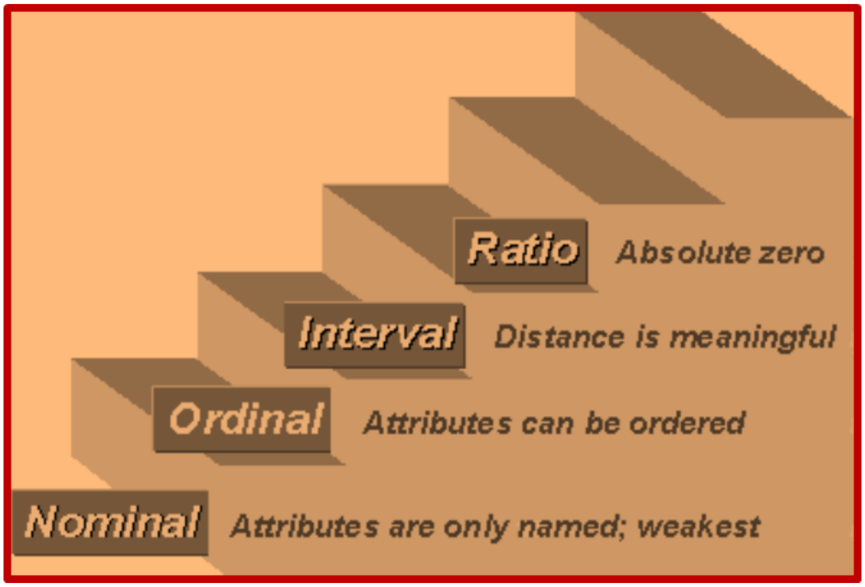
- Understand the various scales of measuring variables used in economic research.
- Know the common errors in measuring variables and their sources.
- Distinguish between the different types of reliability and validity of measurement instruments.

3.0 MAIN CONTENT

3.1 Measurement Scales of Variable

The variables that we will generally encounter fall into one of the following measurement scales or levels of measurement.

Figure 3.2: Measurement scales of variable



Source: Adopted from Weiner (2007)

(a) nominal scale; (b) ordinal scale; (c) interval scale; and (d) ratio scale.

3.1.1 Nominal scale

Nominal scale is simply a system of assigning number symbols to events in order to label them. The usual example of this is the assignment of registration numbers to students or numbers of basketball players in order to identify them. Such numbers cannot be considered to be associated with an ordered scale for their order is of no consequence; the numbers are just convenient labels for the particular class of events and as such have no quantitative value.

Nominal scales provide convenient ways of keeping track of people, objects and events. One cannot do much with the numbers involved. For example, one cannot usefully average these nominal numbers and come up with a meaningful value. Neither can one usefully compare the numbers assigned to one group with the numbers assigned to another. The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed. There is no generally used measure of central tendency or of dispersion for nominal scales. Chi-square test is the most common test of statistical significance that can be utilized, and for the measures of correlation

Nominal scale is the least powerful level of measurement. It indicates no order or distance relationship and has no arithmetic origin. A nominal scale simply describes differences between things by assigning them to categories. Nominal data are, thus, counted data.

3.1.2 Ordinal scale

The lowest level of the ordered scale that is commonly used is the ordinal scale. The ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Rank orders represent ordinal scales and are frequently used in research relating to qualitative phenomena. A student's rank in his graduation class involves the use of an ordinal scale.

One has to be very careful in making statement about scores based on ordinal scales. For instance, if Student A's position in his class is 10 and Student B's position is 40, it cannot be said that A's position is four times as good as that of B. The statement would make no sense at all. Ordinal scales only permit the ranking of items from highest to lowest. Ordinal measures have no absolute values, and the real differences between adjacent ranks may not be equal. All that can be said is that one person is higher or lower on the scale than another, but more precise comparisons cannot be made.

Thus, the use of an ordinal scale implies a statement of 'greater than' or 'less than' (an equality statement is also acceptable) without our being able to state how much greater or less. The real difference between ranks 1 and 2 may be more or less than the difference between ranks 5 and 6. Since the numbers of this scale have only a rank meaning, the appropriate measure of central tendency is the median. A percentile or quartile measure is used for measuring dispersion. Correlations are restricted to various rank order methods. Measures of statistical significance are restricted to the non-parametric methods.

3.1.3 Interval scale

In the case of interval scale, the intervals are adjusted in terms of some rule that has been established as a basis for making the units equal. The units are equal only in so far as one accepts the assumptions on which the rule is based. Interval scales can have an arbitrary zero, but it is not possible to determine for them what may be called an absolute zero or the unique origin.

Interval scales provide more powerful measurement than ordinal scales for interval scale also incorporates the concept of equality of interval. As such more powerful statistical measures can be used with interval scales. Mean is the appropriate measure of central tendency, while standard deviation is the most widely used measure of dispersion. Product moment correlation techniques are appropriate and the generally used tests for statistical significance are the ‘t’ test and ‘F’ test.

3.1.4 Ratio scale

Ratio scale represents the actual amounts of variables. Measures of physical dimensions such as weight, height, distance, etc. are examples. They have an absolute or true zero of measurement. The term ‘absolute zero’ is not as precise as it was once believed to be. We can conceive of an absolute zero of length and similarly we can conceive of an absolute zero of time. For example, the zero point on a centimetre scale indicates the complete absence of length or height. But an absolute zero of temperature is theoretically unobtainable and it remains a concept existing only in the scientist’s mind.

A characteristic difference between the ratio scale and all other scales is that the ratio scale can express values in terms of multiples of fractional parts, and the ratios are true ratios. For example, a metre is a multiple (by 100) of a centimetre distance; a millimetre is a tenth (a fractional part) of a centimetre. The ratios are 1:100 and 1:10. There is no ambiguity in the statements ‘twice as far’, ‘twice as fast’ and ‘twice as heavy’. Of all levels of measurement, the ratio scale is amenable to the greatest range of statistical tests.

Generally, all statistical techniques are usable with ratio scales and all manipulations that one can carry out with real numbers can also be carried out with ratio scale values. Multiplication and division can be used with this scale but not with other scales mentioned above. Geometric and harmonic means can be used as measures of central tendency and coefficients of variation may also be calculated.

Thus, proceeding from the nominal scale (the least precise type of scale) to ratio scale (the most precise), relevant information is obtained increasingly. If the nature of the variables permits, the researcher should use the scale that provides the most precise description.

3.2 Errors in measurement

The measurement error is the difference between the true value and the measured value (Mohajan, 2017). These errors may be positive or negative.

Usually there are three measurement errors in research (Malhotra, 2004, Bajpai & Bajpai 2014):

i. gross errors,

ii. systematic error, that affects the observed score in the same way on every measurement, and

iii. random error; that varies with every measurement. In research a true score theory is represented according to **Bajpai and Bajpai 2014** Mathematically as

$$\boldsymbol{X_O} = \boldsymbol{X_T} + \boldsymbol{X_S} + \boldsymbol{X_R}$$

Where, X_O = the observed score or measurement; X_T = the true score of the characteristic; X_S = systematic error; X_R = random error

If the random error in the above equation is zero, then instrument is termed as reliable and if both systematic error as well as random error are zero then instrument considered as valid. (This leads us to the discussion on the reliability and validity of the measurement instrument used in research). The total measurement error is the sum of the systematic error, which affects the model in a constant fashion, and the random error, which affects the model randomly. Systematic errors occur due to stable factors which influence the observed score in the same way on every occasion that a measurement is made. However, random error occurs due to transient factors which influence the observed score differently each time

Measurement should be precise and unambiguous in an ideal research study. This objective, however, is often not met with in entirety. As such the researcher must be aware about the sources of error in measurement. The following are the possible sources of error in measurement.

a. **Respondent:** At times the respondent may be reluctant to express strong negative feelings or it is just possible that he may have very little knowledge but may not admit his ignorance. Transient factors like fatigue, boredom, anxiety, etc. may limit the ability of the respondent to respond accurately and fully.

b. **Situation:** Situational factors may also come in the way of correct measurement. Any condition which places a strain on interview can have serious effects on the interviewer-respondent rapport. For instance, if someone else is present, he can distort responses by joining in or merely by being present. If the respondent feels that anonymity is not assured, he may be reluctant to express certain feelings.

c. **Measurer:** The interviewer can distort responses by rewording or reordering questions. His behaviour, style and looks may encourage or discourage certain replies from respondents. Careless mechanical processing may distort the findings. Errors may also creep in because of incorrect coding, faulty tabulation and/or statistical calculations, particularly in the data-analysis stage.

d. **Instrument:** Error may arise because of the defective measuring instrument. The use of complex words, beyond the comprehension of the respondent, ambiguous meanings, poor printing, inadequate space for replies, response choice omissions, etc. are a few things that make the measuring instrument defective and may result in measurement errors. Another type of instrument deficiency is the poor sampling of the universe of items of concern.

Reliability and Validity of Measurement instruments used in research

Reliability and validity are the two most important and fundamental features in the evaluation of any measurement instrument or tool for a good research.

3.3 Reliability

Simply put, a reliable measuring instrument is one which gives you the same measurements when you repeatedly measure the same unchanged objects or events. The coefficient of reliability falls between 0 and 1, with perfect reliability equalling 1, and no reliability equalling 0. If a measuring instrument were perfectly reliable, then it would have a perfect positive (r = +1) correlation with the true scores. If you measured an object or event twice, and the true scores did not change, then you would get the same measurement both times. In statistical terms, reliability is analogous to variance (low reliability = high variance). Errors of measurement that affect reliability are random errors.

Reliability is mainly divided into two types as: i) Stability, and ii) Internal consistency reliability.

3.3.1. Stability: It is defined as the ability of a measure to remain the same over time despite uncontrolled testing conditions or respondent themselves. A perfectly stable measure will produce exactly the same scores time after time. Two methods to test stability are: i) test-retest reliability, and ii) parallel-form reliability.

i. **Test – retest reliability:** The reliability coefficient obtained by repetition of the same measure on a second time is called the test-retest reliability. For example, employees of a company may be asked to complete the same questionnaire about employee job satisfaction two times with an interval and correlation coefficient between two set of data will be calculated. If the correlation coefficient is found to be high, better the test-retest reliability. If the coefficients yield above 0.7, are considered acceptable, and coefficients yield above 0.8, are considered very good (Sim & Wright, 2005; Madan & Kensinger, 2017). The interval of the two tests should not be very long as this may affect the reliability of research.

ii. **Parallel-forms reliability:** It is a measure of reliability obtained by administering different versions of an assessment tool to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions. If they are highly correlated, then they are known as parallel-form reliability (DeVellis, 2006). For example, the levels of employee satisfaction of a Company may be assessed with questionnaires, in-depth interviews and focus groups, and the results are highly correlated. Then we may be sure of the measures that they are reasonably reliable (Yarnold, 2014).

3.3.2. Internal Consistency Reliability: It is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results. It examines whether or not the items within a scale or measure are homogeneous [DeVellis, 2006]. The items should "hang together as a set", and be capable of independently measuring the same concept. Consistency can be examined through: i) The inter-item consistency, and ii) Split-half reliability.

i. Inter-item / Inter-rater reliability: It establishes the equivalence of ratings obtained with an instrument when used by different observers. Reliability is determined by the correlation of the scores from two or more independent raters, or the coefficient of agreement of the judgments of the raters. For example, levels of employee motivation of a Company can be assessed using observation method by two different assessors, and inter-rater reliability relates to the extent of difference between the two assessments. The most common internal consistency measure is Cronbach's alpha (α), which is usually interpreted as the mean of all possible split half coefficients. It is a function of the average inter-correlations of items, and the number of items in the scale. In the social sciences, acceptable range of alpha value estimates from 0.7 to 0.8 (Nunnally & Bernstein, 1994).

ii. **Split-half reliability:** It measures the degree of internal consistency by comparing the results of one half of a test with the results from the other half. Half of the items are combined to form one new measure and the other half is combined to form the second new measure. The result is two tests and two new measures testing the same behaviour. It requires only one administration, especially appropriate when the test is very long. If the two halves of the test provide similar results this would suggest that the test has internal reliability. It is a quick and easy way to establish reliability. It can only be effective with large questionnaires in which all questions measure the same construct, but it would not be appropriate for tests which measure different constructs (Chakrabartty, 2013).

3.4 Validity

Validity of a research instrument assesses the extent to which the instrument measures what it is designed to measure and whether the research findings are not spurious. Oliver 2010 sees validity of research as an extent at which requirements of scientific research method have been followed during the process of generating research findings. It is a compulsory requirement for all types of studies. This consist of the way the groups were selected; data were recorded or analyses were performed. It refers to whether a study can be replicated.

Drost (2011) outlined four types of validity that researchers should consider: statistical conclusion validity, internal validity, construct validity, and external validity. Each type answers an important question and is discussed next.

3.4 1. Statistical conclusion validity

Does a relationship exist between the two variables? Statistical conclusion validity pertains to the relationship being tested. Statistical conclusion validity refers to inferences about whether it is reasonable to presume covariation given a specified alpha level and the obtained variances (Cook & Campbell, 1979). There are some major threats to statistical conclusion validity such as low statistical power, violation of assumptions, reliability of measures, reliability of treatment, random irrelevancies in the experimental setting, and random heterogeneity of respondents.

3.4.2 Internal validity

Given that there is a relationship, is the relationship a causal one? Are there no confounding factors in my study? Internal validity speaks to the validity of the research itself. For example, a manager of a company collects an employee job satisfaction survey before Christmas just after everybody received a nice bonus. The results showed that all employees were happy. Again, do the results really indicate job satisfaction in the company or do the results show a bias? There are many threats to internal validity of a research design. Some of these threats are: history, maturation, testing, instrumentation, selection, mortality, diffusion of treatment and compensatory equalisation, rivalry and demoralisation. A discussion of each threat is beyond the scope of this course.

3.4.3 Construct validity

If a relationship is causal, what are the particular cause and effect behaviours or constructs involved in the relationship? Construct validity refers to how well you translated or transformed a concept, idea, or behaviour – that is a construct – into a functioning and operating reality, the operationalisation (Trochim, 2006). To substantiate construct validity involves accumulating evidence in six validity types: face validity, content validity, criterion related validity, concurrent and predictive validity, convergent validity and discriminant validity.

i. *Face Validity*. Face validity is a subjective judgment on the operationalisation of a construct. It is usually used to describe the appearance of validity without empirical testing. So, it is normally considered to be the weakest form of validity. For example, estimating the speed of a car based on its outward appearance (guesswork) is face validity. This is not a very scientific type of validity and it does not depend on established theories for support.

ii. *Content validity*. Content validity is a qualitative means of ensuring that indicators tap the meaning of a concept as defined by the researcher. It ensures that the questionnaire includes adequate set of items that tap the concept. The more the scale items represent the domain of the concept being measured, the greater the content validity. There is no statistical test to determine whether a measure adequately covers a content area, content validity usually depends on the judgment of experts in the field. The unclear and obscure questions can be amended, and the ineffective and nonfunctioning questions can be discarded by the advice of the reviewers. For example, in arithmetic operations, the test problem will be content valid if the researcher focuses on addition, subtraction, multiplication and division, but will be content invalid if the researcher focuses on one aspect of arithmetic alone, addition (say) [Thatcher, 2010].

iii. **Criterion-related validity.** Criterion-related validity is the degree of correspondence between a test measure and one or more external referents (criteria), usually measured by their correlation. For example, suppose we survey employees in a company and ask them to report their salaries. If we had access to their actual salary records, we could assess the validity of the survey (salaries reported by the employees) by correlating the two measures. In this case, the employee records represent an (almost) ideal standard for comparison.

iv. **Concurrent and predictive validity:** When the criterion exists at the same time as the measure, we talk about concurrent validity. Concurrent ability refers to the ability of a test to predict an event in the present. The previous example of employees' salary is an example of concurrent validity. When the criterion occurs in the future, we talk about predictive validity. For example, predictive validity refers to the ability of a test to measure some event or outcome in the future. A good example of predictive validity is the use of students' GMAT scores to predict their successful completion of an MBA program. Another example is to use students' GMAT scores to predict their GPA in a graduate program. We would use correlations to assess the strength of the association between the GMAT score with the criterion (i.e., GPA).

v. ***Convergent Validity:*** Convergent Validity is established when the scores obtained with two different instruments measuring the same concept are highly correlated.

vi. ***Discriminant Validity*** is established when, based on theory, two variables are predicted to be uncorrelated, and the scores obtained by measuring them are indeed empirically found to be so.

3.4.4 External validity

If there is a causal relationship from construct X to construct Y, how generalisable is this relationship across persons, settings, and times? External validity of a study or relationship implies generalising to other persons, settings, and times. Generalising to well-explained target populations should be clearly differentiated from generalising across populations. Each is truly relevant to external validity: the former is critical in determining whether any research objectives which specified populations have been met, and the latter is crucial in determining which different populations have been affected by a treatment to assess how far one can generalise. For example, can the causal relationship observed in a manufacturing plant be replicated in a public institution, in a bureaucracy, or on a military base? This question could be addressed by varying settings and then analysing for a causal relationship within each setting.

SELF ASSESSMENT EXCERCISE

1. Identify the different scales which the dependent and independent variables can be measured.
2. What are the likely sources of error in the measurement of economic variables?

4.0 CONCLUSION

Variables can take different forms depending on their scale of measurement such as nominal, ordinal, interval and ratio scales. The three measurement errors in research includes; gross errors, systematic error and random error. Possible sources of error in measurement includes; **respondent, situation, measurer and instrument.** Reliability and validity are the two most important and fundamental features in the evaluation of any measurement instrument or tool for a good research.

5.0 SUMMARY

Measurement scales of variable are nominal scale, ordinal scale, interval scale and ratio scale. The three measurement errors in research includes; gross errors, systematic error and random error. Possible sources of error in measurement includes; **respondent, situation, measurer and instrument.** Reliability and validity are the two most important and fundamental features in the evaluation of any measurement instrument or tool for a good research.

6.0 TUTOR MARKED ASSIGNMENTS

1. Explain the three measurement errors encountered in economic research.
2. What are the possible sources of measurement error?
3. Differentiate between Reliability and validity.

7.0 REFERENCES / FURTHER READINGS

Allchin, D. (2001). Error types. *Perspectives on Science, 9*(1), 38-58.

Bajpai, S. R., & Bajpai, R. C. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health, 3*(2), 112-115.

Chakrabartty, S. N. (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education, 3*(1), 1-8.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Muffin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Devillis, R. E. (2006). Scale development: Theory and application. *Applied Social Science Research Method Series, Vol. 26,* Newbury Park: SAGE.

Drost, E. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1).

Madan, C. R., & Kensinger, E. A. (2017). Test–retest reliability of brain morphology estimates. *Brain Informatics, 4,* 107–121.

Malhotra, N. K. (2004). *Marketing Research: An Applied Orientation* (4th ed.). New Jersey: Pearson Education, Inc.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University, 17*(3): 58-82.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). NY: Mcgraw-Hill.

Trochim, W. M. K. (2006). Introduction to validity. *Social Research Methods*, retrieved from www.socialresearchmethods.net/kb/introval.php, September 9, 2010.

Weiner, J. (2007). *Measurement: Reliability and Validity Measures*. Bloomberg School of Public Health: Johns Hopkins University

Yarnold, P. R. (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Emergency severity index (version 3) and Canadian triage acuity scale. *Optimal Data Analysis, 3*(4), 50-54.

UNIT THREE: SAMPLING

CONTENT

1.0 INTRODUCTION

In research it is largely difficult to have complete enumeration of all items in the 'population' This necessitates the selection of only a few items or respondents. The selected respondents constitute what is technically called a 'sample' and the selection process is called 'sampling technique.' The survey so conducted is known as 'sample survey'. Algebraically, let the population size be N and if a part of size n (which is $< N$) of this population is selected according to some rule for studying some characteristic of the population, the group consisting of these n units is known as 'sample'. Researcher must prepare a sample design for his study i.e., he must plan how a sample should be selected and of what size such a sample would be. A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of items to be included in the sample i.e., the size of the sample. Sample design is determined before data are collected.

2.0 OBJECTIVES

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

- Identify the different steps in sampling design

- Distinguish between probability and non-probability sampling

3.0 MAIN CONTENT

In research it is largely difficult to have complete enumeration of all items in the ‘population’ This necessitates the selection of only a few items or respondents. The selected respondents constitute what is technically called a ‘sample’ and the selection process is called ‘sampling technique.’ The survey so conducted is known as ‘sample survey’. Algebraically, let the population size be *N* and if a part of size *n* (which is < *N*) of this population is selected according to some rule for studying some characteristic of the population, the group consisting of these *n* units is known as ‘sample’. Researcher must prepare a sample design for his study i.e., he must plan how a sample should be selected and of what size such a sample would be. A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample
design may as well lay down the number of items to be included in the sample i.e., the size of the sample. Sample design is determined before data are collected.

3.1 Steps in Sample Design

While developing a sampling design, the researcher must pay attention to the following points:

(i) **Type of universe:** This is to clearly define the set of objects, technically called the Universe, to be studied. The universe can be finite or infinite.

(ii) **Sampling unit:** Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of
such units that he has to select for his study.

(iii) **Source list:** It is also known as ‘sampling frame’ from which sample is to be drawn. It contains the names of all items of a universe (in case of finite universe only).

(iv) **Size of sample:** This refers to the number of items to be selected from the universe to constitute a sample. This is a major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate. The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed. The size of population must be kept in view for this also limits the sample size. The parameters of interest in a research study must be kept in view, while deciding the size of the sample. Costs too dictate the size of sample that we can draw. As such, budgetary constraint must invariably be taken into consideration when we decide the sample size.

(v) **Parameters of interest:** In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population.

(vi) **Budgetary constraint:** Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.

(vii) **Sampling procedure:** Finally, the researcher must decide the type of sample he will use i.e., he must decide about the technique to be used in selecting the items for the sample. In fact, this technique or procedure stands for the sample design itself. There are several sample designs (explained in the pages that follow) out of which the researcher must choose one for his study. Obviously, he must select that design which, for a given sample size and for a given cost, has a smaller sampling error.

3.2 Types of Sample Design

Sample designs are basically of two types viz., non-probability sampling and probability sampling.

3.2.1 Non-probability sampling

In this type of sampling, items for the sample are selected deliberately by the researcher. Thus, there is always the danger of bias entering into this type of sampling technique. But if the investigators are impartial, work without bias and have the necessary experience so as to take sound judgement, the results obtained from an analysis of deliberately selected sample may be tolerably reliable. Sampling error in this type of sampling cannot be estimated and the element of bias, great or small, is always there. There are two main types of nonprobability sampling design: convenience sampling and purposive sampling.

i. **Convenience sampling** is the least reliable of all sampling designs in terms of generalizability, but sometimes it may be the only viable alternative when quick and timely information is needed, or for exploratory research purposes.

ii. **Purposive sampling** plans fall into two categories: *judgment and quota sampling designs.* Judgment sampling, though restricted in generalizability, may sometimes be the best sampling design choice, especially when there is a limited population that can supply the information needed. Quota sampling is often used on considerations of cost and time and the need to adequately represent minority elements in the population. Although the generalizability of all nonprobability sampling designs is very restricted, they have certain advantages and are sometimes the only viable alternative for the researcher.

3.2.2 Probability sampling

Probability sampling is also known as ‘random sampling’ or ‘chance sampling’. Under this sampling design, every item of the universe has an equal chance of inclusion in the sample. We can measure the errors of estimation or the significance of results obtained from a random sample, and this fact brings out the superiority of random sampling design over the deliberate sampling design. Random sampling ensures the law of Statistical Regularity which states that if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the universe. This is the reason why random sampling is considered as the best technique of selecting a representative sample.

Assuming a certain finite population consisting of six elements (say *a*, *b*, *c*, *d*, *e*, *f*) i.e., $N = 6$. Suppose that we want to take a sample of size $n = 3$ from it. Then there are 20 possible distinct samples of the required size, and they consist of the elements *abc*, *abd*, *abe*, *abf*, *acd*, *ace*, *acf*, *ade*, *adf*, *aef*, *bcd*, *bce*, *bcf*, *bde*, *bdf*, *bef*, *cde*, *cdf*, *cef*, and *def*. If we choose one of these samples in such a way that each has the probability 1/20 of being chosen, we will then call this a random sample.

We can verify the above example by calculating the probabilities. Since we have a finite population of 6 elements and we want to select a sample of size 3, the probability of drawing any one element for our sample in the first draw is 3/6, the probability of drawing one more element in the second draw is 2/5, (the first element drawn is not replaced) and similarly the probability of drawing one more element in the third draw is 1/4. Since these draws are independent, the joint probability of the three elements which constitute our sample is the product of their individual probabilities and this works out to $3/6 \times 2/5 \times 1/4 = 1/20$.

Even this relatively easy method of obtaining a random sample can be simplified in actual practice by the use of random number tables. Various statisticians like Tippett, Yates, Fisher have prepared tables of random numbers which can be used for selecting a random sample.

Complex random sampling designs

Probability sampling under restricted sampling techniques, as stated above, may result in complex random sampling designs. Such designs may as well be called ‘mixed sampling designs’ for many of such designs may represent a combination of probability and non-probability sampling procedures in selecting a sample. Some of the popular complex random sampling designs are as follows:

(i) Systematic sampling: In some instances, the most practical way of sampling is to select every *i*th item on a list. Sampling of this type is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random numbers to pick up the unit with which to start. For instance, if a 4 per cent sample is desired, the first item would be selected randomly from the first twenty-five and thereafter every 25th item would automatically be included in the sample. Thus, in systematic sampling only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals.

(ii) Stratified sampling: If a population from which a sample is to be drawn does not constitute a homogeneous group, stratified sampling technique is generally applied in order to obtain a representative sample. Under stratified sampling the population is divided into several sub-populations that are individually more homogeneous than the total population (the different sub-populations are called 'strata') and then we select items from each stratum to constitute a sample. Since each stratum is more homogeneous than the total population, we are able to get more precise estimates for each stratum and by estimating more accurately each of the component parts, we get a better estimate of the whole. If Pi represents the proportion of population included in stratum i, and n represents the total sample size, the number of elements selected from stratum i is $n \times Pi$. To illustrate it, let us suppose that we want a sample of size $n = 30$ to be drawn from a population of size $N = 8000$ which is divided into three strata of size $N1 = 4000$, $N2 = 2400$ and $N3 = 1600$. Adopting proportional allocation, we shall get the sample sizes as under for the different strata:

For strata with $N_1 = 4000$, we have $P_1 = 4000/8000$

and hence $n1 = n \times P1 = 30\ (4000/8000) = 15$

Similarly, for strata with $N2 = 2400$, we have

$N_2 = n \times P2 = 30\ (2400/8000) = 9$, and

for strata with $N3 = 1600$, we have

$n3 = n \times P3 = 30\ (1600/8000) = 6$.

Thus, using proportional allocation, the sample sizes for different strata are 15, 9 and 6 respectively which is in proportion to the sizes of the strata viz., 4000 : 2400 : 1600. Proportional allocation is considered most efficient and an optimal design when the cost of selecting an item is equal for each stratum, there is no difference in within-stratum variances, and the purpose of sampling happens to be to estimate the population value of some characteristic.

In cases where strata differ not only in size but also in variability and it is considered reasonable to take larger samples from the more variable strata and smaller samples from the less variable strata, we can then account for both (differences in stratum size and differences in stratum variability) by using disproportionate sampling design by requiring:

$n_1/N_1\sigma_1 = n_2 /N_2\sigma_2 = ... = n_k /N_k\sigma_k$

where σ_1, σ_2 , ... and σ_k denote the standard deviations of the k strata, $N1$, $N2$,…, Nk denote the

sizes of the k strata and $n1$, $n2$,…, nk denote the sample sizes of k strata. This is called '*optimum allocation*' in the context of disproportionate sampling. The allocation in such a situation results in the following formula for determining the sample sizes different strata:

$$n_i = \frac{n\ .\ N_i\sigma_i}{N_1\sigma_1+N_2\sigma_2+\cdots+N_k\sigma_k} \quad for\ i = 1, 2 \dots \dots . . and\ k$$

We may illustrate the use of this by an example.

Illustration

A population is divided into three strata so that $N1$ = 5000, $N2$ = 2000 and $N3$ = 3000. Respective standard deviations are:

$\sigma1 = 15$, $\sigma2 = 18$ and $\sigma3 = 5$.

How should a sample of size n = 84 be allocated to the three strata, if we want optimum allocation using disproportionate sampling design?

Solution: Using the disproportionate sampling design for optimum allocation, the sample sizes for different strata will be determined as under:

$$n_i = \frac{n\ .\ N_i\sigma_i}{N_1\sigma_1 + N_2\sigma_2 + \cdots + N_k\sigma_k}$$

For sample size of strata with $N1$ = 5000

$$n_1 = \frac{84(5000)15}{5000(15) + 2000(18) + 3000(5)}$$

$$= \frac{6300000}{126000} = 50$$

For sample size of strata with *N*2 = 2000

$$n_2 = \frac{84(2000)18}{5000(15) + 2000(18) + 3000(5)}$$

$$= \frac{3024000}{126000} = 24$$

For sample size for strata with *N*3 = 3000

$$n_3 = \frac{84(3000)5}{5000(15) + 2000(18) + 3000(5)}$$

$$= \frac{6300000}{126000} = 10$$

It is not necessary that stratification be done keeping in view a single characteristic. Populations are often stratified according to several characteristics known as *cross-stratification*. As an example of mixed sampling, the procedure wherein we first have stratification and then simple random sampling is known as stratified random sampling.

(iii) Cluster sampling: If the total area of interest happens to be a big one, a convenient way in which a sample can be taken is to divide the area into a number of smaller non-overlapping areas and then to randomly select a number of these smaller areas (usually called clusters), with the ultimate sample consisting of all (or samples of) units in these small areas or clusters.

(iv) Area sampling: If clusters happen to be some geographic subdivisions, in that case cluster sampling is better known as area sampling. In other words, cluster designs, where the primary sampling unit represents a cluster of units based on geographic area, are distinguished as area sampling.

(v) Multi-stage sampling: Multi-stage sampling is a further development of the principle of cluster sampling. Suppose we want to investigate the working efficiency of commercial banks in a country and we want to take a sample of few banks for this purpose. The first

stage is to select large primary sampling unit such as states in a country. Then we may select certain districts and interview all banks in the chosen districts. This would represent a two-stage sampling design with the ultimate sampling units being clusters of districts. If instead of taking a census of all banks within the selected districts, we select certain towns and interview all banks in the chosen towns. This would represent a three-stage sampling design. If instead of taking a census of all banks within the selected towns, we randomly sample banks from each selected town, then it is a case of using a four-stage sampling plan. If we select randomly at all
stages, we will have what is known as ‘multi-stage random sampling design’.

Ordinarily multi-stage sampling is applied in big inquires extending to a considerable large geographical area, say, the entire country.

(vii) Sequential sampling: This sampling design is somewhat complex sample design. The ultimate size of the sample under this technique is not fixed in advance, but is determined according to mathematical decision rules on the basis of information yielded as survey progresses. This is usually adopted in case of acceptance sampling plan in context of statistical quality control.

SELF ASSESSMENT EXERCISE

1. Briefly describe the steps in sampling design
2. What are the advantages of probability sampling over the non-probability?

4.0. CONLUSION

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sampling involves certain steps for its successful determination. When selecting a sample, it must be either from the probability or non-probability family of sampling design.

5.0 SUMMARY

While developing a sampling design, the researcher must pay attention to type of universe, the sampling unit, source list, sample size, parameters of interest, the sampling procedure and above all his budget. Sample designs are basically of two types viz., non-probability sampling and probability sampling. In the former, every item of the universe has an equal chance of inclusion in the sample as it will be selected randomly while in the latter case, items for the sample are selected deliberately by the researcher.

6.0. TUTOR-MARKED ASSIGNMENT

1. Briefly describe the steps in sampling design
2. What are the advantages of probability sampling over the non-probability?

7.0 REFERENCES/FURTHER READINGS

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business: A Skill Building Approach* (5th ed.). West Sussex, UK: John Wiley and Sons.

UNIT FOUR: MODEL BUILDING IN ECONOMICS

CONTENTS

1.0 INTRODUCTION

A model is a formal framework for representing the basic features of a complex system by a few central relationships (Samuelson & Nordhaus, 1998). Models take the form of graphs, mathematical equation, and computer programs, they make a series of description from which it assumes how individuals or economic variables behaves. Modeling is a deliberate simplification of reality. Economists use a model to analyze and visualize the economic problem, econometrics is a branch of economics that uses the methods of statistics to measure and estimate quantitative economic relationships. Samuelson and Nordhaus (1998) defines an econometric model as a set of equations representing the behavior of the economy which has been estimated using historical data.

2.0 OBJECTIVES

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

- Know the types of models in economics,
- Reasons for economic models
- Elements of model building in economics.

3.0 MAIN CONTENT

3.1 Types of Models in Economics

Forms of economic models includes the followings:

i. Flow Charts,

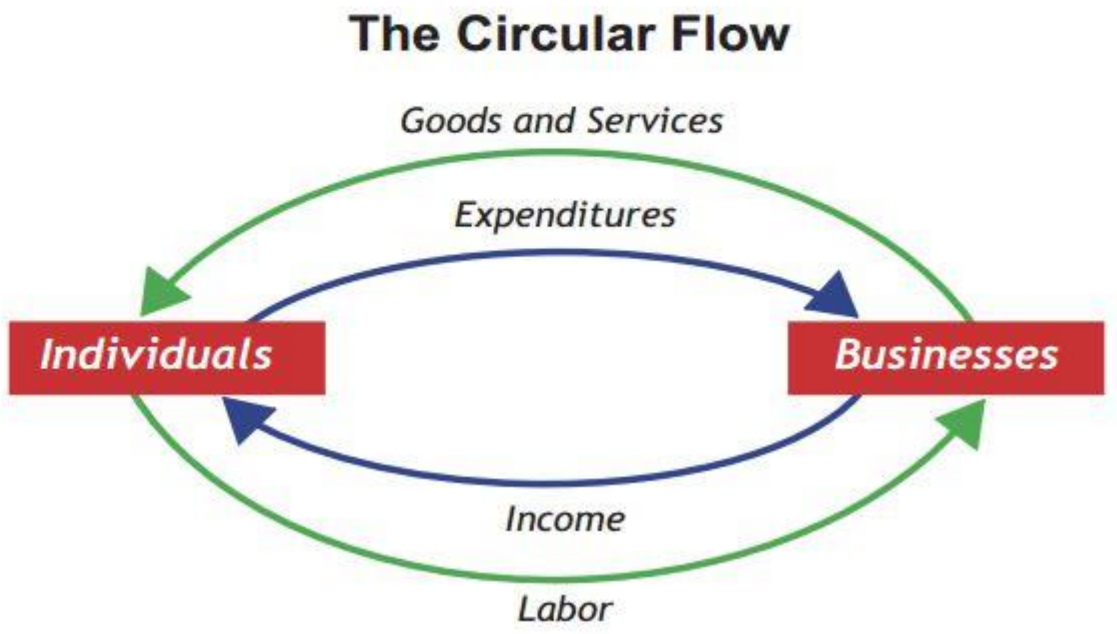
ii. Graphs,

iii. Mathematical models.

Economists use these models in different purposes which depends on many factors such as the type of data involved, how the data can be presented, and what is needed from the model used. This unit explains the main role of these different models.

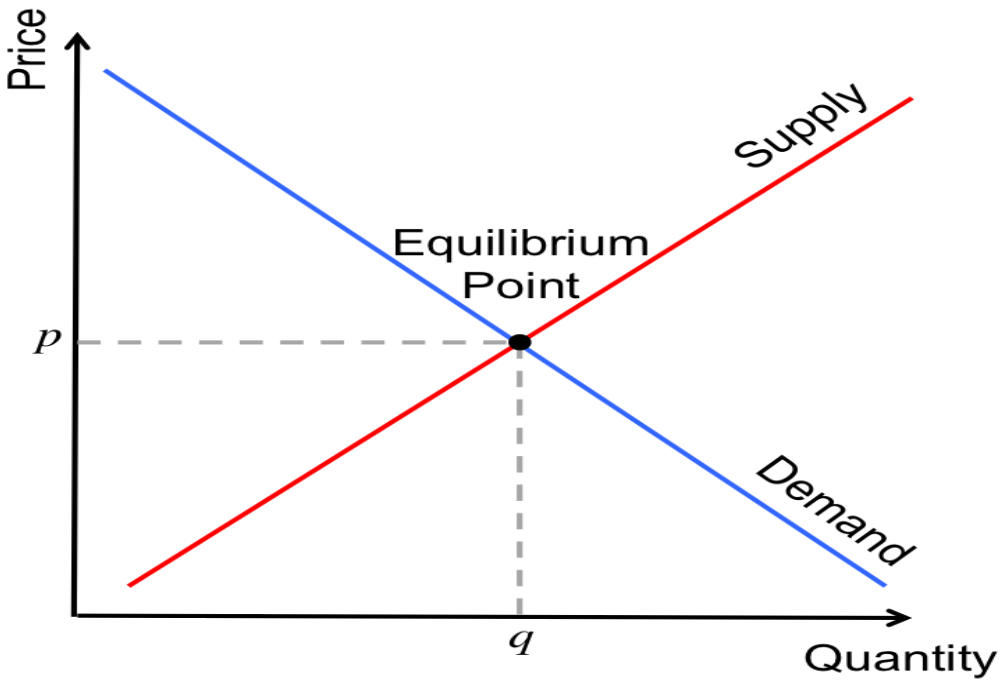
i. Flow Chart

Chart is a drawing which is used to show the association that exist in different stages of a process or parts of a system. In economics, flow chart is a basic model used to show how an economy function, or to explain how an economy is organized, or to illuminate how participants in the economy interact with one another. Above all, it looks at the way money, goods, and services move throughout the economy. Figure 4.1 shows a circular flow of income between individuals and businesses, whereby,the individuals provide labour to the businesses, and the bussinesses pay wages and salaries (income) for the labour, and the individuals use this income to buy goods and services (expenditure) made by the bussinesses.



ii. Graph

Graph is an illustration representing a system of connections or interrelations among two or more things by a number of distinctive dots, lines, bars, etc. In economics graph comprises of a planned drawing, consisting of a line or lines, showing how two or more sets of numbers or variables are related to each other. Numerous kinds of graphs are used in economics, some of them show how variables change over time; others show the relationship between different variables. For instance, figure 4.2 shows the relationship between demand and supply of a normal good. It acertain the law of demand and supply which states that "the lower the price, the higher the quantity demanded and the lower the quantity supplied", it also shows the point at which quantity supplied equals to quantity demanded which is known as the equilibrium point.



iv. Mathematical Model

A mathematical model can be broadly defined as a formulation or equation that expresses the essential features of a physical system or process in mathematical terms. In a very general sense, it can be represented as a functional relationship of the form:

Dependent variable = f(independent variables, parameters, forcing functions). Economists use a mathematical model as a 'theory' to determine a particular problem. Many proven mathematical models are used in economics as the formulas to help economists calculate and analyze the numerical issues easier, they use the scientific way of thinking to develop a new econometric model or a theory to explain the economic system.

When a model is built up, it is a safe way to writing down all the variables that are related to the studied economic issue, otherwise we might forget some influenced variables.

The relationship between expenditure on education and economic growth is expressed below using an econometric model:

$GDP = f(\beta_1 cee, , \beta_2 REE, \beta_3 GFCF, \beta_4 TALF, \mathcal{E}_1)$……..(1)

$GDP = \Phi_o + \beta_1 cee + \beta_2 REE + \beta_3 GFCF + \beta_4 TALF + \mathcal{E}_1$ ………(2)

Where by:

GDP = economic growth
CEE = capital expenditure on education
REE = recurrent expenditure on education
GFCF = gross fixed capital formation
TALF = total active labour force
$\Phi_{o\,=}$ constant
β_1 , $\beta_{2,}\beta_3$ and $\beta_{4\,.}$ = parameters to be estimated
$\mathcal{E}_1$ = forcing function

3.2 Reasons for Economic Model

Economists uses scientific approach to deal with economic problems which begins from scientific observation of a phenomenon, after which hypothesis is built, then the usage of scientific experiment or existing theories to prove the hypothesis and reach a conclusion that can be true or false depending on the results obtained from the experiment. Economists use this method when they think about a particular issue in the economy; they try to simplify an economic issue in the way that everyone will understand and be able to follow their thought; they try to find a formula that can help them calculate the numerical issue; and try to develop a new economic theory to explain the economic behavior in the real world. In order to fulfill these approaches, a tool is needed to assist which is called a 'model'. The ways and manner that economists use models can be classified into three purposes:

i. Explaining an economic process.
ii. Examining an economic issue.
iii. Developing a new economic theory.

i. Explaining an economic process

As the saying "A picture is worth a thousand words", it is true that a picture can express idea better than words or equations. Graphics help economists in many specific purposes: some shows the relationship between observed variables; some shows how the economic process runs; some shows variables trend. Graphs and charts play the main role in this purpose. For example, demand and supply curve shows the relationship between quantity of goods or services demanded and supplied while the production possibility curve shows the relationship between factors of production and technology. These types of models help economists to see the picture of the process, know how data and variables relate to each other and also clarify the idea about the problem. They are used in combination with mathematical expression of the model.

ii. Examining an economic issue

Economists not only want to see how the data or variables looks like or their relationship in the graphic way, but also want to see the trend or changes from observed data or variables. In order to achieve this, a wide range of mathematical models are used to examine economic issues; in some cases, formulas are used to calculate data or analyze it, in other cases, mathematical models are used in the problem-solving process in the forms of equations in order to estimate and forecast a given set of economic data or variables. It should be noted that not only mathematical models are used in the problem-solving process, but also graphics, statistics, etc. are used as the problem-solving tools.

iii. Developing a new economic theory

Economic theories help economists measure changes, understand trend, and predict future result in an economy. How do economists develop a new economic theory? In order to do so, they combine scientific approach, mathematical knowledge, and historical economic data together. Then use the problem-solving process to find the suitable model for the particular problem, after which the model is tested and if confirms to be true, it is used as a new theory. Most economic theories are developed based on or related to the existing theories (or models).

3.3 Elements of Model Building

The use of models is widespread in both the physical and social sciences, economists developed their models as aids to understanding economic issues. These describe for example, the way economic growth changes as a result to shift in public expenditure.

The ultimate goal of models is to learn something about the real world. Economic models incorporate three common basic elements:

i. Ceteris paribus (all things being equal) assumption.
ii. Optimization assumption.
iii. Positive and normative distinction.

i. The Ceteris Paribus Assumption

Economic models attempt to describe relationships between variables, for example, economic growth model, seeks to explain economic growth in relation to the basic determinants of growth such as physical capital, labour and technology. This parsimony in model specification allows the study of economic growth in a simplified setting in which it is possible to understand how the specific forces operate. Although many external variables such as foreign direct invest, income and inflation, affect economic growth, these external variables are held constant in the construction of the model (note that it is not assumed that other factors do not affect economic growth, but rather, such other variables are assumed to be unchanged during the period of study). In this way the effect of only a few forces can be studied in a simplified setting. Such ceteris paribus (other things being equal) assumptions are used in all economic modeling.

i. Optimization Assumptions

Many economic models start from the assumption that the economic agents are rational and aimed at optimizing satisfaction. This assumption is considered as a good starting point for developing economic models, this is because the optimization assumption is useful for generating precise, solvable models which can draw a range of mathematical techniques and are empirically valid, as such are good at explaining reality.

ii. Positive-Normative Distinction

This is the attempt to differentiate carefully between “positive” and “normative” questions. Positive economics describes and explains various economic phenomena or the "what is" scenario, and it is based on fact. Normative economics on the other hand focuses on the value of economic fairness, or what the economy "should be" or "ought to be." For example, an economist engaged in positive analysis might investigate what quantities of capital, labor, and land a consumable goods industry uses in producing a certain amount of consumable goods. The economist might also choose to measure the costs and benefits of dedicating even more resources to consumable goods. But when economists advocate that more resources should be allocated to consumable goods industries, they have implicitly moved into normative analysis.

3.1. SELF ASSESSMENT EXERCISE

Explain how a model is built in Economics.

4.0. CONLUSION

The unit concludes that econometric model is a set of equations representing the behavior of the economy which has been estimated using historical data; model can take the form of graphs, mathematical equation, and computer programs. Model in economics is used in order to analyze and visualize the economic problem and relationships between variables.

5.0 SUMMARY

In this unit, a model is defined as a formal framework for representing the basic features of a complex system by a few central relationships; and that an econometric model as a set of equations representing the behavior of the economy which has been estimated using historical data. Forms of economic models includes graphs and Mathematical expressions. Models in economics are used for explaining an economic process, examining an economic issue and developing a new economic theory. Economic models incorporate three common basic elements; the ceteris paribus (all things being equal) assumption, the assumption that economic decision-makers seek to optimize something and the distinction between “positive” and “normative” questions.

6.0. TUTOR-MARKED ASSIGNMENT

i. Explain the forms of models in economics.

ii. Give the reasons why models are used in economic process.

iii. Discuss the basic elements of model building in economics.

7.0 REFERENCES/FURTHER READINGS

Begg, D., Fischer, S., & Dornbusch, R. (2000). *Economics* (6th ed). London: McGraw-Hill.

Nicholson, W., & Snyder, C. (2012). *Microeconomic Theory Basic Principles and Extensions* (10th ed.). OH, USA: South-Western CENGAGE Learning.

Mankiw, N. G. (2001). *Principles of Economics* (2nd ed.). Orlando: Harcourt College.

Samuelson, P. A., & Nordhaus, W. D. (1998). *Economics* (16th ed.). Irwin: McGraw-Hill.

MODULE FOUR: QUANTITATIVE DATA ANALYSIS

UNIT ONE: BIVARIATE ANALYSIS

CONTENTS

1.0 INTRODUCTION

The variables used in economic research are many and varied. We distinguish between them and establish their relation to each other. In analysing the relations between variables we encounter one or more of the following. A univariate, bivariate and multivariate analyses. A univariate analysis can be established using a **frequency distribution table**; **measures of central tendency** such as the arithmetic mean, the median, and the mode from which normal distribution is when these measures are located at the same value; **measures of dispersion** such as the range, variance and the standard deviation. In the subsequent units we shall treat the bivariate and multivariate analyses in details.

2.0 OBJECTIVES

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

- Know the basic methods of determining the degree of association among variables,
- Understand the rudiments of simple regression
- Determine a cause and effect relationship among two economic variables

3.0 MAIN CONTENT

3.1. Degree of association

This measures and assesses the direction and degree of association in this case between two variables.

3.1.1 Correlation

This is statistically termed correlation coefficients. The commonly used coefficients assume that there is a linear relationship between two variables, either positive or negative. A positive relationship is one where more of one variable is related to more of another, or less of one is related to less of another. For example, higher educational attainment relates to more income. A negative relationship is one where more of one variable is related to less of another or the other way round. For example, more income relates to less anxiety, less income relates to more anxiety. Note that detecting a relationship does not mean you have found an influence or a cause, although it may be that this is the case. Again, graphical displays help to show the analysis.

3.1.2 Scatter graphs

These are a useful type of diagram that graphically shows the relationship between two variables by plotting variable data from cases on a two-dimensional matrix. If the resulted plotted points appear in a scattered and random arrangement, then no association is indicated. If, however they fall into a linear arrangement, a relationship can be assumed, either positive or negative. The closer the points are to a perfect line, the stronger the **association**. A line that is drawn to trace this notional line is called the **line of best fit** or **regression line**. This line can be used to predict one variable value on the basis of the other.

3.1.3 Cross tabulation (contingency tables) is a simple way to display the relationship between variables that have only a few categories. They show the relationships between each of the categories of the variables in both number of responses and percentages.

3.1.4 Nonparametric tests: These are available to assess the relationship between variables measured on a nominal or an ordinal scale. These tests are used when normality of distributions cannot be assumed as in nominal or ordinal data. Spearman's rank correlation and Kendall's rank correlation are used to examine relationships between two ordinal variables. A correlation matrix is used to examine relationships between interval and/or ratio variables. The chi-square *(X^2)* test measures the degree of association or linkage between two variables by comparing the differences between the observed values and expected values to indicate whether or not the observed pattern in the variables is due to chance.

3.2 Difference in Means

When you wish to examine the relationship between a nominal (or ordinal) variable with two categories that is an independent variable, and a dependent variable that is measured at the interval/ratio level then an appropriate then an appropriate procedure and test is to examine the difference in means, and calculate a t-test. To see the direction of the difference in means just examine the respective means for the two groups. For the t-test there is an associated significance level. If the significance level is 0.05 for instance, then the difference in means is statistically significant.

Another common requirement is to look for differences between values obtained under two or more different conditions, e.g. a group before and after a training course, or three groups after different training courses. There are a range of tests that can be applied to discern the **variance** depending on the number of groups. For a single group, say the performance of students on a particular course compared with the mean results of all the other courses in the university you can use the chi-square or the one group *t*-test.

For two groups, e.g. comparing the results from the same course at two different universities, you can use the two group *t*-tests, which compares the means of two groups. There are two types of test, one for paired scores, i.e. where the same persons provided scores under each condition, or for unpaired scores where this is not the case.

3.3 Simple Regression analysis

Regression analysis is the process of developing a mathematical model that can be used to predict one variable by using another variable or variables. Regression involving two variables (bivariate) is known as simple regression. This describes the proportion of variation in one dependent accounted for by one other independent variable. That is to say, it measures the proportion of change in dependent variable occasioned by the explanatory variable. Regression can also be used to predict one variable value on the basis of the other We can explain this using regression line, coefficient of determination.

3.3.1 Regression line: A simple linear regression equation represents a straight line. Indeed, to summarize the relationship between the dependent and independent variables drawn as a straight line through the data points. A linear regression model is based on the linear function such as $Y = b_0 + b_1X$

The parameter b_0 is called the *intercept* and the parameter b_1 is called the *slope* of the regression line. The value of the intercept determines the point where the regression line meets the Y axis of the graph. The value of the slope represents the amount of change in Y when X increases by one unit. Using a hypothetical data on the volume of sales against advertising cost we can show the resulting straight line fitted on it.

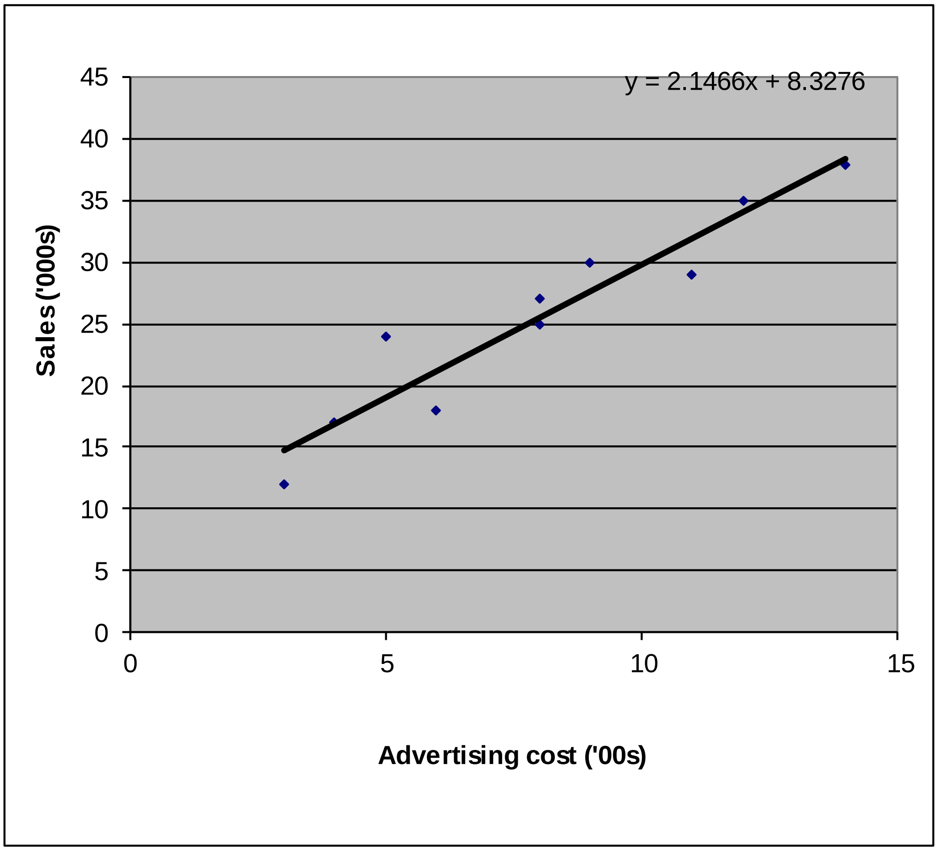


Figure 3.1 Regression line for advertising cost example

As you can see from the above diagram, regression has fitted a straight line on the data. Regression therefore aims to fit a line through the data in order to describe the relationship between two variables. If the relationship between the two variables is linear (like the one in this example), then a straight line is fitted through the data and the data points will lie very close to that line. According to the least squares regression method, a regression line is fitted through the data in such a way so that the sum of the squares of the distances between the data points and the line is minimised. The resulting regression line could be straight or curved depending on the type of the relationship between the two variables.

Once a regression model has been developed it can then be used to make predictions. In this case we can predict the volume of sales for a company based on its advertising cost. Suppose that we want to predict the volume of sales for a company which has spent £1000 on advertising. All we need to do is take this to be the value of X in the regression model and then calculate the corresponding value of Y. In general, it is too risky to attempt to predict a value of Y using an X value which is outside the range of X values of the data collected. That is because the linear relationship that exists between the two variables only covers the existing data and this could change if another range of values was considered. Also note that the prediction should be based on the regression model, which is itself based on the data.

3.2.2 Regression coefficient or coefficient of determination

The coefficient of determination, r^2, provides information about the goodness of fit of the regression model: it is a statistical measure of how well the regression line approximates the real data points. r^2 is the percentage of variance in the dependent variable that is explained by the variation in the independent variable. If r^2 is close to 1, most of the variation in the dependent variable can been explained by the regression model. In other words, the regression model fits the data well. On the other hand, if r^2 is close to 0, most of the data variation cannot be explained by the regression model. In this case, the regression model fits the data poorly.

3.4 Statistical Significance

In calculating statistical significance, a researcher set up null hypothesis that the two variables in the sample are not related. Statistical power, or just power, is the probability of correctly rejecting the null hypothesis based on the level of significance. Statistical power depends on:

i. Alpha (α): the statistical significance criterion used in the test. If alpha moves closer to zero (for instance, if alpha moves from 5% to 1%), then the probability of finding an effect of one variable on the other decreases. This implies that the lower the α (that is, the closer α moves to zero) the lower the power; the higher the alpha, the higher the power.

ii. Effect size: the effect size is the size of a difference or the strength of a relationship *in the population*: a large difference (or a strong relationship) in the population is more likely to be found than a small difference (similarity, relationship).

iii. The size of the sample: at a given level of alpha, increased sample sizes produce more power, because increased sample sizes lead to more accurate parameter estimates. Thus, increased sample sizes lead to a higher probability of finding what we were looking for. However, increasing the sample size can also lead to too much power, because even very small effects will be found to be statistically significant.

3.4 Causality

A causal statement, to no one's surprise, has two components: a cause and an effect. Three commonly accepted conditions must hold for a scientist to claim that *X* causes *Y*:

a. time precedence
b. relationship
c. no spuriousness

For *X* to cause *Y*, *X* must precede *Y* in time. Such time precedence means a causal relationship is fundamentally asymmetric while many statistical measures of relationship are symmetric. Implicit in a causal vocabulary is an active, dynamic process that inherently must take place over time. The second condition for causation is the presence of a functional relationship between cause and effect. The third and final condition for a causal relationship is nonspuriousness. For a relationship between *X* and *Y* to be nonspurious, there must not be a *Z* that causes both *X* and *Y* such that the relationship between *X* and *Y* vanishes once *Z* is controlled.

Causal relationship between two variables x and y can be unidirectional or bidirectional. Unidirectional causality exists when one of the variable causes the other. Bidirectional on the other hand is a causality both ways. That is x causes y and y causes x at the same time. Assuming *x* and *y* to be money supply and GDP, we can say that there is a unidirectional causality running from money supply to GDP when we found that the former statistically causes the latter. However, if we found that both x and y causes each other (GDP stimulates money supply and vice versa) then there is a bidirectional causality.

Even though most economic relationships are causal in nature, regression analysis cannot prove such causality. We don't actually test for theoretical causality; instead, we test for Granger causality. **Granger causality,** or precedence, is a circumstance in which one time series variable consistently and predictably changes before another variable does. If one variable precedes ("Granger causes") another, we still can't be sure that the first variable "causes" the other to change, but we can be fairly sure that the opposite is not the case.

3.4.1 Distinction between correlation and causation

There is an old saying that "correlation does not imply causation,". The presence of correlation does not, in itself, prove causation. To prove causation, three things are needed:

i. Statistically significant correlation must be present between the alleged cause and the effect.

ii. The alleged cause must be present before or at the same time as the effect.

iii. There must be an explanation as to how the alleged cause produces the effect.

3.5. Mathematicatical Analysis of Bivariate relationship

3.5.1 Simple Linear Regression

As doscussed in Section 3.3 above, the major aim of the regression is to find the equation of the least squares regression line of y on x. The **vertical distance** each point is above or below the regression line on Figure 3.1 has been added to the diagram. These distances are called *deviations* or *errors* – they are symbolised as $d_1, d_2, ..., d_n$.

When drawing in a regression line, the aim is to make the line fit the points as closely as possible. We do this by making the **total of the squares of the deviations as small as possible**, i.e. we minimise $\sum d_i^2$. If a line of best fit is found using this principle, it is called the **least-squares regression line**.

Example 1:

The table below shows the cost of advertising *x* (*in 000s*) Naira and the volume of sales in tonnes. The management believe that a linear relationship will exist between the variables.

Cost of Advert, *x* (₦000)	0	1	2	3	4	5	6
Sales, *y (tonnes)*	2.4	4.3	5.0	6.9	9.1	11.4	13.5

The managers may wish to estimate the relationship between advertising and the volume of sales. they could do this by finding the equation of the line of best fit. There is a formula which gives the equation of the line of best fit.

The statistical equation of the simple linear regression line, when only the response variable Y is random, is: $Y = \beta_0 + \beta_1 x + \varepsilon$ (or in terms of each point: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$)
Here β_0 is called the intercept, β_1 the regression slope, ε is the random error with mean 0, x is the regressor (independent variable), and Y the response variable (dependent variable).

The least squares regression line is obtained by finding the values of β_0 and β_1 values (denoted in the solutions as $\hat{\beta}_0$ & $\hat{\beta}_1$) that will minimize the sum of the squared vertical distances from all points to the line: $\Delta = \sum d_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2$
The solutions are found by solving the equations: $\frac{\partial \Delta}{\partial \beta_0} = 0$ and $\frac{\partial \Delta}{\partial \beta_1} = 0$

The equation of the fitted least squares regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (or in terms of each point: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) ----- For simplicity of notations, many books denote the fitted regression equation as: $\hat{Y} = b_0 + b_1 x$
where $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.

Notations: $S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = \sum (x_i - \bar{x})(y_i - \bar{y})$; $S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = \sum (x_i - \bar{x})^2$;
$\overline{x}$ and $\overline{y}$ are the mean values of x and y respectively.

Note 1: Please notice that **in finding the least squares regression line, we do not need to assume any distribution for the random errors ε_i. However, for statistical inference on the model parameters (**β_0 and β_1), it is assumed in our class that the errors have the following three properties:

- Normally distributed errors
- Homoscedasticity (constant error variance $\text{var}(\varepsilon_i) = \sigma^2$ for Y at all levels of X)
- Independent errors (usually checked when data collected over time or space)

The above three properties can be summarized as: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2),\ i = 1, \cdots, n$

Note 2: Please notice that the least squares regression is only suitable when the random errors exist in the dependent variable Y only. If the regression X is also random – it is then referred to as the **Errors in Variable (EIV) regression**.

We can work out the equation for our example as follows:

$$\sum x = 0 + 1 + ... + 6 = 21 \text{ so } \bar{x} = \frac{21}{7} = 3$$

$$\sum y = 2.4 + 4.3 + ... + 13.5 = 52.6 \text{ so } \bar{y} = \frac{52.6}{7} = 7.514...$$

$$\sum xy = (0 \times 2.4) + (1 \times 4.3) + ... + (6 \times 13.5) = 209.4$$

$$\sum x^2 = 0^2 + 1^2 + ... + 6^2 = 91 \text{ so } \bar{x} = \frac{21}{7} = 3$$

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 209.4 - \frac{21 \times 52.6}{7} = 51.6$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 91 - \frac{(21)^2}{7} = 28$$

So, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{51.6}{28} = 1.843$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.514 - 1.843 \times 3 = 1.985$.

So the equation of the regression line is $\hat{y}$ = 1.985 + 1.843*x*.

Assuming you are asked to find out the volume of sales after expending ₦3,500 on adverts:

$\hat{y}$ = 1.985 + 1.843 × 3.5 = 8.44 (3sf)

If you want to forecast the sales after spending ₦8,000 on adverts, we substitute $\hat{y}$ = 8 into the regression equation:

8 = 1.985 + 1.843*x*

Solving this we get: *x* = 3.26 tonnes

Example 2:

The heights and weights of a sample of 11 students are:

Height (m) h	1.36	1.47	1.54	1.56	1.59	1.63	1.66	1.67	1.69	1.74	1.81
Weight (kg) w	52	50	67	62	69	74	59	87	77	73	67

$[n = 11 \quad \sum h = 17.72 \quad \sum h^2 = 28.705 \quad \sum w = 737 \quad \sum w^2 = 50571 \quad \sum hw = 1196.1]$

a) Calculate the regression line of *w* on *h*.

b) Use the regression line to estimate the weight of someone whose height is 1.6m.

Note: Both height and weight are referred to as **random** variables – their values could not have been predicted before the data were collected. If the sampling were repeated again, different values would be obtained for the heights and weights.

Solution:

a) We begin by finding the mean of each variable:

$$\bar{h} = \frac{\sum h}{n} = \frac{17.72}{11} = 1.6109...$$

$$\bar{w} = \frac{\sum w}{n} = \frac{737}{11} = 67$$

Next we find the sums of squares:

$$S_{hh} = \sum h^2 - \frac{\left(\sum h\right)^2}{n} = 28.705 - \frac{17.72^2}{11} = 0.1597$$

$$S_{ww} = \sum w^2 - \frac{\left(\sum w\right)^2}{n} = 50571 - \frac{737^2}{11} = 1192$$

$$S_{hw} = \sum hw - \frac{\sum h \sum w}{n} = 1196.1 - \frac{17.72 \times 737}{11} = 8.86$$

The equation of the least squares regression line is:

$$\hat{w} = \hat{\beta}_0 + \hat{\beta}_1 h$$

where

$$\hat{\beta}_1 = \frac{S_{hw}}{S_{hh}} = \frac{8.86}{0.1597} = 55.5$$

and

$$\hat{\beta}_0 = \bar{w} - \hat{\beta}_1 \bar{h} = 67 - 55.5 \times 1.6109 = -22.4$$

So the equation of the regression line of *w* on *h* is:

$$\hat{w} = -22.4 + 55.5h$$

b) To find the weight for someone that is 1.6m high:

$\hat{w}$ = -22.4 + 55.5×1.6 = 66.4 kg

3.5.2 Simple linear regression and Measures of variation

The objective here is to determine the measures of variation, the goodness-of-fit measure, and the correlation coefficient

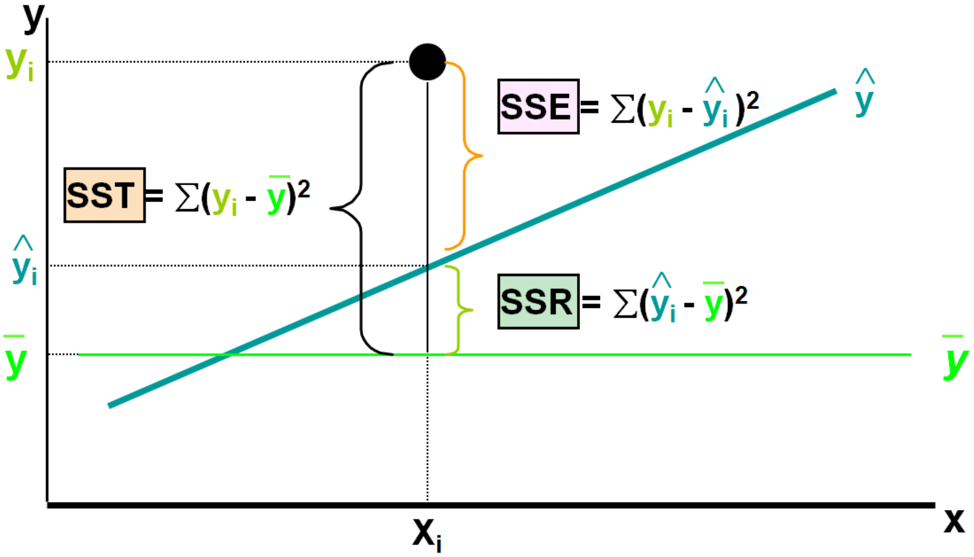
Sums of Squares

- Total sum of squares = Regression sum of squares + Error sum of squares

$SST = SSR + SSE$

(Total variation = Explained variation + Unexplained variation)

- Total sum of squares (Total Variation): $SST = \sum\left(Y_i - \bar{Y}\right)^2$
- Regression sum of squares (Explained Variation by the Regression): $SSR = \sum\left(\hat{Y}_i - \bar{Y}\right)^2$
- Error sum of squares (Unexplained Variation): $SSE = \sum\left(Y_i - \hat{Y}_i\right)^2$



Coefficients of Determination and Correlation

Coefficient of Determination – it is a measure of the regression goodness-of-fit

It also represents the proportion of variation in Y "explained" by the regression on X

$$R^2 = \frac{SSR}{SST};\ 0 \le R^2 \le 1$$

Pearson (Product-Moment) Correlation Coefficient -- measure of the direction and strength of the linear association between Y and X

- The sample correlation is denoted by ***r*** and is closely related to the coefficient of determination as follows:

$$r = sign\left(\hat{\beta}_1\right)\sqrt{R^2}\ ;\ -1 \le r \le 1$$

The sample correlation is indeed defined by the following formula:

$$\boxed{r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2][\sum(y-\bar{y})^2]}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}}}$$

- The corresponding population correlation between Y and X is denoted by $\boldsymbol{\rho}$ and defined by:

$$\rho = \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E\left[(X-\bar{X})(Y-\bar{Y})\right]}{\sqrt{Var(X)Var(Y)}}$$

- Therefore, one can see that in the population correlation definition, both X and Y are assumed to be random. When the joint distribution of X and Y is bivariate normal, one can perform the following t-test to test whether the population correlation is zero:

  - Hypotheses
    H_0: ρ = 0 (no correlation)
    H_A: ρ ≠ 0 (correlation exists)

  - Test statistic

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \overset{H_0}{\sim} t_{n-2}$$

Note: One can show that this t-test is indeed the same t-test in testing the regression slope $\beta_1 = 0$ shown in the following section.

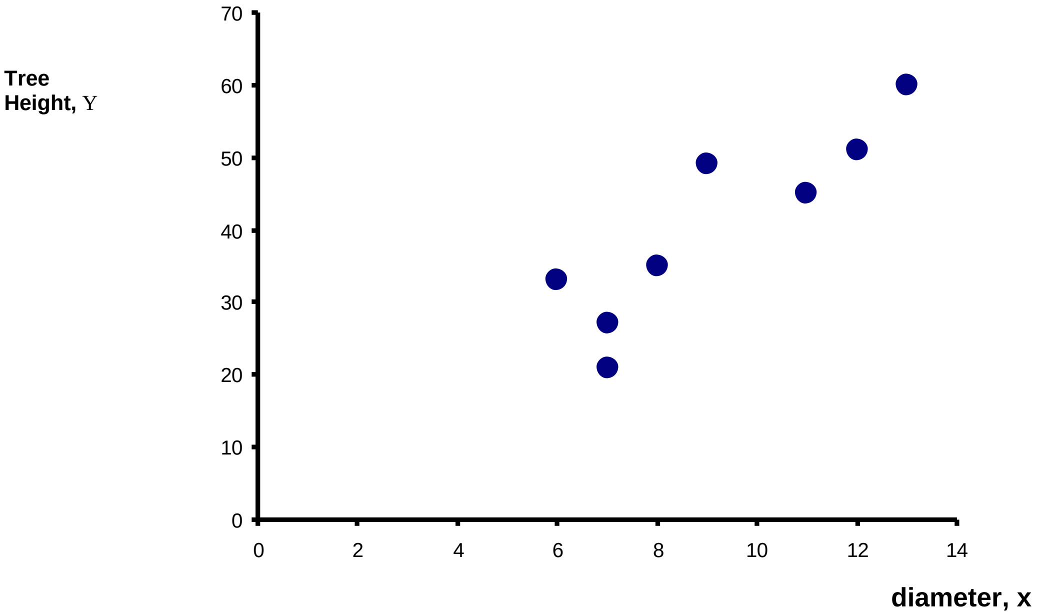
Note: The sample correlation is not an unbiased estimator of the population correlation. You can study this and other properties from the wiki site:
http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Example 3: The following example tabulates the relations between trunk diameter and tree height.

Tree Height	**Trunk Diameter**			
y	**x**	**xy**	**y^2**	**x^2**
35	**8**	**280**	**1225**	**64**
49	**9**	**441**	**2401**	**81**
27	**7**	**189**	**729**	**49**
33	**6**	**198**	**1089**	**36**
60	**13**	**780**	**3600**	**169**

21	**7**	**147**	**441**	**49**
45	**11**	**495**	**2025**	**121**
51	**12**	**612**	**2601**	**144**
Σ=321	**Σ=73**	**Σ=3142**	**Σ=14111**	**Σ=713**

Scatter plot:



$$\begin{aligned} r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \\ &= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}} \\ &= 0.886 \end{aligned}$$

r = 0.886 → relatively strong positive linear association between x and y

Significance Test for Correlation

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

H_0: ρ = 0   (No correlation)
H_1: ρ ≠ 0   (correlation exists)

$$\boxed{t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68}$$

At the significance level α = 0.05, we reject the null hypothesis because $|t_0| = 4.68 \geq t_{6,0.025} = 2.447$ and conclude that there is a linear relationship between tree height and trunk diameter.

3.5.3 Standard Error of the Estimate (Residual Standard Deviation)

- The mean of the random error ε is equal to zero.
- An estimator of the standard deviation of the error ε is given by

$$\hat{\sigma} = s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

3.5.4 Inferences Concerning the Slope

t-test

Test used to determine whether the population based slope parameter (β_1) is equal to a pre-determined value (often, but not necessarily 0). Tests can be one-sided (pre-determined direction) or two-sided (either direction).

2-sided t-test:

- H_0: $\beta_1 = 0$ (no linear relationship)
- H_1: $\beta_1 \neq 0$ (linear relationship does exist)

- **Test statistic:** $t_0 = \frac{b_1 - 0}{s_{b_1}}$

Where $s_{b_1} = \frac{s_\varepsilon}{\sqrt{S_{XX}}} = \frac{s_\varepsilon}{\sqrt{\sum(x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$

At the significance level α, we reject the null hypothesis if $|t_0| \geq t_{n-2,\alpha/2}$

(Note: one can also conduct the one-sided tests if necessary.)

F-test (based on *k* independent variables)

A test based directly on sum of squares that tests the specific hypotheses of whether the slope parameter is 0 (2-sided). The book describes the general case of *k* predictor variables, **for simple linear regression, *k* = 1**.

$$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$$

$$TS: F_{obs} = \frac{MSR}{MSE} = \frac{SSR / k}{SSE / (n - k - 1)}$$

$$RR: F_{obs} \geq F_{\alpha, k, n-k-1}$$

3.5.5 Analysis of Variance (based on *k* Predictor Variables – for simple linear regression, *k* = 1)

Source	**df**	**Sum of Squares**	**Mean Square**	**F**
Regression	k	SSR	$MSR=SSR/k$	$F_{obs}=MSR/MSE$
Error	n-k-1	SSE	$MSE=SSE/(n$-k-$1)$	---
Total	n-1	SST	---	---

100(1-α)% Confidence Interval for the slope parameter, β_1: $\boxed{b_1 \pm t_{n-2,\alpha/2} s_{b_1}}$

- If entire interval is positive, conclude $\beta_1>0$ (Positive association)
- If interval contains 0, conclude (do not reject) $\beta_1=0$ (No association)
- If entire interval is negative, conclude $\beta_1<0$ (Negative association)

Example 4: A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet). A random sample of 10 houses is selected

- Dependent variable (y) = house price in $1000s
- Independent variable (x) = square feet
-

House Price in $1000s (y)	**Square Feet (x)**
245	**1400**
312	**1600**
279	**1700**

308	**1875**
199	**1100**
219	**1550**
405	**2350**
324	**2450**
319	**1425**
255	**1700**

Solution: Regression analysis output:

Regression Statistics	
Multiple R	**0.76211**
R Square	**0.58082**
Adjusted R Square	**0.52842**
Standard Error	**41.33032**
Observations	**10**

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**18934.9348**	**18934.9348**	**11.0848**	**0.01039**
Residual	**8**	**13665.5652**	**1708.1957**		
Total	**9**	**32600.5000**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Estimated house price = 98.24833+ 0.10977(square feet)

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, b_1 = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

$$\boxed{R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082}$$

This means that 58.08% of the variation in house prices is explained by variation in square feet.

$$\boxed{s_\varepsilon = 41.33032}$$

The standard error (estimated standard deviation of the random error) is also given in the output (above).

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic:
  - $$\boxed{t_0 = \frac{b_1 - 0}{s_{b_1}} \approx 3.329}$$

At the significance level α = 0.05, we reject the null hypothesis because $|t_0| = 3.329 \geq t_{8,0.025} = 2.306$ and conclude that there is sufficient evidence that square footage affects house price.

Confidence Interval Estimate of the Slope: $\boxed{b_1 \pm t_{n-2,\alpha/2} s_{b_1}}$

The 95% confidence interval for the slope is (0.0337, 0.1858)

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Predict the price for a house with 2000 square feet:

$$\begin{aligned}\text{house price} &= 98.25 + 0.1098(\text{sq.ft.})\\ &= 98.25 + 0.1098(2000)\\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,85

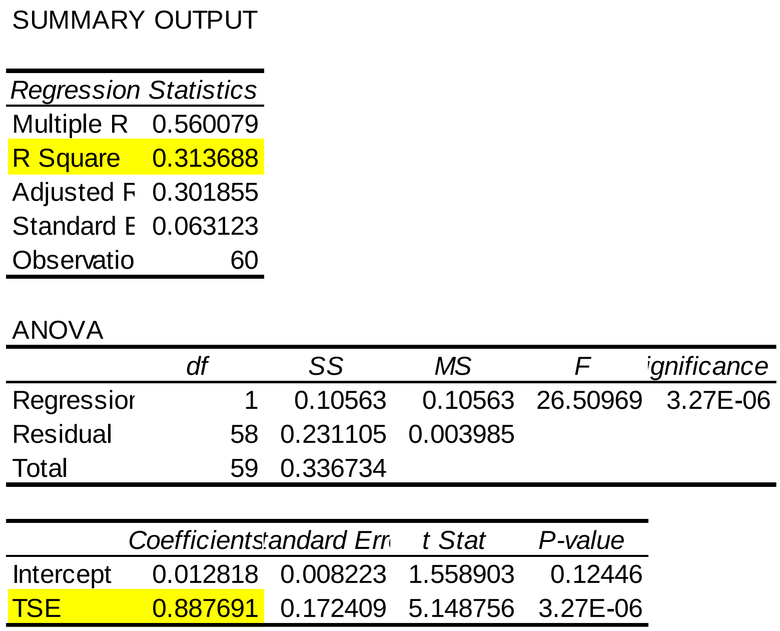
3.5.6 Finance Application: Market Model

One of the most important applications of linear regression is the *market model.* It is assumed that rate of return on a stock (R) is linearly related to the rate of return on the overall market.

$$\mathbf{R = \beta_0 + \beta_1 R_m + \varepsilon}$$

R: Rate of return on a particular stock
R_m: Rate of return on some major stock index
β_1: The beta coefficient measures how sensitive the stock's rate of return is to changes in the level of the overall market.

Example 5: An estimated the market model for Nortel, a stock traded in the Toronto Stock Exchange. Data consisted of monthly percentage return for Nortel and monthly percentage return for all the stocks.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.560079
R Square	0.313688
Adjusted F	0.301855
Standard E	0.063123
Observatio	60

ANOVA

	df	*SS*	*MS*	*F*	*ignificance*
Regressior	1	0.10563	0.10563	26.50969	3.27E-06
Residual	58	0.231105	0.003985		
Total	59	0.336734			

	Coefficients	*tandard Err*	*t Stat*	*P-value*
Intercept	0.012818	0.008223	1.558903	0.12446
TSE	0.887691	0.172409	5.148756	3.27E-06

TSE (estimated regression slope): This is a measure of the stock's market related risk. In this sample, for each 1% increase in the TSE return, the average increase in Nortel's return is .8877%.

R Square (R^2) This is a measure of the total risk embedded in the Nortel stock, that is market-related.

Specifically, 31.37% of the variation in Nortel's return are explained by the variation in the TSE's returns.

SELF ASSESSMENT EXERCISE

1. What are the different forms of bivariate analysis methods in economics?

2. Please prove that:
Total sum of squares = Regression sum of squares + Error sum of squares; that is,
$SST = SSR + SSE$

4.0 CONLUSION

Bivariate analysis considers the properties of two variables in relation to each other. This explores relationships to determine whether a change in one variable corresponds with variation in another variable. A researcher would often like to know how one variable is related to another so as to determine the nature, direction, and significance of the relationships of the variables used in the study A researcher can determine one or a combination of the following relationships.

5.0 SUMMARY

In this unit, a model is defined as a formal framework for representing the basic features of a complex system by a few central relationships; and that an econometric model as a set of equations representing the behavior of the economy which has been estimated using historical data. Forms of economic models includes graphs and Mathematical expressions. Models in economics are used for explaining an economic process, examining an economic issue and developing a new economic theory. Economic models incorporate three common basic elements; the ceteris paribus (all things being equal) assumption, the assumption that economic decision-makers seek to optimize something and the distinction between “positive” and “normative” questions.

6.0. TUTOR-MARKED ASSIGNMENT

1. Explain the simple regression anaysis.
2. What are the different methods of determining the degree of association between two economic variables?

7.0 REFERENCES/FURTHER READINGS

Studenmund, A. H. (2000). *Using Econometrics: A Practical Guide* (4th ed.). Addison-Wesley.

Kenny, D. A. (1979). *Correlation and Causality*. NY: John Wiley

Begg, D., Fischer, S., & Dornbusch, R. (2000). *Economics*, (6th ed.). London: McGraw-Hill.

Greener, S. (2008). *Business Research Methods*. London: Ventus.

Gujarati, D. (2004). Basic Econometrics (4th ed.). NY: McGraw-Hill.

UNIT TWO: MULTIVARIATE ANALYSIS

CONTENTS

1.0 INTRODUCTION

Multivariate analysis is a collection of statistical techniques which helps a researcher to analyse research data whereby a number of observations are available for each object and are mainly important in social science research. These techniques are largely empirical and deal with the reality; they possess the ability to analyse difficult data and also help in various types of decision-making. Multivariate statistical analysis is classified into dependent and independent analysis.

2.0 OBJECTIVES

At the end of this unit, students are expected to know what is meant by multivariate statistical analysis and its basic classifications.

3.0 MAIN CONTENT

3.1 Multivariate Analysis

Multivariate technique is a branch of statistics which aims to generate knowledge by processing quantitative and qualitative data using standard statistical techniques which simultaneously analyse more than two variables on a sample of observations. Multivariate analysis is a collection of methods for analyzing data in which a number of observations are available for each object. These techniques are largely empirical and deal with the reality; they possess the ability to analyse complex data and also help in various types of decision-making. They are particularly important in social science research because social researchers are generally unable to use randomized laboratory experiments, like those used in medicine and natural sciences. Each technique tests the theoretical models of a research question about associations against the observed data. They allow researchers to look at relationships between variables in a predominant way and to quantify the relationship between variables. They can control association between variables by using cross tabulation, partial correlation and multiple regressions, and determine the relationship between the independent and dependent variables or to specify the conditions under which the association takes place.

Multivariate techniques are difficult and consist of high level mathematics that require a statistical program to analyze the data. These statistical programs are generally costly. The results of multivariate analysis are not always easy to interpret and tend to be based on assumptions that may be difficult to assess. For multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors. Standard errors determine how confident you can be in the results, and you can be more confident in the results from a large sample than a small one. Multivariate analysis is classified into dependent and independent analysis.

3.2 Dependent Analysis

3.2.1 Multiple regression analysis. Multiple regression analysis is used to assess and analyze a number of independent variables with the dependent variable linear function of the relationship between the statistical methods. The main objective in using this technique is to predict the variability the dependent variable based on its covariance with all the independent variables. One can predict the level of the dependent phenomenon through multiple regression analysis model, given the levels of independent variables.

A dependent variable y and independent variables $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ x1, is a linear regression relationship:

y=α+$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots\ldots\ldots + \beta_n x_n + \varepsilon$ Where α,β_1, β_2, β_3 ... , β_n are parameters to be estimated, ε is a random variable error. Obtained through experimentsx_1, x_2, $x_3 \ldots\ldots\ldots$ x_n of several sets of data and the corresponding y values.

Multiple regression analysis has the advantage of a phenomenon can be described quantitatively between certain factors and a linear function. The known values of each variable into the regression equation can be obtained estimates of the dependent variable (predictor), which can effectively predict the occurrence and development of a phenomenon. It can be used for continuous variables; it also can be used for dichotomous variables.

3.2.2 Discriminant analysis

It is a technique that allows a researcher to classify individuals or objects into one of two or more mutually exclusive and exhaustive groups on the basis of a set of independent variables. It determines which weightings of quantitative variables best discriminate between two or more than two groups of cases. The analysis creates a discriminant function which is a linear combination of the weightings and scores on these variables. It requires interval independent variables and a nominal dependent variable. For example, suppose that brand preference (say brand x or y) is the dependent variable of interest and its relationship to an individual’s income, age, education, etc. is being investigated, then discriminant analysis should be used. It is considered an appropriate technique when the single dependent variable happens to be non-metric and is to be classified into two or more groups, depending upon its relationship with several independent variables which all happen to be metric. The objective in discriminant analysis happens to be to predict an object’s likelihood of belonging to a particular group based on several independent variables. In case the dependent variable is classified into more than two groups, it is called multiple discriminant analysis while in the case whereby only two groups are to be formed, it is referred to as discriminant analysis.

3.2.3 Multivariate analysis of variance

Multivariate analysis of variance is an extension of bivariate (two variables) analysis of variance in which the ratio of among-groups variance to within-groups variance is calculated on a set of variables instead of a single variable. This technique is considered appropriate when several metric dependent variables are involved in a research study along with many non-metric explanatory variables. (But if the study has only one metric dependent variable and several nonmetric explanatory variables, ANOVA technique is used). In other words, multivariate analysis of variance is specially applied whenever the researcher wants to test hypotheses concerning multivariate differences in group responses to experimental manipulations. For instance, the market researcher may be interested in using one test market and one control market to examine the effect of an advertising campaign on sales as well as awareness, knowledge and attitudes. In that case he/she should use the technique of multivariate analysis of variance in order to achieve the objective.

3.2.4 **Conjoint Analysis**

It is a statistical method for market research which is mainly concern about measuring the relative importance of certain characteristics of a product or service. The product or service is subdivided into inseparable characteristics or functions that are subsequently presented to the consumer in the form of a questionnaire or telephone conversation, for instance. The respondents are asked to select the most favourable or most desirable group of functions or features, depending on the type of Conjoint Analysis. This allows companies to better meet the actual wishes and needs of the consumers. For example, a manufacturer of electronic devices, would like to know whether his/her customers prioritize the sound and image quality over price.

Market research based on conjoint analysis is concerned with why and how people choose between products or services and their properties. Subsequently, the product development team can use the new input, allowing the company to fulfill the customer's wishes in a profitable way. Data and feedback can be obtained by conducting experiments on customers or by conducting questionnaires with the goal to model the purchase process and purchase decision.

3.2.5 **Canonical correlation analysis**

This technique studies the relationship between several predictors (independent variables) and a set of criterion (dependent) variables or between two pairs of vectors. It is used to simultaneously predict a set of dependent variables from their joint co-variance with a set of explanatory variables. Both metric and non-metric data can be used in this multivariate technique. The procedure of using this technique is to obtain a set of weights for the dependent and independent variables in such a way that the linear composite of the criterion variables has a maximum correlation with the linear combination of the explanatory variables. For example, if a researcher wants to relate grade school adjustment to health and physical maturity of the child, he/she can then use canonical correlation analysis, provided the researcher have for each child a number of adjustment scores (such as tests, teacher's ratings, parent's ratings and so on) and also have for each child a number of health and physical maturity scores (such as heart rate, height, weight, index of intensity of illness and so on). The main objective of canonical correlation analysis is to discover factors separately in the two sets of variables such that the multiple correlation between sets of factors will be the maximum possible. Mathematically, in canonical correlation analysis, the weights of the two sets is presented as:

$\beta_1, \beta_2, \beta_3 \ldots\ldots\ldots \beta_n$ and y_1 , $y_2, y_3 \cdots \cdots . y_j$ are determined that the variables *X*= $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots . . \beta_n X_n + \beta$ and *Y* = $y_1 Y_{1+} \, y_2 Y_{2+} \, y_3 Y_3 \cdots \cdots \cdots . y_j Y_j + y$ have a maximum common variance. The process of finding the weights requires factor analyses with two matrices. The resulting canonical correlation solution then gives an overall description of the presence or absence of a relationship between the two sets of variables.

3.2.6 Structural equation modeling

Unlike the other multivariate techniques discussed, structural equation modeling (SEM) examines multiple relationships between sets of variables simultaneously. This represents a family of techniques. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis. For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures. These tools are often used to evaluate many scaled attributes or build summated scales.

3.3 Independent Analysis

3.3.1 Factor analysis

Factor analysis is the analytical process of transforming statistical data (such as measurements) into linear combinations of usually independent variables. This technique is used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms.

3.3.2 Cluster Analysis

A cluster consists of variables that correlate highly with one another and have comparatively low correlations with variables in other clusters. Cluster analysis consists of methods of classifying variables into clusters. The basic objective of cluster analysis is to determine how many mutually and exhaustive groups or clusters (based on the likenesses of profiles among units) really exist in the population and then to tell the composition of such groups. Cluster analysis are judgmental and without statistical inferences. The technique is useful in context of market research studies. Through the use of this technique a researcher can make divisions of market of a product on the basis of several features of the customers such as personality, socio-economic considerations, psychological factors, purchasing habits.

3.3.3 Multidimensional Scaling

This technique allows a researcher to measure an item in more than one dimension at a time. The basic assumption is that people perceive a set of objects as being more or less similar to one another on a number of dimensions (usually uncorrelated with one another) instead of only one.

The technique is important as it allows the researcher to study "The perceptual structure of a set of stimuli and the cognitive processes underlying the development of this structure. It reveals the most noticeable attributes which happen to be the primary determinants for making a specific decision.

3.3.4 Latent Structure Analysis

A family of statistical models. It explains the correlations among observed variables by making assumptions about the hidden (latent) causes of those variables. This type of analysis is appropriate when the variables involved in a study do not possess dependency relationship and happen to be non-metric.

3.4 Mathematical Analysis of Multivariate Relationship

In multiple regression containing *k* variables that we control, or know in advance of outcome, that are used to predict *Y*, the response (dependent variable). The *k* independent variables are labeled $X_1, X_2,...,X_k$. The levels of these variables for the i^{th} case are labeled $X_{1i} ,..., X_{ki}$. Note that simple linear regression is a special case where *k*=1, thus the methods used are just basic extensions of what we have previously done.

$Y_\mathrm{i} = \beta_0 + \beta_1 X_{1\mathrm{i}} + ... + \beta_\mathrm{j} X_{\mathrm{ki}} + \varepsilon_\mathrm{i}$

where β_j is the change in mean for *Y* when variable X_k increases by 1 unit, while holding the *k*-1 remaining independent variables constant (partial regression coefficient). This is also referred to as the slope of *Y* with variable X_k holding the other predictors constant.

3.4.1 Least Squares Fitted (Prediction) Equation (obtained by minimizing *SSE*):

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}$$

3.4.2 Coefficient of Multiple Determination

Proportion of variation in *Y* "explained" by the regression on the *k* independent variables.

$$R^2 = r^2_{Y\bullet 1,\ldots,k} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted-*R*^{2}

Used to compare models with different sets of independent variables in terms of predictive capabilities. Penalizes models with unnecessary or redundant predictors.

$$Adj - R^2 = r^2_{adj} = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSE}{SST}\right) = 1 - \left[\left(1 - r^2_{Y\bullet 1,\ldots,k}\right)\left(\frac{n-1}{n-k-1}\right)\right]$$

3.4.3 F-test for the Overall Model

Used to test whether **any** of the independent variables are linearly associated with *Y*

Analysis of Variance

Total Sum of Squares and df: $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad df_T = n-1$

Regression Sum of Squares: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \qquad df_R = k$

Error Sum of Squares: $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad df_E = n-k-1$

Source	*df*	*SS*	*MS*	*F*
Regression	*k*	*SSR*	*MSR=SSR/k*	F_{obs}=*MSR/MSE*
Error	*n*-*k*-1	*SSE*	*MSE=SSE/*(*n*-*k*-1)	---
Total	*n*-1	*SST*	---	---

F-test for Overall Model

H_0: $\beta_1 = ... = \beta_k = 0$ (*Y* is not linearly associated with **any** of the independent variables)

H_A: Not all $\beta_j = 0$ (At least one of the independent variables is associated with *Y*)

TS: $F_{obs} = \dfrac{MSR}{MSE}$

RR: $F_{obs} \geq F_{\alpha,k,n-k-1}$

P-value: Area in the *F*-distribution to the right of F_{obs}

3.4.4 Inferences Concerning Individual Regression Coefficients

Used to test or estimate the slope of *Y* with respect to X_j, after controlling for all other predictor variables.

t-test for β_j

H_0: $\beta_j = \beta_j^0$ (often, but not necessarily 0)

H_A: $\beta_j \neq \beta_j^0$

TS: $t_{obs} = \dfrac{b_j - \beta_j^0}{S_{b_j}}$

RR: $|t_{obs}| \geq t_{\frac{\alpha}{2},n-k-1}$

P-value: Twice the area in the *t*-distribution to the right of $|t_{obs}|$

(1-α)100% Confidence Interval for β_j

$$b_j \pm t_{\frac{\alpha}{2},n-k-1} S_{b_j}$$

- If entire interval is positive, conclude $\beta_j>0$ (Positive association)
- If interval contains 0, conclude (do not reject) $\beta_j=0$ (No association)
- If entire interval is negative, conclude $\beta_j<0$ (Negative association)

3.4.5 Quadratic Regression Models

When the relation between *Y* and *X* is not linear, it can often be approximated by a quadratic model.

Population Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$

Fitted Model: $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2$

Testing for any association: H_0: $\beta_1=\beta_2=0$ H_A: β_1 and/or $\beta_2 \neq 0$ (F-test)

Testing for a quadratic effect: H_0: $\beta_2=0$ H_A: $\beta_2 \neq 0$ (t-test)

3.4.6 Models with Dummy Variables

Dummy variables are used to include categorical variables in the model. If the variable has *m* levels, we include *m*-1 dummy variables, The simplest case is binary variables with 2 levels, such as gender. The dummy variable takes on the value 1 if the characteristic of interest is present, 0 if it is absent.

a. Model with no interaction

Consider a model with one numeric independent variable (X_1) and one dummy variable (X_2). Then the model for the two groups (characteristic present and absent) have the following relationships with *Y*:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad E(\varepsilon_i) = 0$$

Characteristic Present ($X_{2i} = 1$): $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_{1i}$

Characteristic Absent ($X_{2i}=0$): $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2(0) = \beta_0 + \beta_1 X_{1i}$

Note that the two groups have different *Y*-intercepts, but the slopes with respect to X_1 are equal.

b. Model with interaction

This model allows the slope of Y with respect to X_1 to be different for the two groups with respect to the categorical variable:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i \qquad E(\varepsilon_i) = 0$$

$X_{2i}=1$: $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2(1) + \beta_3 X_{1i}(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i}$

$X_{2i}=0$: $E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2(0) + \beta_3 X_{1i}(0) = \beta_0 + \beta_1 X_{1i}$

Note that the two groups have different Y-intercepts and slopes. More complex models can be fit with multiple numeric predictors and dummy variables.

3.4.7 Multiple regression in Matrix Form

Given the Data:

$$(y_1, x_{11}, x_{12}, \ldots, x_{1p-1}), (y_2, x_{21}, x_{22}, \ldots, x_{2p-1}), \ldots, (y_n, x_{n1}, x_{n2}, \ldots, x_{np-1}).$$

The **multiple linear regression model** in scalar form is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2), \;\; i = 1, \ldots, n.$$

The above linear regression can also be represented in the vector/matrix form. Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

Then,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1p-1} + \varepsilon_1 \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_{p-1} x_{2p-1} + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{np-1} + \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1p-1} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_{p-1} x_{2p-1} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{np-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} =$$

$$\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Estimation:

Least square method:

The least square method is to find the estimate of $\boldsymbol{\beta}$ minimizing the sum of square of residual,

$$S(\boldsymbol{\beta}) = S(\beta_0, \beta_1, \ldots, \beta_{p-1}) = \sum_{i=1}^{n} \varepsilon_i^2 = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{\varepsilon}^{\mathbf{t}}\boldsymbol{\varepsilon}$$

$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

since $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. Expanding $S(\boldsymbol{\beta})$ yields

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}^{\mathbf{t}} - \boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$
$$= \mathbf{y}^{\mathbf{t}}\mathbf{y} - \mathbf{y}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{y} + \boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta}$$
$$= \mathbf{y}^{\mathbf{t}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{y} + \boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta}$$

Note:

For two matrices A and B, $(AB)^t = B^t A^t$ and $\left(A^{-1}\right)^t = \left(A^t\right)^{-1}$

Similar to the procedure in finding the minimum of a function in calculus, the least square estimate ***b*** can be found by solving the equation based on the first derivative of $S(\boldsymbol{\beta})$,

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \partial S(\boldsymbol{\beta}) \big/ \partial \beta_0 \\ \partial S(\boldsymbol{\beta}) \big/ \partial \beta_1 \\ \vdots \\ \partial S(\boldsymbol{\beta}) \big/ \partial \beta_{p-1} \end{bmatrix} = \frac{\partial\left(\mathbf{y}^{\mathbf{t}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{y} + \boldsymbol{\beta}^{\mathbf{t}}\mathbf{X}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^{\mathbf{t}}\mathbf{y} + 2\mathbf{X}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\Leftrightarrow \qquad \mathbf{X}^{\mathbf{t}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathbf{t}}\mathbf{y}$$

$$\Leftrightarrow \mathbf{b} = (\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{y} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

The fitted regression equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}.$$

The fitted values (in vector): $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$

The residuals (in vector): $\mathbf{y}-\hat{\mathbf{y}}=\mathbf{y}-\mathbf{X}\mathbf{b}=\mathbf{e}=\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

Note: (i) $\dfrac{\partial(\boldsymbol{\beta}^t\mathbf{a})}{\partial\boldsymbol{\beta}}=\dfrac{\partial(\sum_{i=1}^{p}\beta_{i-1}a_i)}{\partial\boldsymbol{\beta}}=\mathbf{a},$ where $\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ and $\mathbf{a}=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$.

(ii) $\dfrac{\partial(\boldsymbol{\beta}^t\mathbf{A}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\dfrac{\partial(\sum_{i=1}^{p}\sum_{j=1}^{p}\beta_{i-1}\beta_{j-1}a_{ij})}{\partial\boldsymbol{\beta}}=2\mathbf{A}\boldsymbol{\beta},$ where A is any symmetric $p\times p$ matrix.

Note: Since

$$\left(X^tX\right)^t=X^t\left(X^t\right)^t=X^tX\ ,$$

X^tX is a symmetric matrix.
Also,

$$\left(\left(X^tX\right)^{-1}\right)^t=\left(\left(X^tX\right)^t\right)^{-1}=\left(X^tX\right)^{-1},$$

$\left(X^tX\right)^{-1}$ is a symmetric matrix.

Note: $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^t\mathbf{Y}$ ***is called the normal equation.***

Note: $\mathbf{e}^t\mathbf{X}=(\mathbf{y}^t-\mathbf{b}^t\mathbf{X}^t)\mathbf{X}=\left[\mathbf{y}^t-\mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]\mathbf{X}=\mathbf{y}^t\mathbf{X}-\mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}$
$=\mathbf{y}^t\mathbf{X}-\mathbf{y}^t\mathbf{X}=\mathbf{0}$.

Therefore, if there is intercept, then the first column of X is $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. Then,

$$\mathbf{e}^t\mathbf{X}=\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{n}e_i & \cdots \end{bmatrix}=\mathbf{0}$$

$$\Rightarrow\sum_{i=1}^{n}e_i=0$$

Note: for the linear regression model without the intercept, $\sum_{i=1}^{n}e_i$ might not be equal to 0.

Properties of the least square estimate:
Two useful results:

Let $Z_{n\times 1} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$ **be a** $n\times 1$ **random vector,** $A_{p\times n}$ **is a** $p\times n$ **matrix and**

$C_{n\times 1}$ **is a** $n\times 1$ **vector. Let**

$$E(Z) = \begin{bmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{bmatrix}$$

and

$$V(Z) = \begin{bmatrix} \text{cov}(Z_1, Z_1) & \text{cov}(Z_1, Z_2) & \cdots & \text{cov}(Z_1, Z_n) \\ \text{cov}(Z_2, Z_1) & \text{cov}(Z_2, Z_2) & \cdots & \text{cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Z_n, Z_1) & \text{cov}(Z_n, Z_2) & \cdots & \text{cov}(Z_n, Z_n) \end{bmatrix}$$
$$= \begin{bmatrix} Var(Z_1) & \text{cov}(Z_1, Z_2) & \cdots & \text{cov}(Z_1, Z_n) \\ \text{cov}(Z_2, Z_1) & Var(Z_2) & \cdots & \text{cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Z_n, Z_1) & \text{cov}(Z_n, Z_2) & \cdots & Var(Z_n) \end{bmatrix}.$$

Then
(a) $E(AZ) = AE(Z),\ E(Z + C) = E(Z) + C$.
(b) $V(AZ) = AV(Z)A^t,\ V(Z + C) = V(Z)$
Note:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}, V(\varepsilon) = \sigma^2 I$$

The properties of least square estimate:

1. $E(\mathbf{b}) = \begin{bmatrix} E(b_0) \\ E(b_1) \\ \vdots \\ E(b_{p-1}) \end{bmatrix} = \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$

2. The variance –covariance matrix of the least square estimate b is

$$V(\mathbf{b})=\begin{bmatrix} Var(b_0) & \mathrm{cov}(b_0,b_1) & \cdots & \mathrm{cov}(b_0,b_{p-1}) \\ \mathrm{cov}(b_1,b_0) & Var(b_1) & \cdots & \mathrm{cov}(b_1,b_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(b_{p-1},b_0) & \mathrm{cov}(b_{p-1},b_1) & \cdots & Var(b_{p-1}) \end{bmatrix}$$

$$=\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

[Derivation:]

$$E(\mathbf{b})=E\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\right]=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE(\mathbf{y})=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}=\boldsymbol{\beta}$$

since

$$E(\mathbf{y})=E[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}]=\mathbf{X}\boldsymbol{\beta}+E(\boldsymbol{\varepsilon})=\mathbf{X}\boldsymbol{\beta}+\mathbf{0}=\mathbf{X}\boldsymbol{\beta}.$$

Also,

$$\begin{aligned} V(\mathbf{b}) &= V\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}\right]=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tV(\mathbf{y})\ \left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}=\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \end{aligned}$$

since

$$V(\mathbf{y})=V[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}]=V(\boldsymbol{\varepsilon})=\sigma^2I$$

Example

Heller Company manufactures lawn mowers and related lawn equipment. The managers believe the quantity of lawn mowers sold depends on the price of the mower and the price of a competitor's mower. We have the following data:

Competitor's Price	*Heller's Price*	*Quantity sold*
x_{i1}	x_{i2}	y_i
120	**100**	**102**
140	**110**	**100**
190	**90**	**120**
130	**150**	**77**
155	**210**	**46**
175	**150**	**93**
125	**250**	**26**
145	**270**	**69**
180	**300**	**65**
150	**250**	**85**

The regression model for the above data is

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i\,.$$

The data in matrix form are:

$$\mathbf{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix}=\begin{bmatrix} 102 \\ 100 \\ \vdots \\ 85 \end{bmatrix}, \mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{101} & x_{102} \end{bmatrix}=\begin{bmatrix} 1 & 120 & 100 \\ 1 & 140 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 150 & 250 \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

The least square estimate $\mathbf{b}$ is

$$\mathbf{b}=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}=\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y}=\begin{bmatrix} 66.518 \\ 0.414 \\ -0.269 \end{bmatrix}.$$

The fitted regression equation is

$$\hat{y}=b_0+b_1x_1+b_2x_2=66.518+0.414x_1-0.269x_2.$$

The fitted equation implies ***an increase*** in the competitor's price of 1 unit is associated with ***an increase*** of 0.414 unit in expected quantity sold and ***an increase*** in its own price of 1 unit is associated with ***a decrease*** of 0.269 unit in expected quantity sold.

Suppose now we want to predict the quantity sold in a city where Heller prices it mower at \$160 and the competitor prices its mower at \$170. The quantity sold predicted is

$$66.518+0.414\cdot 170-0.269\cdot 160=93.718.$$

4.0 CONCLUSION

The unit concludes that multivariate analysis is a collection of methods used for analyzing data in which a number of observations are available for each object. These techniques are largely empirical and deal with the reality, they have the ability to analyse difficult data, help in various types of decision-making and are particularly important in social science research.

5.0 SUMMARY

The unit discusses the concept of multivariate statistical analysis and also explained its various types.

6.0 TUTOR-MARKED ASSIGNMENT

i. What is multivariate statistical analysis?

ii. List and explain the classes of multivariate analysis

7.0 REFERENCES/FURTHER READINGS

Bhattarai, K. (2015). *Research Methods for Economics*. UK: University of Hull Business School.

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

Ramayah, T., Ahmad, N.H., Abdul Halim, H., Zainal, S.R.M., & Lo, M. (2010). Discriminant analysis: An illustrated example. *African Journal of Business Management, 4*(9).

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.). NY: John Wiley & Sons.

Shiker, M. A.K. (2012). Multivariate statistical analysis. *British Journal of Science, 6* (1).

MODULE FIVE: RESEARCH ORGANISATION AND REPORTING

UNIT ONE: MANUSCRIPT STRUCTURE AND CONTENTS

CONTENTS

1.0 INTRODUCTION

A scholarly manuscript must be organized and presented clearly and logically. Apart from knowing what goes where in the manuscript but also how to construct each of the elements so as to logically communicate the author's message. A research manuscript usually contains the following key elements:

i. Title
ii. Abstract
iii. Introduction
iv. Literature Review
v. Materials & Methods / Methodology
vi. Results & Discussion
vii. Summary Conclusion and Recommendations

Structure of a research work using chapters should consist of **chapter one, introduction:** This consists of

- background to the study;
- statement of the problem;
- objectives of the study;
- hypotheses of the study;
- significance or justification of the study (The term justification of the study is normally used at proposal stage while significance of the study is used at final defence stage.);
- scope of the study;
- limitations of the study; and
- scheme of chapters.

This consists of background to the study; statement of the problem; objectives of the study;

hypotheses of the study; significance or justification of the study (The term justification of the study is normally used at proposal stage while significance of the study is used at final defence stage.); scope of the study; limitations of the study; and

scheme of chapters.

Chapter two which is theoretical framework and literature review, consists of

- conceptual framework;
- theoretical framework; and
- review of related literature.

Chapter three which is methodology may consist of

- population of the study;
- sources and type of data to be used;
- research design (sampling method used in selecting the objects of study, sample size of the objects);
- method of data collection;
- methods of data analysis (descriptive and inferential methods – descriptive and estimation procedures):
- pre-estimation tests;
- variables measurements;
- model specification; and
- post-estimation tests.

Chapter four consists of

- data presentation,
- Data analysis and interpretation, and
- discussion of findings.

Chapter five is on summary, conclusions and recommendations will consists of

- summary of major findings;
- conclusions;
- policy implications;
- recommendations; and
- suggestions for further study.

2.0 OBJECTIVES

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

- Know how to organize a research work.

3.0 MAIN CONTENT

A research manuscript usually contains the following key elements:

3.1 THE TITLE

The title serves as an advertisement for an article. Just as a newspaper reader skims through the headlines to find what he wants to read, so the busy scientist skimming a journal or any research work will have his curiosity aroused by a business-like title. The title is the first item read by the supervisor, editor and reviewers for journals. They will thus be critical and may even set a limit to length. The shorter the better, but a title must also be clear, and as full as possible. The longer the title, the more likely it is to contain wasted words. For example: A study of …, Observations on …, Some characteristics of…. Some words are also repetitious. Editors and reviewers usually weed these out, but it is better if the author does it himself! Titles are used for indexing articles. Keywords (up to about six) are now asked for by many journals. These are listed on the title page, and make it easier to find article topics, which cannot be put in the title. Abbreviations that are not defined in an index and those that are not familiar to all potential readers should not be used in the title.

An article which scientific encryptions, $CO_2, (CV_U, Z_n)5(OH)r(CO_3)2$ in the title is clearly for specialist readers only. Therefore, except for such readers, formulae should not be used. Also, proprietary names may not be used Slang must not be used. The main reason for this is the difficulty they make for indexing. A good title should be sufficiently descriptive and well-constructed. It should say all that is necessary, with no vague or unnecessary words. It should read easily, without awkward phrases or difficult combinations of sounds. A mistake in the title, may lose your readers; or (even worse) it may deny your paper to someone who could use it. The title must sound (and the words flow) well, when read aloud. A working title must be assigned before one works on a paper, and the author must be prepared to alter or produce a new one as the work proceeds. When the project is finished and one has to decide on a final title, the services of an expert (or other authority) need to be engaged, both for appropriateness, and for good English.

3.2 Abstract

The abstract is a tiny version of the article that extracts the important points.

It allows readers to survey the contents of an article quickly and, like a title, it enables persons interested in the document to retrieve it from abstracting and indexing databases.

A well-prepared abstract can be the most important single paragraph in an article. Most people have their first contact with an article by seeing just the abstract, usually in comparison with several other abstracts, as they are doing a literature search. Readers frequently decide on the basis of the abstract whether to read the entire article. An abstract of a report of an empirical study should describe:

a. *the problem under investigation, in one sentence if possible;*

b. *the essential features of study method*

c. *the basic findings, including effect sizes and confidence intervals and/or statistical significance levels; and*

d. *the conclusions and the implications or applications.*

A good abstract is:

- **accurate:** that correctly reflects the purpose and content of the manuscript. Do not include information that does not appear in the body of the manuscript.
- **Non-evaluative:** Report rather than evaluate
- **coherent and readable:** Write in clear and concise language. Use verbs rather than their noun equivalents and the active rather than the passive voice (e.g., investigated rather than an investigation of; The authors presented the results instead of Results were presented). Use the present tense to describe conclusions drawn or results with continuing applicability; use the past tense to describe specific variables manipulated or outcomes measured.

- **concise:** Be brief, and make each sentence maximally informative, especially the lead sentence. Begin the abstract with the most important points. Do not waste space by repeating the title.

Abstracts seldom make exaggerated claims, do not fall into that trap

3.3 Introduction

The Introduction should take the reader to the body of the paper, as he is keen to find out if the promise made in the abstract is fulfilled. It should tell the reader all he needs to know before he plunges into the melee of interesting and verifiable details. Introducing an article properly should include a full explanation as to why the research was undertaken, indicating its value to the world (outside the research community); description of how the problem was approached and analysed; and a definition of the scope and limitations of the research. Chapter One (introduction) consists of background to the study; statement of the problem; objectives of the study; hypotheses of the study; significance or justification of the study scope of the study; limitations of the study; and scheme of chapters.

3.3.1 Background to the study

This aspect is concerned with the basis for the study, which should contain only the issues that motivate a researcher to conduct study on a given aspect. At the end, one has to link the issues with his topic (project title) in the last paragraph under this section. The author identifies in this section, the wider issues underlying the research problem and question. This is typically the longest section of the first chapter.

3.3.2 Statement of the problem

In empirical studies, this usually involves stating specific question and describing how these were derived from theory or are logically connected to previous data and argumentation. The problem can be detected by looking at and highlighting the efforts been made by various researchers to circumvent the problem under consideration but yet they failed. From then, one draws his research questions. For example, some researchers, in the process of investigating the problem have been unable to capture some critical variables, or they use inappropriate methods of data analysis, or they use small sample size, or the problem has not been investigated in a given country or area at all. Then one should draw research questions from the problem stated and state why the problem deserves new research

3.3.3 Objectives of the study

This section should contain what a researcher wants to achieve or find out in accordance with his research questions. The objectives should contain a broad objective and specific objectives.

3.3.4 Hypotheses of the study

The hypotheses are basically guesses concerning the outcome of the research. They should be measurable and should allow a definite judgment to be made once the data have been collected. When evaluating hypotheses, the null form of the hypothesis is typically used. The formulation of hypotheses is in accordance with the objectives of the study.

3.3.5 Research Questions

The statement of the problem should lend itself to translation into a research question that asks precisely what this study must answer in order to solve the research problem and achieve its purpose.

3.3.6 Scope of the study;

This should contain the boundary within which the research will be conducted. A researcher will indicate only the aspects that his study will contain, all other things are not taken care of by the researcher. He must also justify the scope of the study by given reasons for limiting himself to such aspects.

3.3.7 Limitations of the study;

This section should contain the factors that a researcher expects to reduce the quality of his research findings. You must indicate the measures or steps you intend to take to mitigate the influence of the limitations on the quality of your findings. All personal problems, and the likes should not fall under limitations of the study.

3.3.8 Scheme of chapters;

This section should contain the highlight of all the aspects under each chapter as they are contained in the table of contents. This study has been divided into five chapters. Chapter one consists of background to the study, statement of the problem, objectives of the study, hypotheses of the study, significance of the study, scope of the study, limitations of the study and scheme of chapters.

3.4 Literature Review

This section discusses the previous research and theory in which the researcher discovered and developed the research problem. It will show the relevance of the particular theoretical perspective or framework for identifying the issues, variables, phenomena, or key factors to investigate, including the significance of the problem. It will equally synthesize and critique the literature reviewed, showing both the main foundation points for your work and the opposing viewpoints, controversies in interpretation, or contrary findings relevant to the study. Chapter Two also justifies the selection of the particular methods of data collection, by discussing how the previous research supports (either theoretically or practically, that is, by using them) the use of those methods for obtaining data about the research question of the study. It is mainly divided into three parts namely:

3.4.1 Conceptual framework / literature

One may review different concepts related to the study.

3.4.2 Theoretical framework / literature

This should contain a theory or theories explaining the aspect under study. The theory will be used as a foundation or basis of the study. One should also explain which of theories he reviews would be adopted for the study. Do not blur or blend theoretical frameworks unless they can be authentically integrated and unless the objective of the paper is best served by their integration. In that case, a careful description of all the relevant theories in terms of their major references will be written.

3.4.3 Empirical Literature review

This should contain review of the previous studies in quantitative terms on the aspect under study. Do not simply string one study after another in a random fashion, even if they are well summarized and evaluated: Follow your organizing principle. By following your chosen organizing principle or logic, you will help your reader follow the flow of your own thinking about how you approach the study and its elements. Whatever the organizing principle you chose, follow it strictly and if possible, use section sub-headings to keep the reader oriented. Take into cognizance sample size, type of data an author uses, place of study, the method of data analysis he uses, variable measurement, the results he gets, policy implications and suggestions for further study. Summarize the whole article in your own words. To gain more understanding, read the "literature review" sections of a number of research articles.

3.5 Methodology

The central aim of this section is to show the reader how the researcher arrives at results as well as its reliability and the validity. Therefore, sufficient details need to be provided that will allow for a repetition (if necessary) of the research process. It is important to justify one’s methods. This is especially true if more than one method was possible. If your manuscript is an update of an ongoing or earlier study and the method has been published in detail elsewhere, you may refer the reader to that source and simply give a brief synopsis of the method in this section. This section should consist of population of the study, sources and type of data to be used, research design (sampling method used in selecting the objects of study, sample size of the objects, method of data collection), methods of data analysis (descriptive and inferential methods – descriptive and estimation procedures): pre-estimation tests (give examples) variables measurements, and model specification, and post-estimation tests.

3.6 Data presentation, analysis, interpretation, and discussion of results

This section should contain presentation and analysis of the data collected by testing the hypotheses, interpretation and discussion of results. In the discussion of results section, one is expected to relate his/her findings with those of other authors he/she reviews their works in the review of literature section, and justifies his/her findings in case of divergence.

3.7 Summary conclusion and recommendations

3.7.1 Summary of findings should contain, type of data used, sample size, method of analysis and summary of the study's findings only.

3.7.2 Conclusion should contain conclusions based on the findings of the study. **Policy implications** will consist of the implication of each finding to the dependent variable and the economy in general.

3.7.3 Recommendations consist of the suggestions on how to solve the problem under study based on the findings of the study. **Suggestions for further research** consists of the advice that an author will give for further investigation of the matter on the basis of the limitations of his findings.

4.0 CONCLUSION

In this unit, we explained the step-by-step guide to writing scholarly research article and reports. From the foregoing, we may conclude that scientific research report must be clear, concise, and above all be ordered with a clearly discernible sequence in the way items are arranged for discussion. Writers have to decide beforehand what part goes where, and whether the most important items are positioned at the beginning, in the middle, or at the end. There must be logical coherence between one point and the next.

5.0 SUMMARY

A research manuscript usually contains title, abstract, introduction, literature review, methodology, results and their discussion as well as the summary conclusion and recommendations.

6.0. TUTUR-MARKED ASSIGNMENT

Select any topic of your interest. Write a manuscript suitable for publication.

7.0 REFERENCES/FURTHER READINGS

Abu-Rizaiza, O. (2009). *How to write scientific Articles: A handbook for Non-native Speakers of English (1st ed.)*. Jeddah, KSA: King Abdulaziz University Press.

Bhakar, S. S., & Nathani, N. (2015). *A Handbook on writing Research Paper in Social Sciences*. New Delhi, India: Bharti Publications.

Bhattarai, K. (2015). *Research Methods for Economics*. UK: University of Hull Business School.

Garba, T. (n. d.) Research Method. ECO 903 Lecture notes, *Department of Economics, Usmanu Danfodiyo University Sokoto.*

Kabir, S.M.S. (2016). Writing Research Report. In S. M. S. Kabir (Ed,), *Basic Guidelines for Research: An Introductory Approach for all Disciplines* (1st ed., pp 500-518). Chittagong, Bangladesh: Book Zone.

Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Delhi, India: New Age International.

UNIT TWO: BASICS OF CITING SOURCES

CONTENT

1.0 INTRODUCTION

There are numerous styles of referencing, the most used style in Nigerian research and mostly in the social sciences is the American Psychological Association (APA). The style is a widely used author-date system of referencing or bibliographic citation. Three main reasons for citing sources are:

- Academic Ethics
- Scholarly Credibility
- Source Retrieval.

Sources are cited in two places**:** The In-text Citation and reference list. Also, different sources such as books, periodicals online sources etc. have a varied ways of citation. We shall list and explain each citations and sources with lucid examples.

2.0 OBJECTIVES

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

- Know how to make in-text citation.
- Know how to report sources on the reference list

3.0 MAIN CONTENT

3.1 In-text Citation

This is a short citation in the body of your work. It entails giving a short citation (author last name and year of publication) in the body of your paper. Even though you have put someone else's ideas or information in your own words (i.e. paraphrased), you still need to show where the original idea or information came from. This is all part of the academic writing process. When citing in text within an assignment, use the author/s (or editor/s) last name followed by the year of publication. In-text citation comes in two forms: the running text and parenthetical (citations in brackets). Here are examples:

3.1.1 Running text

i. ***Single author:*** Adhama (2020) investigated the effect of inflation on income………

ii. ***More than one author:***

Mustapha and Adhama (2011) stated that human capital development is a vital determinant of growth.

Note: To cite in *running text* means to place the author's surname in the sentence outside the parentheses but always place date in parentheses.

3.1.2 Parenthetical

Single author example: Angola is the second largest contributor of carbon emission in Africa (Adhama, 2020).

More than one author example:

Diversification of the economy is the only solution to Nigeria's over dependence on crude oil (Akin & Ndubueze, 2018).

Note: In this case the authors surname(s) are followed by a comma (,) along with the date of publication are enclosed in parentheses. Also ampersand "&" not "and" is used in parenthetical in-text citation.

In the subsequent subsections we shall discuss two in-text citations using multiple authors.

3.2 Reference list

This a full citation on the reference page at the end of your document, where you will provide author(s), publication date, title and publisher information. Everything you have cited in text appears in your reference list. The exception is when citing a personal communication. Personal communications are cited in text but do not appear in the reference list.

3.2.1 Four Elements of a reference list

APA reference list is easier to learn by noting its four basic elements that appear in the following order:

1. Author 2. Publication date 3. Title 4. Publisher information

In the following reference list entry for a journal article, the four elements are illustrated in Table 3.1.

Table 3.1: Four elements of a reference list entry

1. Author	2. Date	3. Title	4. Publisher information
Aliyu, M. J.	(2020).	Inflation and economic growth.	*Journal of Economic Research*, *4*(11), pp 90-101 .

Authors can be single or multiple. We have discussed this in detail in the following subsections. The date is usually expressed only as a year—the year of publication—and is placed in parentheses after the Author Element. *This element cannot be left blank.* If no publication year is provided, type the letters **n.d.** inside parentheses. Title Elements can be title of a work published *inside* another work (for example, a chapter inside a book or an article inside a journal or title of a work published as by itself—not inside another work. The fourth element, Publisher, varies according to type of source. For **Books** the publisher information element contains: (a) City, (b) State, and (c) Publisher name. For **Periodicals**: The publication information element contains: (a) *Title of journal, (b) volume number*, (c) issue number (if provided), (d) page number range, and (e) doi number, (if provided). Each element ends with a period. Commas and other punctuation marks may appear *within* an element. In the case of the *Author* element, the period after the author's initial serves as the final period for the element.

3.2.2 Formatting a reference list

1. The reference list is arranged in alphabetical order of the authors' last names.

2. If there is more than one work by the same author, order them by publication date – oldest to newest (therefore a 2004 publication would appear before a 2008 publication).

3. If there is no author the title moves to that position and the entry is alphabetised by the first significant word, excluding words such as "A" or "The". If the title is long, it may be shortened when citing in text.

4. Use "&" instead of "and" when listing multiple authors of a source.

5. The first line of the reference list entry is left-hand justified, while all subsequent lines are consistently indented.

6. Capitalise only the first word of the title and of the subtitle, if there is one, plus any proper names – i.e. only those words that would normally be capitalised.

7. *Italicise* the *title* of the book, the *title* of the journal/serial and the *title* of the web document.

8. Do not create separate lists for each type of information source. Books, articles, web documents, brochures, etc. are all arranged alphabetically in one list.

3.3 Referencing a Book

The following are some of the guidelines in referencing a book:

1. Author/s or Editor/s last name (surname) appears first, followed by initials example, (Bloggs, J.).

2. Year of publication in brackets (2010).

3. Full title of the book. Capitalise only the first word of the title and the subtitle, if any, and proper names. Italicise the title. Use a colon (:) between the title and subtitle.

4. Include the edition number, if applicable, in brackets after the title or subtitle (3rd ed.) or (Rev. ed.). Note: No full stop, after the title, if there is an edition.

5. Place of publication. Always include the city and 2-letter state code when published inside the USA, and the city & country, if published outside the USA (Fort Bragg, CA or Auckland, New Zealand or Benalla, Australia or Weybridge, England). If there are two or more places included in the source, then use the first one listed.

6. Publisher's name. Provide this as briefly as possible. Do not use terms such as Publishers, Co., or Inc. but include the words Books & Press. When the author and the publisher are the same, use the word *Author* as the name of the publisher.

Examples:

Book – one author

Collier, A. (2008). *The world of tourism and travel*. Rosedale, New Zealand: Pearson Education New Zealand.

Book – editor(s) & edition

Collins, C., & Jackson, S. (Eds.). (2007). *Sport in Aotearoa/New Zealand society* (2nd ed.). South Melbourne, Australia: Thomson.

Book – author & publisher are the same

Mid Central District Health Board. (2008). *District annual plan 2008/09*. Palmerston North, New Zealand: Author.

Chapter in an edited book

Dear, J., & Underwood, M. (2007). What is the role of exercise in the prevention of back pain? In D. MacAulay & T. Best (Eds.), *Evidence-based sports medicine* (2nd ed., pp. 257-280). Malden, MA: Blackwell.

3.4 Serial/journal articles

The following are some of the guidelines in referencing a journal article and other periodicals:

1. Author/s last name (surname) first, followed by initials.
2. Year of publication in brackets. (2012)
3. Title of article. Capitalize only the first word of the title and the subtitle, if any, and proper names. Use a colon (:) between the title and subtitle.
4. Title of the serial/journal in full in *italics.*

5. Volume number, in *italics*. Do not use “Vol.” before the number.

6. Issue number. This is bracketed immediately after the volume number but not italicized.

7. Include all page numbers.

8. Include any Digital Object Identifiers [DOI].

Examples

Serial / journal article (print)

Mark, P. (2010). The Relationship between Public Expenditure and Human Capital Development in Nigeria*: Journal of Economics, 10*(7), 2.

Serial / journal article (online from a database – e.g. EBSCO or Newztext)

Musa, A., John, C., Amin, S., & Bashir, A. (2019). Analysis of the Relationship between Inflation and Economic Growth in Nigeria. *Dutse Journal of Social Sciences and Economic Research, 2(4),* 1543-1610. Retrieved from http://www.djse.com/bw/journal.asp?ref=0243-1057

Newspaper article

Mukhtar, L. (2019, May 2). Nigeria Fights Depression. *Daily Trust*, p. 2.

Newspaper article (no author)

Unemployment in Nigeria. (2015, March 3). *Daily Trust*, p.11.

In-text citation:

Shorten the title and enclose in quotation marks.

(“Unemployment in Nigeria”, 2015).

Newspaper (online)

Rabiu, M. (2018, April 6). Small Scale Enterprises and Unemployment in Nigeria. *The Guardian*. Retrieved from http://www.stuff.co.nz/small/enterprises /6038621/Small scale-could-decrease-unemployment

Online– DOI

Many publishers, databases and online journals use DOIs. They are alpha-numeric codes that usually appear on the first page of the article. Copy the DOI exactly as it appears.

Garba, K., James, M., & Mark, Z. (2018). Green House Gas Emission and Environmental Degradation in Nigeria. *Journal of Social Science, 3*(7), 453-569. doi:10.1178/j.jsosc.2018.03.007

Conference paper (online)

James, I. (2014). Nexus between Human Capital Development and Economic Growth: ARDL Approach. Conference proceedings of the National Open University Nigeria, Abuja.Retrievedfromhttp://www.noun.ed.ng/conferences/papers/Proceedings_2014.pdf

3.5 Internet sources

The following are some of the guidelines in referencing a journal article and other periodicals:

1. Author/s of the document or information – individual or organisation/corporate author.
2. Date of publication. If no date is available, use (n.d.).
3. Title of the document or webpage in *italics.*
4. Complete & correct web address/URL.

Examples:

Software (including apps)

UBM Medica. (2010). iMIMS (Version1.2.0) [Mobile application software]. Retrieved from http://itunes.apple.com

Thesis (online)

Mann, D. L. (2010). Vision and expertise for interceptive actions in sport (Doctoral dissertation, The University of New South Wales, Sydney, Australia). Retrieved fromhttp://handle.unsw.edu.au/1959.4/44704

3.6 Multiple Authors

If a source has only **one or two authors**, name the authors and publication date each time the source is cited. If a source has **3 to 5 authors,** name all the authors in the first in-text citation; but in all other in-text citations, give only the name of the first author, followed by *et al.* (no italics) and the publication date..

For example, a work with the following four authors: Faiza Adhama Mukhtar; Aminu Muhammad Mustapha; Henry Jack; and Ngozi Ndubuisi Adewale will feature on a reference list entry like this . . .

Mukhtar, F. A., Mustapha, A. M., Jack, H. & Adewale, N. N. (2015).

It would be cited in the text the first time as:

Mukhtar, Mustapha, Jack, and Adewale (2015) recommends that . . .

and thereafter would be cited this way:

Mukhtar et al. (2015) recommends that. . .

If a source has **6 or more authors**, name only the first author, followed by *et al.,* in *all* in-text citations, even the first citation.

For example, a reference list entry like this . . .

Morgan, I, Abdulmalik, M K., Inuwa, S., Chun, H., Lee, L., Mark, Z., & Sheba, R. (2001) would be cited each time like this:

Morgan et al. (2001) examined . . . –

or- parenthetical as. . . (Morgan et al., 2001)

3.7 Themes

A literature review, for example, is organized by themes. Instead of separately summarizing each article you read, you should write about themes that emerged from your review of the literature. For instance, you might write something like this:

Nigerian government recognized the need to issue out credit to small scale enterprises as a way to cut down unemployment in the country.

In that sentence you are writing in your own word, stating an assertion for which you found support in the literature. You need to cite each source you read that supported that assertion. A good place to cite in a case like this is at the end of the sentence in parentheses, like this:

Nigerian government recognized the need to issue out credit to small scale enterprises as a way to cut down unemployment in the country (Azubilke (2020); Benedict & Ahmad, 2013; Clement, 2016; Johnson, 2017).

Note that the sources are arranged alphabetical.

SELF ASSESSMENT EXERCISE

Search for scholarly articles and observe their formatting of in-text and reference lists.

4.0. CONCLUSION

In academic writing, sources are cited in two places**:** the In-text citation and reference list. Also, different sources such as books, periodicals online sources etc. have a varied ways of citation.

5.0. SUMMARY

In this unit we explained the basics of citing sources in academic writing using the APA style of referencing.

6.0. TUTUR-MARKED ASSIGNMENT

Review the general rules of referencing with relevant in-text and reference list entry examples for books and serial/journal with multiple authors using the 6th ed. of APA

7.0 REFERENCES/FURTHER READINGS

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Prentice, C. (2013). *Introduction to the APA* (6th ed.). Minnesota: Saint Mary's University.

UCOL Student Success Team. (2015). *A Guide to APA 6th ed. Referencing Style.* Matauranga: Author.