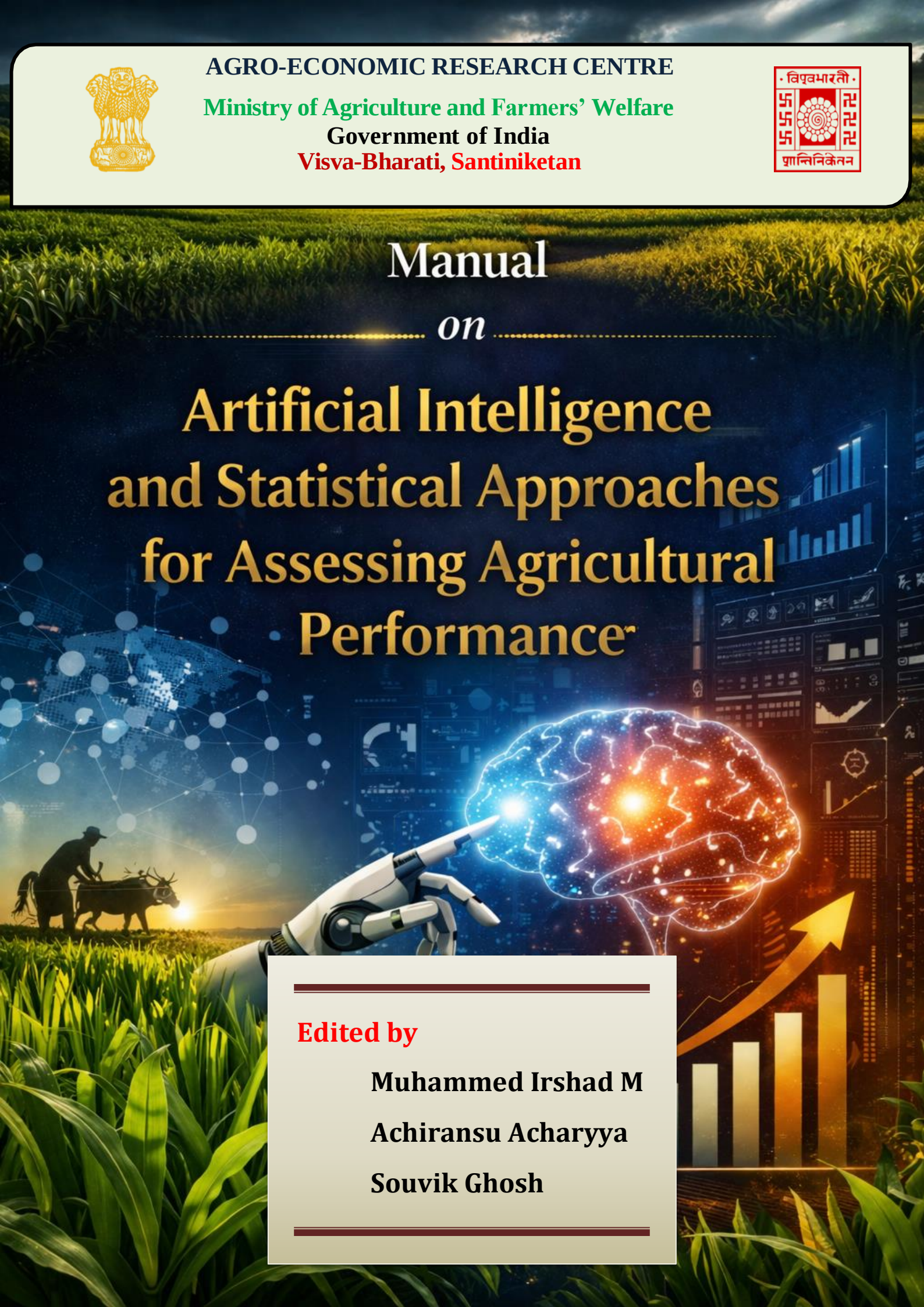
AGRO-ECONOMIC RESEARCH CENTRE
Ministry of Agriculture and Farmers' Welfare
Government of India
Visva-Bharati, Santiniketan

Manual on Artificial Intelligence and Statistical Approaches for Assessing Agricultural Performance



Edited by

Muhammed Irshad M

Achiransu Acharyya

Souvik Ghosh

Manual

on

Artificial Intelligence and Statistical Approaches for Assessing Agricultural Performance

Edited by

Muhammed Irshad M

Achiransu Acharyya

Souvik Ghosh

Agro-Economic Research Centre

(For the States of West Bengal, Sikkim, and Andaman & Nicobar Islands)

Ministry of Agriculture and Farmers Welfare, Government of India

Visva Bharati

Santiniketan

West Bengal

2026

Citation:

Irshad, M. M., Acharyya, A., & Ghosh, S. (Eds.). (2026). *Manual on Artificial Intelligence and Statistical Methods for Assessing Agricultural Performance*. Agro-Economic Research Centre (for the States of West Bengal, Sikkim and Andaman & Nicobar Islands), Visva-Bharati, Santiniketan, West Bengal, India.

ISBN: 978-81-989525-0-9

Assistance

Dr. Sridev Adak
Mr. Saptarsi Chakraborty
Dr. Rishav Mukherjee
Mr. Mehdi Hasan
Dr. Sreejit Roy

Secretarial Services

Mr. Nityananda Maji
Mr. Munshi Abdul Khaleque
Mr. Deb Sankar Das
Mr. Dibyendu Mondal
Mr. Parag Mitra
Mr. Bimal Kumar Singha
Mr. Sunil Hansda
Mrs. Susmita Biswas

Published By: Director (Hony.), Agro-Economic Research Centre, Visva-Bharati, Santiniketan, West Bengal, E-mail: dir.aerc@visva-bharati.ac.in

Preface

This manual on ***Artificial Intelligence and Statistical Approaches for Assessing Agricultural Performance*** is prepared to serve served as a comprehensive guide to undertake statistical analysis in agro economic research. The manual outlines the research design, methodologies, and data analysis techniques relevant to agricultural performance assessment.

The manual covers key aspects of research design, including quantitative, qualitative, and mixed-method approaches, hypothesis formulation, research strategies, and sampling techniques. It also outlines data collection methods such as questionnaires, interviews, observation, and sociometry, along with approaches to minimize errors and biases. It further discusses data preparation, including handling missing values, detecting outliers, and testing assumptions, along with descriptive statistics and exploratory data analysis. Regression analysis, including linear and nonlinear models and diagnostics, is also addressed. Additionally, the manual includes time series modelling using ARIMA, SARIMA, and ARFIMA models, an overview of R software, and applications of Artificial Intelligence techniques for classification, regression, and forecasting.

The editors sincerely express their heartfelt gratitude to Dr. Ranjit Kumar Paul, National Fellow, ICAR-IASRI, and Dr. Md Yeasin, Scientist, ICAR-IASRI, for their valuable contributions in providing essential content for this manual.

Our sincere gratitude is extended to Dr. Probir Kumar Ghosh, Hon'ble Vice-Chancellor, Visva Bharati, for his kind guidance and encouragement. We also acknowledge the support and valuable guidance of the Adviser (AER Division), Ministry of Agriculture & Farmers' Welfare, Government of India, New Delhi, in our endeavours. The support of other officials in the Ministry has also been invaluable in undertaking the activities of the Centre.

This manual will be useful for scholars, professionals, researchers, and policymakers through providing practical insights and methodological guidance for the application of artificial intelligence and statistical approaches in agro-economic research and the assessment of agricultural performance.

Professor Souvik Ghosh,
Director (Hony.)
Agro-Economic Research Centre
Visva Bharati

RESEARCH DESIGN

Prof. Souvik Ghosh
Hony. Director
Agro-Economic Research Centre
Visva Bharati
Email: dir.aerc@visva-bharati.ac.in

Research Design

A research design is the plan or proposal to conduct research which entails the intersection of philosophy, strategies of inquiry and specific methods. Choosing a research design involves various intricate decisions based on the purpose and objectives of the study. Creswell (2009) has suggested a framework for selecting an appropriate research design based on philosophical paradigms, along with strategies of inquiry and research methods.

- It provides answers to various questions.
- It acts as a standard and guidepost which helps the researcher in measuring his shortcomings and deviations in actual research.
- It helps in carrying out research validity objectively, accurately, and economically

.

Four parts of a research design:

1. Sampling design
2. Observational design
3. Statistical design
4. Operational design.

Steps in preparing research design:

- Setting up of objectives and formulation of a research problem
- Review of literature
- Selection of hypothesis
- Designing the experiment
- Processing, analysis, and interpretation
- Reporting and publications.

The first step in designing social research is to perceive a problem: either theoretical or applied.

This may happen –

- When there is a gap in the results of earlier investigation.
- When the results of several enquiries disagree.
- When a fact exists in the form of a bit of unexplained information.
- When there is a desire for innovation.

Types of Research Strategies/ Designs

Strategies of inquiry are types of quantitative, qualitative, and mixed methods, designs, or models that provide specific direction for procedures in a research design.

I. Qualitative Research Strategies

Quantitative research can be of experimental and non-experimental strategies.

A. Experimental strategies

The researcher manipulates the conditions of what the subjects will experience by structuring actual situations, introducing, or controlling certain variables in order to measure their effect on each other, or by systematically imposing or withholding specified conditions. Experimental strategies are of two types - between group and within group designs (Creswell, 2012)

a. Between-group designs

The between-group design is used when the researcher wishes to compare two or more groups.

(i) True experimental design

Cause-effect analytical experiments which involve random assignment of subjects and conditions to the subjects for objectively measuring the research phenomenon by controlling external interference and errors. For example, the influence of subsidies on adoption of organic farming may be studied by assigning subsidy (experimental group) and without subsidy (control group) and assessing adoption by comparing both groups.

(ii) Quasi-experimental design

Quasi-experiments are non-randomised experiments conducted in non-laboratory situations where independent variables are manipulated to assess their influence on dependent variables. Quasi-experiments differ from true experiments in that subjects are not randomly assigned to conditions. It has both pre-, post-, as well as post-only test designs

Quasi-experimental designs are of three types (Creswell, 2009):

- Non-equivalent (Pre-Test and Post-Test) Control-Group Design – The experimental and control groups are selected without random assignment. Only the experimental group receives the treatment.
- Single-Group Interrupted Time-Series Design – The researcher performs pre- and post-tests in a single group prior to, and after, treatments.
- Control-Group Interrupted Time-Series Design – A modified version of single group time-series design where two non-random groups of participants (control and treatment) are studied over a period of time.

(iii) Factorial designs

In the factorial design, the researcher studies two or more categorical, independent variables that are analysed at two or more levels (Vogt, 2005). The purpose of this design is to study the independent and simultaneous effects of two or more independent treatment variables on an outcome.

b. Within-group designs

Within-group design is employed when the total number of participants is very limited. This design analyses the variations present within the group.

(i) Time series experiments – In this design, studies are conducted on a single group over a period of time, through a series of pre-test and post-test observations. For example, diffusion of new agricultural technologies guided by a series of technology interventions can be studied within a specific geographical or social system over years.

(ii) Repeated measures experiment – In a repeated measures design, multiple treatments are administered on a single group and the researcher compares its performance across these experiments.

(iii) Single-subject designs – Single-case designs are usually single-subject or group research (without control), used to evaluate the extent to which a causal relation exists between introduction of an 'intervention' and change in a specific dependent variable. For example, a study focusing on the influence of quality seed material provision to a single farmer or a group of farmers focusing on the livelihood security aspect may be termed as a single-case research design.

B. Non-experimental strategies

In non-experimental designs, the researcher may examine the relationship between things without manipulating the conditions. Examples include:

(i) Descriptive research – Descriptive research provides a detailed summary of an existing phenomenon by assigning numbers to characteristics of objects or subjects involved.

(ii) Comparative research – In comparative designs, the researcher investigates the presence of difference between two or more groups on the phenomenon studied. Comparative research is often multidisciplinary and utilises quantitative techniques to study the phenomenon. Comparative research methods are used extensively in cross-cultural studies.

(iii) Correlational research – Assessing the direction and strength of association between two or more phenomena, without manipulating them. Types of correlational designs: concurrent or explanatory and predictive (Creswell, 2009).

(iv) Survey research design - The survey research design provides a quantitative description of trends, attitudes, or opinions of a population by studying a sample of that population. It includes cross-sectional and longitudinal studies using questionnaires or structured interviews for data collection, with the intent of generalizing from a sample to a population (Creswell, 2009).

Non-experimental research may be divided into Exploratory research and Descriptive research

Exploratory research: It aims at exploring the possibility of doing research on a certain subject where due to paucity of existing knowledge framing and testing of hypotheses are difficult. The aim of such studies is to collect enough data. The researcher needs to be very receptive and alert in identifying the clues. Data are generally collected from the secondary sources, surveys and experienced individuals on insight stimulating case studies.

Descriptive research: It aims at answering what and why of the current state of some system. Description and explanation are its two aims. Having described the system, the researcher may also be interested in explaining why the system arrived at its present state. He may investigate the factors responsible for it. When a researcher thus identifies a problem based on its association with some related phenomena, it is called **Diagnostic research.**

A descriptive research is generally enabled to make any predictions for the simple reason that it does not provide setting up and testing of hypothesis through controlled experiments. As such, it is unable to prove cause and effect relationship between any two parameters / phenomena. This is achieved by carrying out experimental research.

A descriptive research can be either static, dynamic of historical in nature. It is static when it involves a single measurement of the phenomena in question. e.g. a telephone survey assessing the public's attitude toward energy crisis. It is dynamic in nature if it goes beyond the simple measurement of variable and examines the relationship among variables by using either cross sectional or longitudinal design.

A cross-sectional study is one which collects data about various variables of the sample at one point of time in order to uncover relationships existing among those variables. Thus, a study to examine the relationship between job satisfaction and style of leadership or between similarity of preferences. Although experimental control is not there, researchers still incline to speak w.r.t. independent and dependent variables. For e.g: difference in the independent variable (style of leadership) are used to explain difference in the dependent variable (Job satisfaction).

A longitudinal/panel study is one which collects data about the sample survey /same sample over a period of time so that possible relationships among variables can be revealed by examining the changes take place during that time. In the area of marketing, 'a consumer panel' is an example of longitudinal study in which the same sample of households is studied for one or more aspects of consumer behaviour for duration of time. It is a type of time series study. The common subject which is again and again studied constitute a panel for the researcher.

Historical research: It is a critical evaluation and exanimation of past events, development, and experiences. Here, the subject matter for study is always past. Data are mostly drawn from the secondary sources.

- It helps in understanding the roots of all modern theories.
- It helps in establishing the genuineness of the sources of data.

This research suffers from two limitations –

- The history never repeats itself.
- The findings suffer from bias because in the absence of any methodological control the researcher may misinterpret the data in order to validate his hypothesis.

Cross-sectional survey

In a cross-sectional survey design, the researcher collects data at one point in time. Cross-sectional designs are of several types:

• Single group study measuring current attitudes, beliefs, opinions, or practices;
• Comparing attitudes, beliefs, opinions, or practices between groups.

Longitudinal survey

Collecting data over a period of time from single or multiple groups. There are three types of longitudinal designs:

- Trend study – Conducted every year on a specific aspect, but with a different sample. However, the sample size remains the same. For example, assessing trends in farmers' input utilisation in potato over the years.

- Cohort – A group of subjects are identified and a specific phenomenon is studied over a period of time to assess changes. Although the same population is studied each year, the sample from that population is different for each year. For example, studying food grain consumption patterns through cohorts over a period of time.

- Panels – An identical sample selected at the beginning is used for collecting data every year to assess the changes over time. For example, studying career progression of agricultural students who have graduated from a particular university.

Ex post facto

Ex post facto study or after-the-fact research is a category of research in which the investigation starts after the phenomenon occurred, without any intervention from the researcher. It is primarily a quasi-experimental study examining how an independent variable, present prior to the study in the participants, affects a dependent variable. This lack of direct control over the independent variable and the non-random selection of participants are the

most important differences between ex post facto research and the true experimental research design. Most extension research follows the ex post facto approach.

II. Qualitative Research Strategies

Qualitative designs are concerned with describing or interpreting a phenomenon without manipulating its conditions. It is broadly classified as interactive or non-interactive, based on researchers' involvement in the inquiry.

a. Interactive

(i) Ethnography – Strategy of inquiry in which the researcher studies an intact cultural group in a natural setting over a prolonged period of time by collecting, primarily, observational and interview data

(ii) Phenomenological research – The researcher attempts to understand and explain how an individual or a group of individuals experience a particular phenomenon from the individual's or individuals' own Perspective through in-depth interviews.

(iii) Grounded theory – Grounded theory is a qualitative strategy for developing theories where the researcher derives an abstract theory of a process, action, or interaction from the views of participants.

(iv) Narrative research – In narrative research, the researchers explain the lives of individuals, collect, and tell stories about people's lives, and write narratives of individual experiences. Narrative research can be effectively used in documenting indigenous technical knowledge, develop success stories of agricultural technologies, etc., using autobiographies, narrative interviews, and oral histories.

b. Non-interactive

(i) Concept analysis - Concept analysis is a research strategy where concepts, their characteristics and relations to other concepts are examined for the purpose of identifying different meanings of the same concept. For example, concepts such as food security, poverty, livelihood security, etc., can be assessed by comparing the meanings attributed by various stakeholders like farmers, farm women, farm youth, local traders, panchayat officials, etc.
(ii) Historical analysis - Historical analysis is a method of interpreting and understanding the past through a disciplined and systematic analysis. It involves a detailed examination of the 'traces' of past through artefacts, texts, images, and old buildings, etc. For example, indigenous technical knowledge can be collected and elaborated upon by key informant interviews, village records, artefacts, old photographs, and drawings made by elders.

(iii) Policy analysis - Policy analysis is a research strategy for generating information which helps in formulating and implementing policies and then assessing their impact. It uses both quantitative and qualitative methods for collecting and analyzing the information related to policy.

III. Mixed Methods Strategies

The mixed methods research design is a procedure for collecting, analyzing, and combining both quantitative and qualitative methods in a single study or a series of studies to understand a research problem (Creswell and Plano Clark, 2011). Combining quantitative and qualitative strategies will provide a comprehensive view of the phenomenon under study. Various types of mixed-methods researches are described below:

(i) **Convergent parallel design –** Quantitative and qualitative data are collected simultaneously, analysed, and interpreted in order to gain a comprehensive view of the phenomenon. It is possible to offset the biases, errors and gaps in the data collected through one form in this design.

(ii) **Explanatory sequential design –** In this design, quantitative data is collected first and qualitative data is gathered at a later stage for explaining the results of quantitative analysis. Quantitative data results provide a general idea about the phenomenon and qualitative data refines and explains this view.

(iii) **Exploratory sequential design –** In this method, qualitative data is collected first to explore the phenomenon under study, and then the quantitative data are gathered to explain the relationships among elements of the phenomenon explained through qualitative data. This method is widely used in developing new scales.

(iv) **Embedded design –** In this design, both quantitative and qualitative data are collected simultaneously or sequentially, and one form of data is used only as supportive material to justify the results from another set of data. Both first and second set of data are either qualitative or quantitative.

(v) **Transformative design –** This approach uses any one of the above stated mixed methods designs, but fits the data within a transformative framework (Creswell and Plano Clark, 2011). Examples of transformative frameworks are feminism, gender, ethnicity, disability, and racism.

(vi) **Multiphase design –** The researchers or a team of researchers study a problem through a series of separate studies.

IV Action Research Designs

Action research is a problem-oriented design where the researcher systematically gathers information about field practice for improving the effectiveness of field work. Action research can be of two types:

- Practical action research – A small-scale research work narrowly focuses on a specific field problem or issue undertaken in a specific area

- Participatory action research – A social process in which the researcher deliberately explores the relationship between the individual and other people for the purpose of improving the quality of life

- Choosing a research design involves various intricate decisions based on the purpose and objectives of the study.
- A research design is based on the philosophical paradigm of the research, strategies of inquiry and methods.
- The quantitative, qualitative, and mixed method designs provide wide choice and greater flexibility for conducting quality extension research.
- Though extension research is conducted largely as ex post facto, experimental strategies will help to develop theories.

Source: Sivakumar et al. (2017)

Types of Social Research Commonly Followed

There are several types of social research, out of which the following are commonly followed

1. Ex-post facto research
2. Experimental research in laboratories
3. Field experiments
4. Simulation
5. Field studies
6. Survey research
7. Case study

1. Ex-post facto research

The researcher must try to link the relationship. of dependent variable by his observation with the independent variables that have already occurred. The researcher's main function is to study the independent variables in respect of their possible relations to and effects on the dependent variable. A systematic empirical enquiry in which the scientist does not have direct control on independent variables because their manifestation has already occurred or because they are inherently non-manipulable

Limitations:

1. Inability to manipulate the independent Variables.

2. Lack of power to randomize

3. The danger of improper and inappropriate inferences and interpretations despite the drawbacks only this method explores new fields for scientific investigation. For analysing problems in sociosphere by experimental method ex-post facto method can be applied.

2. Experimental research in laboratories

In this method of research, an attempt is made to keep the variance of all or nearly all the possible influential independent variables to the minimum, which is not immediate concern of investigation. This can be achieved by isolating the research in a physical situation, apart from the routine or ordinary living and the manipulating of one or more independent variable under rigorously specified, operationalised, controlled conditions.

Objectives:

- To explore the relations under pure uncontaminated conditions.
- To test the predictions derived from theory and other researches.
- To refine theories and hypotheses and to assist building theoretical Statement.
- High internal validity.
- Controls all extraneous variables.
- Manipulates independent variables.

In short Experimental research in laboratory condition tries to establish validity of hypotheses derived from theory, to explore precise interactions of variables and to control variance under research reconditions.

Weakness:

1. Lack of strength of independent variables.

2. Artificiality of situations.

3. Chances of wrong interpretations.

4. Laboratory setting are contrived situations.

3. Field experiment / experimental research in field

Lesser control is there over independent variables. These are manipulated by the researcher up to the extent which does not destroy the realistic situation. Manipulation is performed under carefully controlled conditions.

It is appropriate for studying complex social influences, processes, and changes in life like settings. The dynamic interrelations of small groups have been fruitfully studied in field experiments. It is also suitable for testing theories and the solutions of problems. Field experiment is a research study done in a realistic situation on which one or more independent

variables are manipulated by the experimenter under as carefully controlled conditions as the situation permits.

Two types of field experiments:

i) Conducted in institutional settings.

ii) Conducted in residential settings.

Limitations:

- Investigator faces many practical difficulties and has no justifiable reason for using contaminated independent variable though he has power to manipulate.
- Attitude of the researcher determine the field of study and its findings.
- Problem of randomization.

4. Simulation

As in the case of a laboratory experiment, in a simulation study, also the setting is deliberately structured to mirror important dimensions of some naturally occurring system. The only difference between the two is that whereas in a laboratory experiment the system being studied is more generic, whereas in a simulation, it is specific.

Advantages:

- There is greater realism in the setting
- There is greater amount of control over external sources of variance
- The researcher enjoys greater ability to manipulate independent variables
- There is higher participant development.

Disadvantage:

- This method is expensive
- The high degree of participant involvement increases the risk of the subjects being psychologically harmed during study.

5. Field Study

Any systematic study which aims at discovering the relations and interactions among variables or testing hypothesis in natural living/Live Situations like committees, school, factories, organisations, institutions, etc and which is ex-post facto in nature is a field study. The essential factor that distinguishes a field study from field experiment is the design of research. Though both carry out in natural setting, a field experiment involves actual manipulation of independent variables by the experimenter in order to find out cause and effect relations with dependent variable whereas in a field study, a researcher has no control whatsoever the independent variables. Since data in a field study are collected at only one point of time, only correlational or cross-sectional analyses of the data are possible. Field study serves several purposes: exploratory, descriptive and hypothesis testing.

Advantages:

1. Very much realistic.

2. Data on large no of variables can be obtained.

Limitations:

1. Cooperation of subjects/organisation difficulties.

2. Ex-post facto in which independent variables are not manipulated, therefore, causal inferences cannot be drawn.

3. Data are likely to contain unknown sampling biases

4. Measurement is not as precise due to influence of confounding variables.

5. The "cross rate" proportion of irrelevant data may be high.

Field studies are of following types:

- Exploratory field study for gaining familiarity with a system to enable the researcher to define a research problem or to develop hypotheses about some phenomenon of the system.
- Descriptive field study: answers what and why of the current state of some system. Seeks to explain why the system arrived at its present state.
- Hypothesis testing field shady: To achieve the desired aims of hypothesis testing, a preliminary, methodological and measurement investigation must be undertaken.

Any scientific study large or small that systematically pursue relations and best hypothesis, those are ex-post facto, that are made in actual life situation will be considered field studies. Any ex-post facto scientific enquiry which is conducted with the aim of exploring the relations and interactions among sociological, psychological, and educational variables in real socio-sphere can be called field study.

Objective:

- Discovers significant variables in the given situations.
- Discovers relations among the variables.
- Lays groundwork for the more systematic and rigorous testing of hypothesis.

Merits:

- Off all types of studies, field studies are close to the real life.
- Field studies are strong in realism significances, strength of variables, theory orientation and heuristic quality.

Demerits:

- Ex-post facto character
- Lack of precision in measuring variables

6. Survey Research

It is extensively used by social and natural scientists. It studies large and small populations by selecting and studying samples chosen from the population to investigate the relative incidence, distribution, and interrelations of sociological and psychological variables. Surveys covered under these definitions are often called sample survey probably because survey research was developed as a separate research, activity along with the development and improvement of sampling procedures.

The social scientific nature of survey research is revealed by the nature of its variables which can be classified as sociological facts, opinions, and attitudes. Sociological facts are attributes of an individual that spring from their membership in social group- sex, education, occupation, race etc. Psychological variables include behaviour, attitude, and opinion. Economic variables include income, saving, asset, liability, expenditure, political variables include political alliance, feelings, affiliation, etc.

Survey research, therefore, does not necessarily refer to sociological and psychological research alone but encompass variables of economic, political, religious nature for the estimation of the incidence and distribution of welfare activities from the economic point of view.

Types of survey:

As per methodological variations of collecting information, surveys can be of many categories:

- Personal Interview
- Mailed Questionnaire
- Panel, Conversation
- Group discussion, etc.

Merit:

- It has great amount of objectivity.
- Provide large amount of information.
- Findings and conclusions have large/great reality.
- Information is accurate within the sampling error.
- It throws some light on important problem. which would otherwise have been hidden.

Demerits:

- Information pertains to only periphery level (sampling level).
- No scope for manipulation of independent variables.
- More time and money involved.
- subject to the Sampling error.
- Mostly “one shot" studies.

Sample survey is an ex-post facto research, in which the researcher simply collects data about certain sociological, psychological characteristics of a sample that represents a population in natural setting.

Comparison between Field study and Sample Survey is given below:

Field study	Sample survey
Directly measure social, psychological factors. Provide detailed and independent picture of interrelations of groups than does the survey.	Direct study of ongoing social and psychological process is not done. Inferences are drawn indirectly through statistical end results.
Restricted in scope and is little concerned with the questions of sampling.	Extensive in scope. Sampling has to be done carefully.

Like field study, a sample survey is also a form of ex-post facto research in which the researcher simply collects data about certain sociological or psychological characteristics of a sample that represents a known population in natural settings. Like field study, a sample survey also aims at:

i. Exploring the existence of some phenomena.
ii. Describing a phenomenon.
iii. Testing a hypothesis.
iv. To predict future conditions
v. To evaluate social programmes
vi. To develop social indicators

Advantages of sample survey:

- Use of a larger representative sample in a survey reduces problem of sample bias and allows the researcher to generalize his results to the parent population
- Data collection can take place in any setting.
- Data are obtained directly from respondents.
- A variety of data collection techniques (interviews, questionnaire & observations) can be used alone or in combination.
- Surveys often yield information that suggests new hypothesis.

Steps in conducting a survey:

1. Planning a survey

- Statement of objectives.
- Knowledge about similar studies.
- Preparation of work schedule and budget.
- Arrangement of publicity.
- Determination of sample
- Preparation of questionnaire/schedule.
- Selection, training, and supervision of field investigators

2. Collection of the facts

3. Analysis of the facts

4. Presenting the facts

5. Reporting

Pilot Survey:

It is the whole survey operation in miniature. It is a careful empirical checking of all phases of the study from the collection of data to their tabulation and analysis. The magnitude of this survey will depend on the available resources of time, money, and personnel, but it should be large enough to permit an initial analysis of the adequacy of questionnaire and of the training, instructions, and supervision of interviewers under field conditions. A pilot survey is thus a short of "dress rehearsal" which reveals to the planners and interviewer's weaknesses and provides feedback for necessary corrections.

Advantages:

- It tells about completeness, accuracy, and convenience of the sampling frame from which it is proposed to select the sample.
- It unfolds the variability within the population to be surveyed.
- It helps in bringing inadequacies of the draft questionnaire.
- It helps in bringing inadequacies of the draft questionnaire.
- It allays interviewer's fear and bests stamina and skill/efficiency
- It helps in making estimates of required time & expand.

7. Case Study

It is a method of intensively exploring and analysing the life of single social unit - (be it a person, a family, institution, culture group, even an entire community). In this method no attempt is made to exercise experimental or statistical controls and phenomena relating to the units studied in natural settings. The information are gathered from variety of sources.

This method is useful in following cases:

- To present rare, remarkable, typical evidence
- To exemplify or illustrate a concept in depth which otherwise is difficult.
- To demonstrate a technique (team building
- To establish a pool of data that may be useful at a future point in time. To serve as an inductive or hypothesis-generating vehicle

Advantages:

- Very intensive in nature
- Data Collection is flexible
- Data collected in natural setting
- Less expensive method

Limitations:

- Lacks internal validity
- Cannot serve as a base for generalization.

- Qualitative analysis subject to varying interpretations.
- More time consuming
- Causal relationships are not possible to establish.

Comparison of different social research methods is given below:

			Lab experiment	Simulation	Field experiment	Field study	Sample survey	Case study
A	Efficiency	Ease	L	L	L	L	M	H
		Speed	M	M	L	L	M	H
		Cost-initial cost of setting	M	L-H	M-H	M	H	L
		Cost- marginal cost per subject	L	L-H	M	M	L-M	L
B	Informational adequacy	Potential for controlling confounding variables and testing causal hypothesis	H	M-H	M	L	L	N
		Artifacts -potential for experimenter demand	H	M	M-H	M	L	H
		-potential for evaluation apprehension	H	M-H	M-H	M	L	H
C	Generalisability		L-H	L-H	L	L	H	L
D	Naturalness of setting		L	M-H	H	H	H	H

N: NONE, L: LOW, M: MODERATE, H: HIGH

MAX MIN CON Principle

- Maximise the true variance / the variance of a study's independent variable (in order to make the study externally Valid)
- Minimise the error variance / the variance caused by unreliability of measures
- Controlling of the extraneous variables/ confounding variables (in order to make the study internally valid)

Four ways to control extraneous variance:

- To eliminate variables as variable. This can be done by choosing subjects so that they are homogeneous as far as possible on that independent variable
- By thorough randomisation
- To build it right into the design as an independent variable thus achieving control and yielding additional research information about the effect of the variable on the dependent variable.
- To match the subject through division and randomization.

Hypothesis

"Hypo" means 'less than' and 'thesis' means a 'a generally held view'. Hypothesis is a less than generally held view. As a matter of fact, scientific investigations start with hunch.

Hypothesis is a tentative generalization of the validity of which remains to be tested. In other words, it is a testable statement of potential relationship between two or more variables.

Functions and Roles of Hypothesis:

1. Hypotheses are working instrument and can be deduced from theory or from other hypothesis.
2. Hypothesis provide directions to research by putting up scientific question.
3. It directs the search for data in order to answer the potential research question.
4. It serves as an instrument (important link between theory and investigation).
5. Helps in drawing proper conclusion.

Qualities of a workable hypothesis:

- Testability
- Simplicity
- specificity
- Clarity
- Relevancy

Hypotheses are statements about the relations between variables.

Carry clear implication for testing the stated relations

Type I Error: Committed when null-hypothesis is rejected that is true and H_1 is accepted which is false.

Type II Error: Committed when null hypothesis is accepted that is false and H_1 is rejected which is true.

Forms of hypothesis:

Declarative: It states the relationship in that form, researcher expects.

Null hypothesis: States no relation between two variables

Interrogative: In question form

SAMPLING TECHNIQUES

Prof. Souvik Ghosh
Hony. Director
Agro-Economic Research Centre
Visva Bharati
Email: dir.aerc@visva-bharati.ac.in

Probability Sampling

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wish to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes what defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality to be released to the customer, or should be sentenced for scrap or rework due to poor quality. In this case, the batch is the population. Although the population of interest often consists of physical objects, sometimes we need to sample over time, space, or some combination of these dimensions. For instance, an investigation of supermarket staffing could examine checkout line length at various times, or a study on endangered penguins might aim to understand their usage· of various hunting grounds over time. For the time dimension, the focus may be on periods or discrete occasions, in other cases, our 'population' may be even less tangible.

Probability

A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household.

For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income.

People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.) In the above example, not everybody has the same probability of selection; what makes it a

probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and
2. Involves random selection at some point.

Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'under covered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest which forms the criteria for selection. Hence, because the selection of elements is non-random, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door {e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities. On-probability sampling

Methods include accidental sampling, quota sampling and purposive sampling. In addition, nonresponse effects may turn any probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

Sampling methods

Within any of the types of frames identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

Simple random sampling

Simple random sampling: In a simple random sample (SRS) of a give size, all such subsets of the frame are given an equal probability. Furthermore, any given pair of elements has the same chance of selection as any other such pair (and similarly for triples and so on). This minimises bias and simplifies analysis of results. In particular the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple randon1 sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to over represent one sex and under represent the other. (Systematic and stratified techniques), attempt to overcome this problem by "using information about the population" to choose a more "representative" sample.

SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in "research questions specific" to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide subsamples of the population. "Stratified sampling" addresses this weakness of SRS.

Simple random sample

In statistics, a simple random sample is a subset of individuals (a sample) chosen from a larger set (a population). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset is process and technique is known as simple random individual and should not be confused with systematic random. A simple random sample is an unbiased surveying technique.

Simple random sampling is a basic type of sampling, since it can be a component of other more complex sampling methods. The principle of simple random sampling is that every object has the same probability of being chosen. For example, suppose N college students want to get a ticket for a basketball game, but there are only $X < N$ tickets for them, so they decide to have a fair way to see who gets to go. Then, everybody is given a number in the range from 0 to N-1, and random numbers are generated, either electronically or from a table of random numbers. Numbers outside the range from 0 to N-1 are ignored, as are any numbers previously selected. The first X numbers would identify the lucky ticket winners.

In small populations and often in large ones, such sampling is typically done "without replacement", i.e., one deliberately avoids choosing any member of the population more than once. Although simple random sampling can be conducted with replacement instead, this is less common and would normally be described more fully as simple random sampling with replacement. Sampling done without replacement is no longer independent, but still satisfies exchangeability, hence many results still hold. Further, for a small sample from a large

population, sampling without replacement is approximately ‘the same as sampling with replacement, since the odds of choosing the same individual twice is low.

An unbiased random selection of individuals is important so that if a large number of samples were drawn, the average sample would accurately represent the population. However, this does not guarantee that a particular sample is a perfect representation of the population. Simple random sampling merely allows one to draw externally valid conclusions about the entire population based on the sample.

Conceptually, simple random sampling is the simplest of the probability sampling techniques. It requires a complete sampling frame, which may not be available or feasible to construct for large populations. Even if a complete frame is available, more efficient approaches may be possible if other useful information is available about the units in the population.

Advantages are that it is free of classification error, and it requires minimum advance knowledge of the population other than the frame. Its simplicity also makes it relatively easy to interpret data collected in this manner. For these reasons, simple random sampling best suits situations where not much information is available about the population and data collection can be efficiently conducted on randomly distributed items, or where the cost of sampling is small enough to make efficiency less important than simplicity, if these conditions do not hold, stratified sampling or cluster sampling may be a better choice.

Distinction between a systematic random sample and a simple random sample

Consider a school with 1000 students, divided equally into boys and girls, and suppose that a researcher wants to select 100 of them for further study. All their names might be put in a bucket and then 100 names might be pulled out. Not only does each person have an equal chance of being selected, we can also easily calculate the probability of a given person being chosen, since we know the sample size (n) and the population (N):

In the case that any given person can only be selected once (i.e., after selection a person is removed from the selection pool):

$$P = 1 - \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \cdots \cdot \frac{N-n}{N-(n-1)}$$

$$\overset{\text{Canceling:}}{=} 1 - \frac{N-n}{N}$$

$$= \frac{n}{N}$$

$$= \frac{100}{1000}$$

$$= 10\%$$

2. In the case that any selected person is returned to the selection pool (i.e., can be picked more than once):

$$P = 1 - \left(1 - \frac{1}{N}\right)^n = 1 - \left(\frac{999}{1000}\right)^{100} = 0.0952\ldots \approx 9.5\%$$

This means that every student in the school has in any case approximately a 1 in 10 chance of being selected using this method. Further, all combinations of 100 students have the same probability of selection.

If a systematic pattern is introduced into random sampling, it is referred to as "systematic (random) sampling". An example would be if the students in the school had numbers attached to their names ranging from 0001 to 1000, and we chose a random starting point, e.g. 0533, and then picked every 10th name thereafter to give us our sample of 100 (starting over with 0003 after reaching 0993). In this sense, this technique is similar to cluster sampling, since the choice of the first unit will determine the remainder. This is no longer simple random sampling, because some combinations of 100 students have a larger selection probability than others - for instance, {3, 13, 23, ..., 993} has a 1/10 chance of selection, while {1, 2, 3, .. . , 100} cannot be selected under this method.

Systematic sampling

Systematic sampling relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards. In this case, k= (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 1O').

As long as the starting point is randomized, systematic sampling is a type of sampling, It is easy to implement and the stratification induced can make it efficient, if the variable by which the list is order is correlated with the variable of interest.'Every10th' sampling is especially useful for efficient sampling from databases. For example, suppose we wish to sample people from a long street that starts in a poor area (house No. 1) and ends in an expensive district (house No. 1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e.g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end; by randomly selecting the start between #1 and #10, this bias is eliminated. However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less

accurate than simple random sampling. For example, consider a street where the odd-numbered houses are all on the north (expensive) side of the road, and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible to get a representative sample; either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side.

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to quantify that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses - but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is not 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4, 14, 24, 994} has a one-in-ten probability of selection, but the set {4, 13, 24, 34,} has zero probability of selection. Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

PPS- SAMPLING

Probability proportional to size (PPS) is a sampling technique for use· with surveys or mini-surveys in which the probability of selecting a sampling unit (e.g., village, zone, district, health centre) is proportional to the size of its population. It gives a probability (i.e., random, representative) sample. It is most useful when the sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice versa. This method also facilitates planning for field work because a pre-determined number of respondents are interviewed in each unit selected, and staff can be allocated accordingly

Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected· [(1)] There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified

sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

A stratified sampling approach is most effective when three conditions are met

1. Variability within strata are minimized
2. Variability between strata are maximized
3. The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods

1. Focuses on important subpopulations and ignores irrelevant ones.
2. Allows use of different sampling techniques for different subpopulations.
3. Improves the accuracy/efficiency of estimation.
4. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

1. Requires selection of relevant stratification variables which can be difficult.
2. Is not useful when there are no homogeneous subgroups. 3. Can be expensive to implement.

Probability-proportional-to-size sampling

In some cases, the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above. Another option is probability proportional to size ('PPS') sampling, in which the selection probability for each element is set

to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under- represented due to chance variation in selections.

Systematic sampling theory can be used to create a probability proportionate to size sample. This is done by treating each count with the size variable as a single sampling unit Samples are then identified by selecting at even intervals among these counts within the size variable. This method is sometimes called PPS-sequential or monetary unit sampling in the case of audits or forensic sampling.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to 150013) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample sizes by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available-for instance, a survey attempting to measure the number of guest-nights spent. In hotels might use each hotel's number of rooms as an auxiliary variable. In some cases, an older measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates. (6)

Cluster sampling

Sometimes it is more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. Nearly all samples are in some sense 'clustered' in time - although this is rarely taken 1 to account in the analysis.) For instance, if surveying house_ ho1ds within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame; with an element-level frame created only for the selected clusters. In the example above 1 the sample only requires a block-level city map for initial selections, and then a household-level map of the 100 selected blocks, rather than a household-level map of the whole city. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster

variation. For this reason, cluster sampling requires a larger sample than SRS to achieve the same level of accuracy - but cost savings from clustering might still make this a cheaper option.

Cluster sampling is commonly implemented as multistage sampling. This is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, · for 'instance) selected at the last step of this procedure are then surveyed. This technique, thus, is essentially the process of taking random subsamples of preceding random samples.

Multistage sampling can substantially reduce sampling costs, where the complete population list would need to be constructed (before other sampling methods could be applied). By eliminating the work involved in describing clusters that are not selected, multistage sampling can reduce the large costs associated with traditional cluster sampling. (6)

Multistage sampling

Multistage sampling is a complex form of cluster sampling. Cluster sampling is a type of sampling which involves dividing the population into groups (or clusters). Then, one or more clusters are chosen at random and everyone within the chosen cluster is sampled. Using all the san1ple elen1ents in all the selected clusters may be prohibitively expensive or unnecessary. Under these circumstances, multistage cluster sampling becomes useful. Instead of using all the elements contained in the selected clusters, the researcher randomly selects elen1ents fron1each cluster. Constructing the clusters is the first stage. Deciding what elements within the cluster to use is the second stage. The technique is used frequently when a complete list of all members of the population does not exist and is inappropriate.

In some cases, several levels of cluster selection may be applied before the final sample elements are reached. For example, household surveys conducted by the Australian Bureau of Statistics begin by dividing metropolitan regions into ‘collection districts’ and selecting some of these collection districts (first stage). The selected collection districts are then divided into blocks, and blocks are chosen from within each selected collection district (second stage). Next, dwellings are listed within each selected block, and some of these dwellings are selected (third stage). This method makes it unnecessary to create a list of every dwelling in the region and necessary only for selected blocks. In re1note areas, an additional stage of clustering is used, in order to reduce travel requirements.

Although cluster sampling and stratified sampling bear some superficial sin1ilarities, they are substantially different. In stratified sampling, a random sample is drawn from all the strata, where in cluster sampling only the selected clusters are studied, either in single- or multi-stage.

Advantages

1. Cost and speed that the survey can be done in
2. Convenience of finding the survey sample
3. Normally more accurate than cluster sampling for the same size sample

Disadvantages

1. Not as accurate as SRS if the sample is the san1e size
2. More testing is difficult to do.

Steps for using sample size tables

I. Postulate the effect size of interest, α and β
II. Check sample size table
III. Select the table corresponding to the selected α
IV. Locate the row corresponding to the desired power
V. Locate the column corresponding to the estimated effect size.
VI. The intersection of the column and row is the minimum sample size required

Sampling errors and biases

Sampling errors and biases are induced by the sample design. They include:

- Selection bias: When the true selection probabilities differ from those assumed in calculating the results.
- Random sampling error: Random variation in the results due to the elen1ents in the sample being selected at random.

Non-Probability Sampling

Sampling is the use of a subset of the population to represent the whole population. In any form of research, true sample is always difficult to achieve. Most researchers are bounded by time, money and workforce and because of these limitations, it is almost impossible to randomly sample the entire population and it is often necessary to employ another sampling technique, the non-probability sampling. In contrast with probability sampling, non-probability sample is not a product of a randomized selection processes.

Non-probability sampling represents a group of sampling techniques that help researchers to select units from a population that they are interested in studying. Collectively, these units form the sample that the researcher studies. A core characteristic of non-probability sampling techniques is that samples are selected based on the subjective judgments of the researcher, rather than random selection (i.e., probabilistic methods), which is the cornerstone of probability sampling techniques. Non-probability sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected. The downside of this method is that an unknown proportion of the entire population was not sampled. This entails that the sample may or may not represent the

entire population accurately. Therefore, the results of the research cannot be ·used in generalization pertaining to the entire population.

Though some researchers may view non-probability sampling techniques as inferior to probability sampling techniques, there are strong theoretical and practical reasons for their use.

When to Use Non-Probability Sampling?

- This type of sampling can be used when demonstrating that a particular trait exists in the population.
- It can also be used when the researcher aims to do a qualitative, pilot or exploratory study.
- It can be used when randomization is impossible like when the population is almost limitless.
- It can be used when the research does not aim to generate results that will be used to create generalizations pertaining to the entire population.
- It is also useful when the researcher has limited budget, time and workforce.
- This technique can also be used in an initial study which will be carried out again using a randomized, probability sampling.

Further, this note is divided into two sections:

- ✓ Principles of non-probability sampling, &
- ✓ Types of non-probability sampling

Principles of Non-Probability Sampling:

There are theoretical and practical reasons for using non-probability sampling. In addition, we need to decide whether non-probability sampling is appropriate based on the research strategy that has been chosen to guide a particular research project.

Theoretical Reasons:

Non-probability sampling represents a valuable group of sampling techniques that can be used in research that follows qualitative, mixed methods, and even quantitative research designs. Despite this, for researchers following a quantitative research design, non-probability sampling techniques can often be viewed as an inferior alternative to probability sampling techniques. Non-probability sampling techniques can often be viewed in such a way because units are not selected for inclusion in a sample based on random selection, unlike probability sampling techniques. As a result, researchers following a quantitative research design often feel that they are forced to use non-probability sampling techniques because of some inability to use probability sampling (e.g., the lack of access to a list of the population being studied). However, this is not the case for researchers following a qualitative research design.

When following a qualitative research design, non-probability sampling techniques, such as purposive sampling, can provide researchers with strong theoretical reasons for their choice of

units (or cases) to be included in their sample. Rather than using probabilistic methods (i.e., random selection) to generate a sample, non-probability sampling requires researchers to use their subjective judgments, drawing on theory (i.e., the academic literature) and practice (i.e., the experience of the researcher and the evolutionary nature of the research process). Unlike probability sampling, the goal is not to achieve objectivity in the selection of samples, or necessarily attempt to make generalizations (i.e., statistical inferences) from the sample being studied to the wider population of interest. Instead, researchers following a qualitative research design tend to be interested in the intricacies of the sample being studied. Whilst making generalizations from the sample to the population under study may be desirable, it is more often a secondary consideration. Even whether this is desired, there are additional problems of bias and transferability (or validity).

Practical Reasons:

Non-probability sampling is often used because the procedures used to select units for inclusion in a sample are much easier, quicker and cheaper when compared with probability sampling. This is especially the case for convenience sampling. For students doing dissertations at the undergraduate and master's level, such practicalities often lead to the use of non-probability sampling techniques.

As mentioned, for researchers following a quantitative research design, non- probability sampling techniques can often be viewed as an inferior alternative to probability sampling techniques. However, where it is not possible to use probability sampling, non-probability sampling at least provides a viable alternative that can be used. As such, it ensures that research following a quantitative research design is not simply abandoned because (a) it cannot meet the criteria of probability sampling and/ or (b) meeting such criteria is excessively costly or time consuming, such that it would not be sponsored. This could significantly diminish the potential for researchers to study certain types of population, such as those populations that are hidden or hard-to-reach (e.g., drug addicts, prostitutes), where a list of the population simply does not exist. Here, snowball sampling, a type of non-probability sampling technique, provides a solution.

Non-probability sampling can also be particularly useful in exploratory research where the aim is to find out if a problem or issue even exists in a quick and inexpensive way. After all, you may have a theory that such a problem or issue exists, but there is limited or no research that currently supports such a theory. Where your main desire is to find out is if such a problem or issue even exists, the potential sampling bias of certain non-- probability sampling techniques can .be used as a tool to help you. For example, you may choose to select only those units to be included in your sample that you feel will exhibit the problem or issue you are interested in finding. If this problem or issue does not exist even in your biased sample, it is unlikely to be present if you selected a relatively unbiased sample (whether using another non-probability sampling technique; or even a probability sampling technique). This would help you to avoid a potentially more time consuming and expensive piece of research looking into a potential problem or issue that actually doesn't exist. It may also be considered an ethical approach to finding out whether a problem or issue is worth examining in more depth, since fewer participants are subjected to a research project unnecessarily.

Types of Non-Probability Sampling

When considering using non-probability sampling, it is important to consider how the choice of research strategy will influence whether this is an appropriate decision. Even if non-probability sampling fits with the research strategy, it is important to choose the appropriate type of non-- probability sampling techniques. There are five types of non-probability sampling technique that are used by researchers: quota sampling, convenience sampling, purposive sampling, self-selection sampling and snowball sampling.

Quota Sampling

Quota sampling is a non-probability sampling technique wherein the assembled sample has the same proportions of individuals as the entire population with respect to known characteristics, traits or focused phenomenon.

In addition to this, the researcher must make sure that the composition of the final sample to be used in the study meets the research's quota criteria.

Step-by-step Quota Sampling:

- ✓ The first step in non-probability quota sampling is to divide the population into exclusive sub-groups.
- ✓ Then, the researcher must identify the proportions of these subgroups in the population; this same proportion will be applied in the sampling process.
- ✓ Finally, the researcher selects subjects from the various subgroups while taking into consideration the proportions noted in the previous step.
- ✓ The final step ensures that the sample is representative of the entire population. It also allows the researcher to study traits and characteristics that are noted for each subgroup.

Example of Quota Samples:

In a study wherein the researcher likes to compare the academic performance of the different high school class levels, its relationship with gender and socioeconomic status·, the researcher first identifies the subgroups.

Usually, the subgroups are the characteristics or variables of the study. The researcher divides the entire population into class levels, intersected with gender and socioeconomic status. Then, he takes note of the proportions of these subgroups in the entire population and then samples each subgroup accordingly.

When to Use Quota Samples:

The main reason why researchers choose quota samples is that it allows the researchers to sample a subgroup that is of great interest to the study. If a study aims to investigate a trait or a characteristic of a certain subgroup, this type of sampling is the ideal technique. Quota sampling also allows the researchers to observe relationships between subgroups. In some

studies, traits of a certain subgroup interact with other traits of another subgroup. In such cases, it is also necessary for the researcher to use this type of sampling technique.

Disadvantages of Quota Samples:

It may appear that this type of sampling technique is totally representative of the population. In some cases, it is not. Keep in mind that only the selected traits of the population were taken into account in forming the subgroups.

In the process of sampling these subgroups, other traits in the sample may be overrepresented. In a study that considers gender, socioeconomic status and religion as the basis of the subgroups, the final sample may have skewed representation of age, race, educational attainment, marital status and a lot more.

Convenience Sampling

Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher. ·

The subjects are selected just because they are easiest to recruit for the study and the researcher did not consider selecting subjects that are representative of the entire population.

In all forms of research, it would be ideal to test the entire population, but in most cases, the population is just too large that it is impossible to include every individual. This is the reason why most researchers rely on sampling techniques like convenience sampling, the most common of all sampling techniques. Many researchers prefer this sampling technique because it is fast, inexpensive, easy and the subjects are readily available.

Examples of Convenience Sampling:

One of the most common examples of convenience sampling is using student volunteers as subjects for the research. Another example is using subjects· that are selected from a clinic, a class or an institution that is easily accessible to the researcher. A more concrete example is choosing five people from a class or choosing the first five names from the list of patients.

In these examples, the researcher inadvertently excludes a great proportion of the population. A convenience sample is either a collection of subjects that are accessible or a self-selection of individuals willing to participate which is exemplified by your volunteers.

When to Use Convenience Sampling:

Researchers use convenience sampling not just because it is easy to use, but because it also has other research advantages.

In pilot studies, convenience sample is usually used because it allows the researcher to obtain basic data and trends regarding his study without the complications of using a randomized sample. This sampling technique is also useful in documenting that a particular quality of a substance or phenomenon occurs within a given sample. Such studies are also very useful for detecting relationships among different phenomena.

Criticisms of Convenience Sampling:

The most obvious criticism about convenience sampling is sampling bias and that the sample is not representative of the entire population. This· may be the biggest disadvantage when using a convenience sample because it leads to more problems and criticisms.

Systematic bias stems from sampling bias. This refers to a constant difference between the results from the sample and the theoretical results from the entire population. It is not rare that the results from a study that uses a convenience sample differ significantly with the results from the entire population. A consequence of having systematic bias is obtaining skewed results.

Another significant criticism about using a convenience sample is the limitation in generalization and inference making about the entire population. Since the sample is not representative of the population, the results of the study cannot speak for the entire population. This results to a low external validity of the study.

Notes on Convenience Sampling:

When using convenience sampling, it is necessary to describe how the sample would differ from an ideal sample that was randomly selected. It is also necessary to describe the individuals who might be left out during the selection process or the individuals who are overrepresented in the sample.

In connection to this, it is better to describe the possible effects of the people who were left out or the subjects that are over-represented to the results. This will allow the readers of the research to get a good grasp of the sample that was being tested. It will also enable the readers to estimate the possible difference between the results of convenience sampling and the results from the entire population.

Judgmental Sampling or Purposive Sampling

Judgmental sampling, also known as Purposive sampling, Selective or Subjective and Authoritative sampling, is a non-probability sampling technique where the researcher selects units to be sampled based on their knowledge and professional judgment.

Purposive sampling is used in cases where the specialty of an authority can selects a more representative sample that can bring more accurate results than by using other probability sampling techniques. The process involves nothing but purposely handpicking individuals from the population based on the authority's or the researcher's knowledge and judgment.

Example of Judgmental Sampling:

In a study wherein a researcher wants to know what it takes to graduate summa cum laude in college, the only people who can give the researcher first hand advise are the individuals who graduated summa cum laude. With this very specific and very limited pool of individuals that can be considered as a subject, the researcher must use judgmental sampling.

When to Use Judgmental Sampling:

Judgmental sampling design is usually used when a limited number of individuals possess the trait of interest. It is the only viable sampling technique in obtaining information from a very specific group of people. It is also possible to use judgmental sampling if the researcher knows a reliable professional or authority that he thinks is capable of assembling a representative sample.

Setbacks of Judgmental Sampling:

The two main weaknesses of authoritative sampling are with the authority and 1.n the sampling process; both of which pertain to the reliability and the bias that accompanies the sampling technique.

Unfortunately, there is usually no way to evaluate the reliability of the expert or the authority. The best way to avoid sampling error brought by the expert is to choose the most experienced authority in the field of interest.

When it comes to the sampling process, it is usually biased since no randomization was used in obtaining the sample. It is also worth noting that the members of the population did not have equal chances of being selected. The consequence of this is the misrepresentation of the entire population which will then limit generalizations of the results of the study.

Sequential Sampling Method

Sequential sampling is a non-probability sampling technique wherein the researcher picks a single or a group of subjects in a given time interval, conducts his study, analyzes the results then picks another group of subjects if needed and so on.

This sampling technique gives the researcher limitless chances of fine tuning his research methods and gaining a vital insight into the study that he is currently pursuing.

Difference of Sequential Sampling from All Other Sampling Techniques:

If we are to consider all the other sampling techniques in research, we will all come to a conclusion that the experiment and the data analysis will either boil down to accepting the null hypothesis or disproving the null hypothesis while accepting the alternative hypothesis.

In sequential sampling technique, there exists another step, a third option. The researcher can accept the null hypothesis, accept his alternative hypothesis, or select another pool of subjects and conduct the experiment once again. This entails that the researcher can obtain limitless number of subjects before finally making a decision whether to accept his null or alternative hypothesis.

Advantages of Sequential Sampling:

The researcher has a limitless option when it comes to sample size and sampling schedule. The sample size can be relatively small of excessively large depending on the decision making of the researcher. Sampling schedule is also complete dependent to the researcher

since a second group of samples can only be obtained after conducting the experiment to the initial group of samples.

As mentioned above, this sampling technique enables the researcher to fine- tune his research methods and results analysis. Due to the repetitive nature of this sampling method, minor changes and adjustments can be done during the initial parts of the study to correct and hone the research method.

There is very little effort in the part of the researcher when performing this sampling technique. It is not expensive, not time consuming and not workforce extensive.

Disadvantages of Sequential Sampling:

This sampling method is hardly representative of the entire population. Its only hope of approaching representativeness is when the researcher chose to use a very large sample size significant enough to represent a big fraction of the entire population.

The sampling technique is also hardly randomized. This contributes to the very little degree representativeness of the sampling technique.

Due to the aforementioned disadvantages, results from this sampling technique cannot be used to create conclusions and interpretations pertaining to the entire population.

Snowball Sampling

Snowball sampling is a non-probability sampling technique that is used by researchers to identify potential subjects in studies where subjects are hard to locate.

Researchers use this sampling method if the sample for the study is very rare or is limited to a very small subgroup of the population. This type of sampling technique works like chain referral. After observing the initial subject, the researcher asks for assistance from the subject to help identify people with a similar trait of interest**.**

The process of snowball sampling is much like asking your subjects to nominate another person with the same trait as your next subject. The researcher then observes the nominated subjects and continues in the same way until the obtaining sufficient number of subjects.

For example, if obtaining subjects for a study that wants to observe a rare disease, the researcher may opt to use snowball sampling since it will be difficult to obtain subjects. It is also possible that the patients with the same disease have a support group; being able to observe one ·of the members as your initial subject will then lead you to more subjects for the study.

Advantages or Snowball Sampling:

- The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods.
- The process is cheap, simple and cost-efficient.
- This sampling technique needs little planning and fewer workforces compared to other sampling techniques.

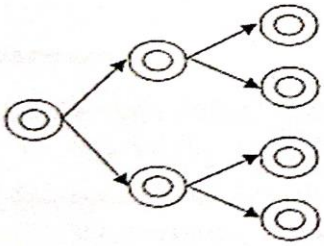
Disadvantages or Snowball Sampling:

- The researcher has little control over the sampling method. The subjects that the researcher can obtain rely mainly on the previous subjects that were observed.
- Representativeness of the sample is not guaranteed. The researcher has no idea of the true distribution of the population and of the sample.
- Sampling bias is also a fear of researchers when using this sampling technique. Initial subjects tend to nominate people that they know well. Because of this, it is highly possible that the subjects share the same traits and characteristics, thus, it is possible that he samples that the researcher will obtain is only a small subgroup of the entire population.

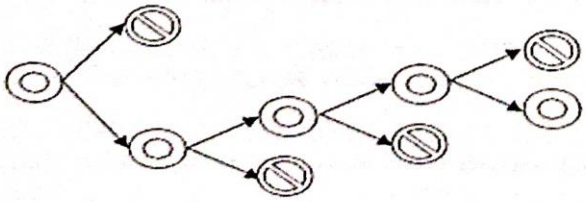
Types of Snowball Sampling:

i. Linear Snowball Sampling

ii. Exponential Non-Discriminative Snowball Sampling



iii. Exponential Discriminative Snowball Sampling



DATA COLLECTION METHODS

Prof. Souvik Ghosh
Hony. Director
Agro-Economic Research Centre
Visva Bharati
Email: dir.aerc@visva-bharati.ac.in

Data Collection Techniques

The data collection is undertaken through various methods like questionnaire, interview-schedule, observation, projective techniques and sociometry.

Questionnaire and Schedule Survey

A Questionnaire is a form containing a series of questions and providing space for their replies to be filled in by the respondent himself. Schedule is a set of questions which are asked and filled in by an interviewer in a face-to-face situation with another person (respondent.).

Questionnaire is filled in by a respondent without any direct oral explanation or interpretation from the investigator. A schedule is filled in by the investigator himself who can, if necessary, explain any point to the respondent on the spot.

Sl. No.	Schedule	Questionnaire
1	Filled in by the investigator, never mailed to respondent. Presence of investigator/ interviewer enlivens the atmosphere.	Filled in by the respondents, mailed to respondents. Lacks personal touch.
2	It is generally used where the survey is to be conducted of a small geographic area.	It is generally used where the field o enquiry is large. It can be mailed to distant places.
3	Involves greater expenditure of time and money	Much time and money are saved in its use
4	It can be used even where the respondents are illiterate as the investigator can explain to them any point on the spot	It cannot be used where the respondents are illiterate. Possibility of having several entries incomplete
5	The wordings may not be in term of questions	Wordings are in the form of questions
6	In its designing convenience of investigator in handling it the field should be the main consideration	Convenience of the respondents should be the main considerations

Types of Questionnaire/ Schedule

Structured- Questions are fixed.

Unstructured- only broad areas flexibility is advantage here. However, replies cannot be easily compared, quantitative analyses is difficult.

Guideline for designing a questionnaire / schedule

1. The questionnaire must be intimately related to the objectives of the investigation.
2. It should be brief and clear and simple.
3. Questions outside the respondent's experience should not be asked.
4. In asking questions about past events too much reliance should not be placed the respondent's memory.
5 Questions which are likely to arouse bias in the respondent should be avoided.
6. The structure of the question should be according to the forum in which the responses are to be recorded.
7. There should be logical sequence of questions in the questionnaire.
R. It is desirable to incorporate the cross checks.
9. A mailed questionnaire should always be accompanied with a letter.
10. Each questionnaire should have an appropriate ending to it.

Pretesting a Questionnaire

1. It reveals to the investigator in advance questions which either is not understood or is misunderstood by the respondents or which arouse defensiveness in them.

2. It helps to decide the proper form and structure of each question - whether it should be dichotomous, multiple choice, open-end or some other combination.

3. It helps to resolve many mechanical problems of measurement.

4. It helps to improve questionnaire design in terms of format, quality of instructions, need for filter on screening questions, amount of spacing required on the page, and the use of special symbols for colour-coded pages for directing the interviewer physically through the questionnaire

5. It gives firm estimates of the amount of time, money, personnel, and equipment required to process the main study data efficiently and successfully.

Nonresponse due to:

- Non-coverage
- Not-at-homes
- Unable to answer
- The hardcore-refuse to be interviewed

Methods of Interviews:

- Invasion method.

- Immersion method

Types of Interviews

- Standardised structural formal interview
- Unstandardised, Unstructured or informal interview
- Semi standardised, semi structured interview.

Interview process

- Preparatory thinking.
- Developing rapport with the respondent.
- Carry the interview forward.
- Rewarding the interview.
- Closing the interview.

Factors affecting the interview: Inherent to the interviewer, to the interviewee, general situation/environment, content of the interview.

Observation method

It is a very important technique of data collection in use in experimental and non-experimental, social and anthropological research. In the strict sense it implies the use of the eyes rather than of the ears. The investigator obtains data by watching and noticing the phenomena as they occur with regard to their cause and effect or mutual relations. An eg. of data collection by observation is the investigation of the expenditure and living conditions of landless labour in rural areas by actually living/staying in that area.

Observation can be of three types:

Simple or uncontrolled observation: The data collected by two observers cannot be compared. The observer's bias is the crucial weakness.

Systematic or controlled observation: it tries to remove the observer's bias by using various control techniques. Best known e.g. is the interaction-analysis techniques-observer is required to place his observations in pre-determined categories.

Mass observation: Record mass/collective behaviour of people in public places.

Uncontrolled observations are of 3 types

- Participant observation
- Non-participant observation.
- Quasi participant observation (the observer assumes several roles, sometimes a participant, sometimes an interviewer). (Non-participant and Quasi-participant observation can be either Direct or Indirect observation)

Rules/guidelines for observation

1. The observer should first of all formulate his hypothesis. This would provide desirable focus to the study and guide him the facts to be observed.

2. The observer should try to probe further exploring the interrelationships between the facts / observed phenomena.

3. The observer should keep record of all occurrences in diary, may be supplemented with photographs.

4. The record of observations should be kept separate from the record of interpretations.

5. A continuous analysis of record should be made.

6. Any biasness should be avoided.

7. Uncontrolled observations should as far as possible be supplemented by some other tools such as questionnaires, tests, quantitative analyses.

Projective techniques

- Rorschach ink blot tests.
- Sentence completion tests
- Word association test
- Thematic apperception test (TAT)
- Role playing/ psychodrama (respondent plays himself) / sociodramas (he/she acts out roles of others)
- Error choice/ information test

In these techniques, the true nature of the subject matter is concealed and the respondent is not aware of it and purpose of the study.

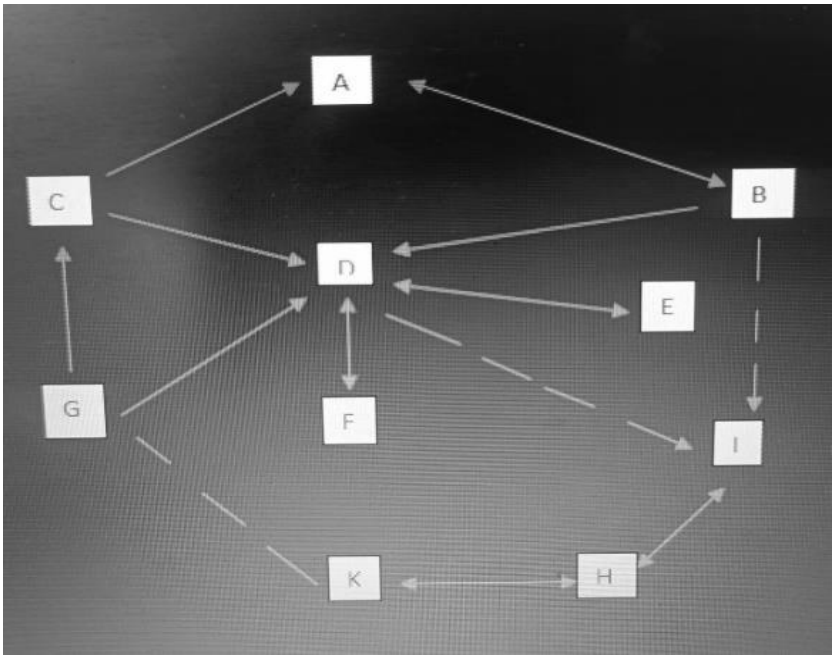
Sociometry

Interpersonal relationship in terms of attraction or repulsion. It is a method used for the discovery and manipulation of social configurations by measuring the attractions and repulsions between individuals in a group at some point in time. The basic technique is the socio-metric test. It explores thee each individual positions in the social structure.

Sociogram

The diagrammatic presentation of socio metric data is the sociogram. Sociogram is a diagrammatic/diagnostic presentation/ visual presentation of different patterns of choices and rejection noted in the responses made by the individual members in the groups.

Developed by Guttman.



CLIQUE: Subset of persons interact intensively (ABCD)
STAR: Person consulted by maximum number of people (D)
ARISTOTLE LEADER: Person whom the 'star' contact (EF)
ISOLATE: People do not prefer, least consulted (G)

Data Collection Tools

Data collection procedures involve the systematic gathering of information or data from individuals, groups, or sources for research or analysis purposes. Various methods are employed depending on the nature of the study, research objectives, and available resources. Some common data collection procedures include personal interview schedules, questionnaires, and online data collection techniques.

1. Personal Interview Schedule:

A personal interview schedule involves face-to-face interaction between the interviewer and the respondent. The interviewer asks questions orally and records the respondent's answers.

Types of Interview Schedules:

Structured Interviews: In structured interviews, the interviewer follows a predetermined set of questions in a fixed order. This format ensures consistency across interviews and facilitates quantitative analysis of responses. Structured interviews are often used in survey research and quantitative studies.

Semi-Structured Interviews: Semi-structured interviews combine predefined questions with the flexibility to explore topics in more depth. While there is a general outline, interviewers have the freedom to probe or follow up on responses to gather richer qualitative data. Semi-

structured interviews are commonly used in qualitative research to explore complex phenomena and gain insights into participants' perspectives.

Unstructured Interviews: Unstructured interviews are open-ended and flexible, allowing interviewers to explore topics freely without a predetermined set of questions. The conversation flows naturally, guided by the interviewer's interests and the participant's responses. Unstructured interviews are valuable for generating rich qualitative data but may be more challenging to analyse due to their open-ended nature.

Clinical Interviews: Clinical interviews are used in psychology, counselling, and clinical research to assess individuals' psychological or emotional functioning. They involve in-depth exploration of personal experiences, emotions, and behaviours to diagnose mental health conditions, develop treatment plans, or conduct therapeutic interventions.

Advantages:

- **In-depth Information:** Allows for probing and clarification of responses, facilitating in-depth understanding.
- **Flexibility:** Interviewers can adapt questions or follow-up based on the respondent's answers, allowing for flexibility in data collection.
- **Non-verbal Cues:** Enables observation of non-verbal cues, which can provide additional insights into responses.
- **High Response Rates:** Personal interaction may lead to higher response rates compared to other methods.

Disadvantages:

- **Time-Consuming:** Can be time-consuming and resource-intensive, especially for large sample sizes.
- **Costly:** Requires trained interviewers and resources for travel, increasing costs.
- **Interviewer Bias:** Interviewer bias may influence responses, affecting data quality.
- **Social Desirability Bias:** Respondents may provide socially desirable responses, impacting data validity.

2. Questionnaire:

A questionnaire is a written or printed set of questions designed to elicit specific information from respondents. It can be administered in person, via mail, or electronically.

Types of Questionnaires:

Structured Questionnaires: Structured questionnaires consist of closed-ended questions with predefined response options, such as multiple-choice, Likert scale, or rating scale questions. Respondents select from the provided choices, making data collection and analysis more

straightforward. Structured questionnaires are commonly used in surveys and quantitative research.

Semi-Structured Questionnaires: Semi-structured questionnaires combine closed-ended questions with open-ended questions that allow respondents to provide additional information or insights. This format provides a balance between structured and unstructured approaches, enabling researchers to collect both quantitative and qualitative data.

Open-Ended Questionnaires: Open-ended questionnaires consist entirely of open-ended questions, where respondents provide free-text responses without predefined options. This format allows for in-depth exploration of topics and provides rich qualitative data but may be more challenging to analyse due to the variability in responses.

Likert Scale Questionnaires: Likert scale questionnaires use a series of statements or items to assess respondents' attitudes, opinions, or perceptions. Respondents indicate their level of agreement or disagreement with each statement using a predefined scale (e.g., strongly agree, agree, neutral, disagree, strongly disagree). Likert scale questionnaires are widely used in social sciences and market research.

Diagnostic Questionnaires: Diagnostic questionnaires are used in clinical settings to assess symptoms, behaviours, or characteristics associated with specific conditions or disorders. These questionnaires typically include a series of standardized questions or items designed to screen for or diagnose mental health or medical conditions.

Exit Questionnaires: Exit questionnaires are administered to individuals upon completion of a program, service, or event to gather feedback and assess their experiences. These questionnaires help organizations evaluate program effectiveness, identify areas for improvement, and make informed decisions about future initiatives.

Advantages:

- **Cost-Effective:** Economical method, especially for large-scale data collection.
- **Anonymity:** Allows respondents to provide honest responses without fear of judgment or bias.
- **Standardization:** Ensures consistency in data collection and analysis, facilitating comparisons.
- **Efficiency:** Can be administered to multiple respondents simultaneously, saving time.

Disadvantages:

- **Limited Depth:** Provides limited opportunity for clarification or probing, potentially resulting in shallow responses.
- **Low Response Rates:** Relies on self-administration, leading to lower response rates compared to interviews.

- **Incomplete Responses:** Respondents may skip questions or provide incomplete answers, affecting data quality.
- **Difficulty in Understanding:** Complex or ambiguous questions may lead to misunderstandings or misinterpretations.

3. Online Data Collection Techniques:

Online data collection techniques involve using the internet and digital platforms to collect data from respondents. This can include online surveys, web-based questionnaires, or data mining from social media platforms.

Online data collection refers to the process of gathering data from respondents using digital platforms and internet-based tools. Here is an overview of the process:

Setup:

- Selection of Platform: Choose an online survey platform or data collection tool that meets your research needs and objectives.
- Questionnaire Design: Design a questionnaire or survey instrument tailored to your research questions and objectives using the platform's built-in features and templates.
- Distribution Channels: Determine the channels for distributing the survey, such as email, social media, or website embeds.

Data Collection:

- Distribution: Share the survey link or questionnaire with targeted respondents via selected channels, ensuring accessibility and ease of participation.
- Response Monitoring: Monitor responses in real-time as they are collected, tracking response rates and engagement metrics.
- Reminders: Send out reminders to non-respondents to encourage participation and maximize response rates.

Data Management:

- Data Storage: Collect and store survey responses securely on the online platform, ensuring compliance with data protection regulations.
- Data Export: Export survey data in various formats (e.g., Excel, CSV) for further analysis and reporting.
- Data Cleaning: Clean and pre-process the collected data to identify and correct any errors or inconsistencies.

Online Data Collection Tools

1. Google Forms:

Google Forms is a free online survey tool provided by Google. It allows users to create customized surveys and questionnaires with various question types, including multiple-choice, short answer, and Likert scale questions. Responses are automatically collected and stored in a Google Sheets spread sheet.

Procedure:

- Create a Form: Sign in to your Google account, go to Google Forms, and click on the "+ Blank" button to create a new form.
- Add Questions: Use the form editor to add questions to your form. Choose from various question types and customize options as needed.
- Share the Form: Once your form is ready, you can share it with respondents via email, social media, or by generating a link. You can also embed the form on a website.
- Collect Responses: Responses are automatically collected in a Google Sheets spread sheet linked to the form. You can view and analyze responses directly in Google Sheets or export data to other formats for analysis.

2. SurveyMonkey:

SurveyMonkey is a popular online survey platform that offers a range of features for creating and administering surveys. It provides customizable survey templates, advanced question branching, and robust analytics tools for analyzing survey data.

Procedure:

- Sign Up/Login: Create a SurveyMonkey account or log in to your existing account.
- Create a Survey: Click on the "Create Survey" button and choose a survey template or start from scratch.
- Design the Survey: Use the survey editor to add questions, customize question types, and configure survey settings.
- Distribute the Survey: Share the survey with respondents via email, social media, or embed it on a website. SurveyMonkey also offers options for purchasing responses from a targeted audience.
- Analyze Responses: Survey responses are collected and stored securely in SurveyMonkey. Use the built-in analytics tools to analyze survey data, generate reports, and gain insights into respondent trends and preferences.

3. Qualtrics:

Qualtrics is a comprehensive online survey platform used by businesses, academic institutions, and organizations for data collection, analysis, and insights. It offers advanced survey design features, robust data analytics capabilities, and options for collaboration and integration with other tools.

Procedure:

- Login/Sign Up: Log in to your Qualtrics account or sign up for a new account.
- Create a Survey: Click on "Create Survey" to start building your survey. Qualtrics offers a range of survey templates and question types to choose from.
- Customize Survey: Use the survey editor to customize survey questions, design, and layout according to your research objectives and branding requirements.
- Distribute Survey: Distribute the survey to respondents via email, social media, or by generating a link. Qualtrics also offers options for targeting specific audiences and managing survey distribution.
- Analyze Results: Qualtrics provides powerful analytics tools for analyzing survey data, generating reports, and visualizing insights. Use advanced features such as statistical analysis, text analysis, and data segmentation to gain deeper insights into respondent behavior and preferences.

4. Typeform:

Typeform is an online survey and form-building tool known for its modern and user-friendly interface. It offers a wide range of question types, including conversational forms, multiple-choice questions, and rating scales. Typeform also provides options for customization, branding, and integration with other tools and platforms.

Procedure:

- Create a Form: Log in to your Typeform account and click on "Create" to start building your form. Choose from various form templates or create a new form from scratch.
- Design the Form: Use the drag-and-drop form builder to add questions, customize question types, and design the layout and branding of your form.
- Share the Form: Share the form with respondents by generating a unique link, embedding it on a website, or integrating it with other platforms such as email marketing tools or social media.
- Collect Responses: Responses are collected and stored securely in Typeform. Use the analytics dashboard to view and analyze responses in real-time, track completion rates, and gain insights into respondent behaviour.

- Integrate with Other Tools: Typeform offers integration with a range of third-party tools and platforms, allowing you to automate workflows, sync data, and streamline data collection and analysis processes.

Advantages of online data collection:

- **Convenience:** Provides convenience for respondents, who can participate from anywhere with internet access.
- **Cost-Effective:** Eliminates the need for printing, postage, and manual data entry, reducing costs.
- **Automation:** Allows for automation of data collection, storage, and analysis, saving time and effort.
- **Global Reach:** Enables access to a wider and more diverse pool of respondents, enhancing generalizability.

Disadvantages of online data collection:

- **Sampling Bias:** May result in sampling bias, as not all individuals have access to the internet or specific online platforms.
- **Security Concerns:** Data security and privacy concerns may arise, particularly with sensitive information.
- **Response Bias:** Respondents may provide biased or inaccurate responses, influenced by the online environment.
- **Technical Issues:** Technical glitches or compatibility issues may affect data collection or respondent experience.

The Process of Collecting Data

Data collection is a crucial step in the research process, and it involves several stages, including the selection, training, supervision, and evaluation of field investigators.

Selection of Field Investigators:

- Criteria: Select field investigators based on criteria such as experience, qualifications, language proficiency, and familiarity with the research context.
- Recruitment: Advertise job openings, conduct interviews, and assess candidates' suitability for the role.
- Training: Provide training on data collection protocols, research objectives, ethical considerations, and data management procedures.

Training of Field Investigators:

- Orientation: Orient field investigators to the research project, including its goals, objectives, and expected outcomes.

- Protocol Training: Train investigators on data collection instruments, techniques, and procedures to ensure consistency and accuracy.
- Role-Play: Conduct role-playing exercises to simulate data collection scenarios and practice effective communication and interaction with respondents.

Supervision of Field Investigators:

- Monitoring: Supervise field investigators closely to ensure adherence to data collection protocols and ethical standards.
- Support: Provide on-going support, guidance, and feedback to address any challenges or issues encountered during data collection.
- Quality Assurance: Implement quality assurance measures, such as spot checks, to verify the accuracy and reliability of collected data.

Evaluation of Field Investigators:

- Performance Evaluation: Evaluate field investigators' performance based on criteria such as data quality, timeliness, and adherence to protocols.
- Feedback: Provide constructive feedback and coaching to help improve performance and address any areas for improvement.
- Recognition: Recognize and reward outstanding performance to motivate field investigators and foster a culture of excellence.

Errors and Biases during Data Collection:

Despite careful planning and execution, data collection is susceptible to errors and biases that can affect the validity and reliability of the collected data. Here are some common errors and biases to be aware of:

Sampling Bias:

Definition: Sampling bias occurs when the sample selected for data collection is not representative of the target population, leading to skewed or un-generalizable results.

Mitigation: Use random sampling techniques, stratification, or oversampling to minimize sampling bias and ensure representativeness.

Response Bias:

Definition: Response bias occurs when respondents provide inaccurate or biased responses due to social desirability, acquiescence, or other factors.

Mitigation: Design surveys with clear, unbiased questions, avoid leading or loaded language, and use anonymous response formats to minimize response bias.

Measurement Error:

Definition: Measurement error occurs when there are inaccuracies or inconsistencies in the measurement process, leading to unreliable or invalid data.

Mitigation: Implement standardized data collection protocols, provide adequate training and supervision to field investigators, and conduct pilot testing to identify and address potential sources of measurement error.

Non-Response Bias:

Definition: Non-response bias occurs when certain groups of respondents are disproportionately less likely to participate in data collection, leading to biased results.

Mitigation: Implement strategies to minimize non-response bias, such as follow-up reminders, incentives for participation, and targeted outreach efforts to underrepresented groups.

Self-Selection Bias:

Definition: Self-selection bias occurs when individuals voluntarily choose to participate in data collection, leading to a non-random sample that may not be representative of the target population.

Mitigation: Implement random sampling techniques or use appropriate statistical adjustments to account for self-selection bias and improve the generalizability of results.

ERRORS AND BIASES

Prof. Souvik Ghosh
Hony. Director
Agro-Economic Research Centre
Visva Bharati
Email: dir.aerc@visva-bharati.ac.in

Errors

- Any difference between the average values that were obtained through a study and the true average values of the phenomenon of a particular population being target.
- Describes how much the results of a study are short of, or exceeds, the real values of the attribute of the population, by comprehending all the flaws in a research study.
- For example, if the study reveals that the adoption of good practices in dairy farming among small dairy farmers is 25% when the actual level is only 20%, the difference could be from a whole range of different biases and errors but the total level of error in this study is considered to be 5%.

Biases

- Bias refers only to error that is systematic in nature, i.e., data's value systematically differs from the true value of the population of interest
- Introduced at various stages of survey – designing, executing, data entry and analysis, and created errors.

Sampling Error

- Sampling error comprises the differences between the sample and the population that are due solely to the particular units that happen to have been selected.
- Occurs when the sample is too small to adequately infer survey results.
- Sources – by chance and through sampling bias.

Sampling Bias

- Tendency to favour the selection of units that have particular characteristics.
- Cause – poor sampling plan.

Non-sampling or Measurement Error

- Error that results solely from the manner in which the observations are made.
- The non-sampling error can be measurement and non-response errors.

(a) Measurement error

The result of poor question, wording or questions being presented in such a way that inaccurate or uninterpretable answers are obtained.

The interviewer error may be survey interviewer error and process error.

Survey Interviewer Error

(i) Falsification error

Deliberate falsification of survey item responses in any way, including filling out partial answers, for any reason. It can be of following types:

Types

1. Fabricating an interview – the recording of data that are not provided by a designated survey respondent and reporting them as answers of that respondent;
2. Deliberately misreporting disposition codes and falsifying process data (e.g., the recording of a refusal case as ineligible for the sample; reporting a fictitious contact attempt);
3. Deliberately miscoding the answer to a question in order to avoid follow up questions;
4. Deliberately interviewing a non-sampled person in order to reduce effort required to complete an interview; or
5. Intentionally misrepresenting the data collection process to the survey management.

How to avoid or control: Periodic monitoring of data collection process and selective re-interviewing to detect and deter falsification

(ii) Questioning error

Interviewer using leading questions or different wording from one respondent to the next

How to avoid or control: Using structured questions; periodic observation of interviewers during work

(iii) Surrogate respondent error

Deliberate falsification of any portion of subject selection process or specified respondent for Interview

How to avoid or control: Periodic monitoring of data collection process and select re-interviewing to detect and deter Falsification

(iv) Response option error

Interviewer failing to follow response option instructions correctly by reading, or not reading, response options

How to avoid or control: Comprehensive training of interviewers. Clear instructions should be given on option-reading policy included in the interviewer's question list

(v) Question error

Question error occurs when interviewers change the wording or sequence of questions, or when respondents do not provide an answer that relates to the topic construct upon which the question was constructed

How to avoid or control: Developing a coding scheme for assessing the behaviours displayed by interviewers and respondents during a question-and-answer survey process is one way that researchers can check for question error at the source

Process Error

(i) Instruction error

- Interviewer skipping or incorrectly paraphrasing instructions that result in respondent deviation from instructions
- How to avoid or control
- Providing clear, concise, and unambiguous instructions
- Consistent and complete follow-through by interviewers

(ii) Recording error

- Interviewer deliberately abbreviating or omitting portions of verbal responses to unstructured questions
- How to avoid or control
- Supervisor must monitor interview discussion and separately record responses for comparison
- Review of all responses for outliers

(iii) Interpretation error

- Difficulties in interpreting oral or written responses given to open-ended question
- How to avoid or control
- Avoid using unstructured questions unless optional answers are provided to respondents

(iv) Scale interpretation error

- Interviewer deliberately simplifying scale options, such as using numbers instead of scale options; failure to note reverse scaling
- How to avoid or control
- If scale cards are used, the wording and numbers used on the card should also be included as used on the interviewers working survey

(v) Data capture and recording error

- Process error that is often attributed to deliberate or accidental mistakes made by interviewers in reporting respondents' answers to unstructured (open-ended) questions

(vi) Editing error

- After data collection, the researcher employs a procedure that inadequately locates errors.

(vii) Coding error

- Coding error occurs as responses to open-ended questions are classified and assigned a form that can be used in tabulating and processing survey data

Non-response error

- Non-response error occurs when sampling units selected for a sample are not interviewed
- Sampled units typically do not respond because they are unable, unavailable, or unwilling to do so.

Why problematic

- Introduces systematic bias into the data. This results in poorer data quality and can significantly bias any of the estimates derived from the data.
- Missing completely at random (MCAR) or being systematic.
- MCAR - If non-response is MCAR, then there is no underlying reason why certain sampling units failed to complete the survey (No bias).
- Systematic bias, as discussed above, occurs when there is some underlying reason why sampling units do not participate in the survey. This will bias any results based upon the data to the extent to which respondents differ from non-respondents on variables of importance to the analysis.

Minimizing non-response

Call backs/reminders – Researchers should contact sampling units multiple times during the data collection period with reminders to complete the survey.

Refusal conversions – If an individual has explicitly refused to complete the survey but has not asked the researcher to cease additional contact, a common tactic is to employ staff skilled at response conversion techniques to convince that respondent to participate.

Incentives – Respondents usually feel no obligation to complete a survey. They usually do not know a researcher and do not care much about the survey itself, if at all. Frankly, respondents are doing the researcher a favour by participating. Offering an incentive to participants can be an extra boost that convinces many to participate.

Oversampling – If there are certain sub-groups that a researcher suspects will show lower response rates, a common technique is to over-sample that group.

Response Bias in Questionnaire Survey

When the researchers employ a self-report questionnaire for measuring psychological attributes, the following errors and biases are noticed:

(1) Ordering effects

- The order in which questions are asked in a questionnaire can have a significant effect on the results.
- The preceding questions provide the context in which the respondent answers an item, and changing this context can make a large difference in the survey results. This is called 'priming'.
- How to avoid or control – Randomize the questions so that respondents are not answering all questions in the same order.

(2) Wording effects

- Given alternative responses, the respondent's choices tend to be sensitive to the language used to express an alternative.
- For example, when respondents are asked, "Should the Government permit GMO crops?" most respondents said, "Yes". But, when asked, "Should the Government ban GMO crops?" most respondents still said, "Yes". They said so mainly because of their obedience to Government, and not based on their opinion.
- How to avoid or control: Use only neutral questions

(3) Scaling effects

- Response types are arranged in a certain order so that it can influence respondents to respond in the same order.
- For example, when studying farmers' attitudes towards adopting eco-friendly production practices, the researcher provides response types beginning with 'Strongly Agree' to 'Strongly Disagree' for all questions. It is quite likely that the respondents will choose only 'Strongly Agree' for all statements.
- How to avoid or control: Randomise the question order so that the responses will be independent of scaling effects.

(4) Respondent fatigue

- The fatigue generated from a long or complex questionnaire will lead to non-response or random response.
- How to avoid - Styling and colouring – use plenty of colour combinations, graphics, logos, etc.

(5) Social desirability bias

Respondent bias created by the unwillingness to provide honest answers stems from the participant's natural desire to provide socially acceptable answers in order to avoid embarrassment or to please the organization conducting the study.

How to avoid – Avoid using questions which will create social desirability bias.

(6) Non-attitudes

- Respondents may not have an opinion on an issue, but will feel obliged to express one in a questionnaire.
- The attitudes supposed to be measured may not in fact exist in any coherent form.
- How to avoid – Rechecking and triangulation

(7) Acquiescence

- Yea-saying bias – Respondents' tendency to agree with items regardless of their content.
- How to avoid – Rechecking and triangulation

(8) Leniency or harshness

- Individuals systematically may respond more negatively or positively, regardless of the question posed.
- This response tendency is specific and consistent with the respondent.

(9) Critical event and recency

- Associated with respondents' recall.
- Critical event response bias occurs when a dramatic event is given a greater weight in the evaluation than routinely occurring events.
- Recency response bias occurs when events or information presented more recently are weighted more heavily by the respondent than events or information presented in the more distant past.

(10) Halo effect

Halo effect occurs when a participant's response to a previous question serves as a trigger for determining responses to subsequent questions. For example, if a question that asks people if they have heard of 'the profitability of Noni cultivation' and then, later in the questionnaire they are asked to type the names of all commercial crops suitable for their region, there is a good chance that 'Noni' will get a higher result in the latter question than otherwise.

(11) Extreme response style

Extreme response style is the 'tendency to endorse the most extreme response categories regardless of the item content.' When given a survey with a 5-point (Likert) scale, a particular respondent will always respond using one of the two end points, either a '1' or a '5'" even though the respondents' view of some items may not be either extreme.

(12) Midpoint response style

Reflects the 'tendency to use the middle scale category (or most moderate response alternative) regardless of the content'. When provided with the same 5-point (Likert) scale, a particular respondent repeatedly responds with the neutral mid-point alternative, '3', even though the respondent's view of many items may not be neutral.

- Survey errors are caused by sampling and non-sampling sources.
- Sampling errors are caused by inadequacies and errors during the sampling process.
- The non-sampling errors are researcher-, instrument- or respondent-based.
- The self-report questionnaire can create a halo effect, social desirability bias, central tendency bias, respondent fatigue bias, acquiescence, etc.
- Managing errors and biases require a systematic approach in selecting, designing, executing, and coding data.

DATA EXPLORATION AND PREPARATIONS

Prof. Souvik Ghosh
Hony. Director
Agro-Economic Research Centre
Visva Bharati
Email: dir.aerc@visva-bharati.ac.in

Data exploration and preparation is an essential step in extension research which makes the data amenable to statistical analysis.

In extension research, the data generated from field research are often subjected to biases and errors which need to be rectified before analysing them.

Purpose of data exploration

- Identifying patterns in the data – representativeness of respondents, sampling adequacy;
- Locating and correcting errors – missing data, outliers;
- Choosing right data analytical tools by examining whether the data satisfy the necessary statistical assumptions.

Steps in data exploration and preparation

A. Variable identification

B. Missing data treatment

C. Outlier treatment

D. Testing assumptions for statistical analysis

E. Data transformation

A. Variable Identification

After collection, the data can be entered in the Spread sheet (example, MS EXCEL, SPSS) by identifying its nature (dependent or dependent), type (string or numerical) and variable category (categorical and continuous).

The labels/names of variables and values for each variable (example, in an interval scale, 1 – 'Strongly disagree' to 5 – 'Strongly agree') are entered.

B. Missing Data Treatment

- Missing data refers to absence of valid values on one or more variables in the data sheet.
- Missing data can reduce the sample size and power of a model leading to estimates with poor model fit.
- Purpose - to identify the patterns and relationships underlying missing data so that the original data distribution pattern can be maintained.

Reasons for missing data

- Data entry – When the data is transferred from a questionnaire to a data sheet, a few values maybe missed by the data entry operator.
- Data extraction – When the data is extracted from a database or from an Excel spreadsheet to statistical software, like SPSS, a few values may be missed due to extraction errors.
- Data collection – The errors occurring during data collection are difficult to correct. They can be categorized under four types

(a) Missing completely at random

The propensity for a data point to be missing is completely random;

For example, the data on 'land holding' may be missing across the gender categories randomly.

(b) Missing randomly

The propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

For example, if a respondent didn't answer a question about 'Annual income of the family', it is because

(i) She may be a housewife who didn't know the exact annual income of her husband; or
(ii) She may not be aware of income from secondary sources other than the ones she already knows; or
(iii) She didn't want to disclose the real income.

(c) Missing values that depend on unobserved predictors

- Where the missing values are not random and are related to the unobserved input variable.
- For example, when conducting effectiveness of multimedia module on 'Ornamental Fish Farming', some participants may not answer the post-test of knowledge as they couldn't understand some part of the module.

(d) Missing that depends on the missing value itself

- Where the probability of missing value is directly correlated with missing value itself.
- For example, people with higher or lower annual incomes are likely to skip the question related to annual income.

Treatment of missing values

Missing values can be ignored, deleted, or remedied based on the nature of missing values and nature of the study. Various methods for treating missing data are

1. Ignore the missing values

- If the missing data are about 10% for an individual case or variable, it can be ignored;
- If the cases with no missing data are sufficient for selected data analysis method.

2. Deletion

- Variables with 15% of missing data can be deleted. However, if the missing data exceeds 20%, they can be remedied;
- Cases with missing data in dependent variable are deleted, as they are likely to affect the relationships specified for the study;
- Conduct the statistical analysis with and without the case or variable to check the differences in the coefficients. If there is no marked difference, delete the specific case or variable;
- Deletion methods are used when the nature of missing data is 'Missing completely at random'.

Methods

(i) List-wise deletion or complete-case analysis

- Removes all data for a case or variable that has one or more missing values;
- Used when (a) data are 'Missing completely at random'; (b) deletion will not affect the power of the analysis even with reduced sample size.

(ii) Pair-wise deletion or using all available data approach

- Analysis is performed with all cases in which the variables of interest are present;
- Advantage – Sufficient cases are available for analysis;
- Disadvantage – Uses different sample size for different variables.

3. Mean/Mode/Median Imputation

- Imputation – Method for filling the missing data with estimates from valid values of other cases;
- Uses relationships identified from the valid values of the data for estimating the missing values;
- Popular method where the missing data for a specific attribute are replaced with the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

Types

(i) Generalized Imputation

The mean or median for all non-missing values of that variable are calculated and substituted for missing values;

Methods

1. Series mean – Replaces missing values with the mean for the entire series;

2. Mean of nearby points – Replaces missing values with the mean of valid surrounding values;

3. Median of nearby points – Replaces missing values with the median of valid surrounding values;

4. Linear interpolation – Replaces missing values using linear interpolation. The last valid value before the missing value and the first valid value after the missing value are used for the interpolation. If the first or last case in the series has a missing value, the missing value is not replaced;

5. Linear trend at point – Replaces missing values with the linear trend for that point. The existing series is regressed on an index variable scaled 1 to n. Missing values are replaced with their predicted values.

(ii) Similar case Imputation

The average is calculated for each category of the variable (e.g., gender – male and female; race – Indian and non-Indian) and missing values in the respective categories are substituted with estimated values.

4. Prediction Model

- Creating a prediction model from the valid data to estimate values for substituting missing data;
- Statistical techniques, such as multiple regression, ANOVA, Logistic regression and various modelling techniques are used to estimate missing values.

5. K Nearest-Neighbour (KNN) Imputation

- The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing, for example, knowledge;
- The similarity of two attributes is determined using a distance function.

C. Outlier Treatment

Outliers in the social data can create serious issues that affect accuracy of the results of data analysis.

Outliers

- Outliers – Observations with a unique combination of characteristics identifiable as distinctly different from the other observations in the data sheet;
- Such an observation appears far away from the other observations and diverges from the overall pattern of the sample;
- Unusually high or low value on a variable.

Causes of outliers

- **Data Entry Errors** – Human errors caused during data collection, recording, or data entry. Typographical mistakes can add or delete a Zero in the number value and this will create outliers
- **Measurement Error** – Occurs when using a faulty instrument
- **Experimental Error** – Errors caused when conducting an experiment
- **Intentional Outlier** – Reporting false information in the self-reported measures which involves sensitive data (for example, drug abuse data)
- **Data Processing Error** – Manipulation or extraction errors when capturing data into statistical software from the spread sheet
- **Sampling Error** – Choosing an unrepresentative sample
- **Natural Outlier** – Inbuilt into the measurement system.

Impact of outliers

- Increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed they can decrease normality
- They can bias or influence estimates that may be of substantive interest;
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Sources of outliers

- Procedural error – Data entry error or mistake in data coding
- Extraordinary event – Any extraordinary event that occurred in the study area which affects the response (e.g., cyclone wipes out the crop and the yield of improved variety under this disastrous condition is very low)
- Extraordinary observation – A few respondents who are exceptionally different from others (e.g., a small farmer who uses Internet to gather recent developments in agriculture is likely to score high on the INNOVATIVENESS scale)
- Unique in combination – Retain them for analysis.

Types of outliers

(a) Global outlier or point anomaly: Observations which deviate from the rest of the entire data set

(b) Contextual outlier or conditional outlier

- An outlier deviates significantly based on a selected context or behaviour;
- For example, an illiterate farmer showing a high level of comprehension of a printed extension folder;
- Used mostly in time-series data and spatial data.

Two types

- Contextual attributes: defines the context (for example, time and location);
- Behavioural attributes: characteristics of the object, used in outlier evaluation (example, knowledge and intelligence).

(c) Collective outliers

- A group of observations which collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers;
- For example, a small farmer showing attributes of large farmers, where several variables of this specific case deviate from the average of other small farmer cases;
- Collective outliers have been explored for sequence data, graph data and spatial data.

(d) Real outliers

- Real outlying observations which are of interest to the researcher;
- Retained for analysis.

(e) Erroneous outliers

- An observation designated incorrectly as an outlier, due to some inherent problem, or some catastrophic failure.

Methods for detecting outliers

Outliers can be detected in two ways:

(i) Based on whether user-labelled examples of outliers can be obtained

1. Supervised methods: The training data sets are available and can classify any unusual observation in normal or outlier class by comparing it against a developed model;

2. Semi-supervised methods: The training data sets are available only for normal class and any unusual observation is compared against the normal class;

3. Unsupervised methods: do not require training data, and thus are most widely applicable. Make the implicit assumption that the data from the normal instances are far more frequent as compared to the outliers.

(ii) Based on assumptions about normal data and outliers

1. Statistical methods

- To detect the outlying observations, and analyze them to study the complete dataset based on them;
- Assume that the normal data follow some statistical model and the data not following the model are outliers;
- Methods – Box plot, maximum likelihood, Grubbs test, Mahalanobis distance and Chi Square test.

2. Proximity-based methods

- An object is an outlier if the nearest neighbours of the object are far away, i.e., the proximity of the object significantly deviates from the proximity of most of the other objects in the same data set;
- Three methods – K Nearest Neighbour analysis (KNN), Clustering method and density based methods.

3. Parametric techniques

- Fitting the data into a model and identifying outliers;
- Popular method – Regression analysis.

4. Non-parametric methods

- No assumptions about distribution of data;
- Methods – histograms, Kernel Density Function or Kernel Feature Space.

5. Distance-based Methods

- Most widely accepted and frequently used techniques in machine learning and data mining;
- Based on Nearest Neighbour principle.

6. Density-based Methods

- Complex method based on Local Outlier Factor (LOF).

7. Clustering-based methods

- Partitioning Clustering Method – Partitioning methods, various centroid based methods, mediods based methods, PAM, CLARA, k-means, and CLARANS, etc., methods are used;

- Two types – Agglomerative methods and divisive methods;
- Hierarchical Methods – MST clustering, CURE, CHAMALEON and BIRCH.

Statistical methods for detecting outliers

(i) Univariate outliers

- Concerned with a single variable;
- The distribution of observation for every variable is examined as well as select cases falling at the outer ranges (high – low).

Methods

(i) Box plot

Identification of outliers

- Any value, which is beyond the range of -1.5 x Inter Quartile Range to 1.5 x Inter Quartile Range are outliers
- Any value which is out of range of 5th and 95th percentile can be considered as an outlier
- Data points, three or more standard deviation away from mean are considered an outlier.

(ii) Trimmed mean method

- Used to determine the extent of a problem likely to be created by the outliers;
- The original mean of a variable is compared with the 5% trimmed mean (the new mean calculated after the top and bottom 5 percent of cases are removed from the distribution)
- If both means are similar, it can be concluded that the outlying values are not too different from the distribution, which can be retained

(ii) Bivariate and multivariate outlier detection

Bivariate detection

- Required for correlation, regression and other analyses involving two variables;
- Outliers can be identified through Scatter plots.

Multivariate detection

- Required for conducting confirmatory factor analysis in Structural Equation Modelling;
- Methods – Mahalanobis D2 , Cook's D and Madira's coefficient of multivariate kurtosis (For Structural Equation Modelling);
- Thresholds for D2 / df – Small sample 2.5; large sample – 4

Treatment of outliers

(i) Ignore outliers: If the global outliers are satisfying 5% trimmed mean criteria, they may be retained.

(ii) Deleting observations

- Delete outlier values only if it is due to data entry error, data processing error, or if outlier observations are very small in number.
- The contextual outliers may be deleted.

(iii) Transforming and binning values

- Transforming variables can also eliminate outliers
- Natural log of a value reduces the variation caused by extreme values
- Binning is also a form of variable transformation
- Decision Tree algorithm allows to deal with outliers well due to binning of variable
- We can also use the process of assigning weights to different observations.

D. Testing Statistical Assumptions

- Testing statistical assumptions for data analysis is an essential step in the data exploration and preparation process.
- As many statistical methods work on specific assumptions, it is necessary to test the data to check that it complies with these distributional requirements.

Testing assumptions for statistical analysis

- The inferential statistical methods are based on certain assumptions representing the requirements of the statistical theory.
- Four major assumptions – normality, linearity and absence or correlated errors.

I. Normality testing

- Many parametric statistics require normally distributed data for performing analysis;
- Type of normality – Univariate (all parametric analysis) and Multivariate (Structural Equation Modelling).

Methods for testing normality

a. Univariate normality

i. Skewness and kurtosis

- If skewness and kurtosis exceed the absolute value of 1, we assume that there is non-normality

ii. Normal probability plot

- The normal probability plot (Chambers et al., 1983) is a graphical technique for assessing whether or not a data set is approximately normally distributed;
- The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line;
- Departures from this straight line indicate non-normality.

b. Bivariate normality

- Scatter plots are widely used to detect bivariate normality.

c. Multivariate normality

- The multivariate normality of the data is assessed through Mahalanobis Distance (D2) and Mardia's multivariate kurtosis.
- Mahalanobis Distance – A detailed description of calculating Mahalanobis Distance is described in Arifin (2015).
- Mardia's multivariate kurtosis – The Madira's coefficient can be estimated through SPSS AMOS and a detailed procedure.

II. Linearity

- The variable should be linearly related with other variables.
- Important assumption for correlation-based techniques.
- Used for multiple regression, logistic regression, factor analysis and structural equation modelling.

III. Multicollinearity

- Indicates the presence of high correlation between independent variables.
- Undesirable character which causes inaccurate prediction.
- Essential requirement for OLS estimation (Multiple regression).
- Presence of multicollinearity – Tolerance value < 0.2 and VIF > 5.

5. Data Transformations

Purpose

- To correct errors in the data;
- To improve the relationship among variables.

Methods of transformations

Non-normality

- Flat distribution – Inverse (1/X or 1/Y);
- Negatively skewed transformations – Squared or cubed;
- Positive skewness – Logarithm or square root.

- Testing the data to check if it satisfies statistical assumption is an important component of data exploration and preparation phase.
- Univariate and multivariate normalities are necessary assumptions for factor and confirmatory factor analyses.

Source: Sivakumar et al. (2017)

REFERENCES

- Babbie, E. (2008). The basics of social research. 4th ed. Belmont, CA, USA; Thompson Wordsworth.
- Creswell, J.W. (2009.) Research design: Qualitative, quantitative, and mixed methods approach. Third edition. Thousand Oaks: Sage Publications.
- Creswell, J.W. (2012.) Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Fourth edition. Pearson.
- Harvey, D. (2015, October). ACE8001: What do we mean by Research? and Can we hope to do genuine Social Science Research. Retrieved from: https://www.staff.ncl.ac.uk/david.harvey/AEF801/What.html
- ICAR (2020). Handbook of Agricultural Extension. Indian Council of Agricultural Research, New Delhi.
- Kerlinger, F.N. and Lee, H.B. (2000.). Foundations of behavioral research. Fourth Edition. Fort Worth: TX: Harcourt College Publishers.
- Mertens, D.M. and Wilson, A.T. (2012.) Program Evaluation: Theory and Practice. New York, NY: Guilford.
- Ray, G.L. and Mondal, Sagar, (2011). Research Methods in Social Sciences and Extension Education, Kalyani Publishers.
- Sivakumar, P.S., Sontakki, B.S., Sulaiman, R.V., Saravanan, R. and Mittal, N. (eds). (2017). Good Practices in Agricultural Extension Research. Manual on Good Practices in Extension Research and Evaluation. Agricultural Extension in South Asia. Centre for Research on Innovation and Science and Policy (CRISP), Hyderabad. India.
- Tripathi, P.C. (1991). A Text Book of Research Methodology in Social Sciences. Sultan Chand & Sons, New Delhi

DESCRIPTIVE STATISTICS AND EXPLORATORY DATA ANALYSIS

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

Introduction

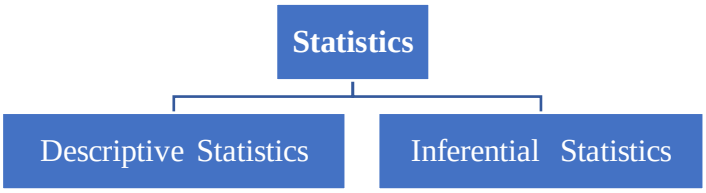
The word 'Statistics' has been derived from the Latin word '**Status**' or the Italian word '**Statista**' or the German word '**Statistik**' each of which means 'political state'. Statistics is a broad concept featuring applications in a wide range of areas. Statistics, in general, can be defined as the process for collecting, analyzing, interpreting, and making conclusions from data. In other terms, statistics is the approach established by scientists and mathematicians for analyzing and deriving conclusions from acquired data. Everything that has anything to do with the collection, processing, interpretation, and presentation of data falls within the scope of statistics.

Definition of statistics: Statistics is a branch of mathematics that deals with collecting, organizing, summarizing, presenting, and analyzing data as well as providing valid results and interpreting towards reasonable decisions.

Statisticians, in other words, give methodologies for

- **Design:** Planning and conducting out research projects.
- **Description:** Data summarization and exploration.
- **Inference:** Making predictions and inferences about the data

Statistics can be divided into two sections; one is descriptive statistics and another is inferential statistics.



Descriptive statistics helps describe, show or summarize data in a meaningful way. Descriptive statistics provides us with tools, tables, graphs, averages, ranges, correlations for organizing and summarizing data. Examples: measures of central tendency, measures of dispersion, skewness, kurtosis etc.

Inferential statistics helps to understand the properties of the population by observing the sample values. Inferential statistics deals with the estimation of parameters and test of hypothesis. In this section we briefly discussed the descriptive statistics such as measures of central tendency, measures of dispersion, skewness, and kurtosis.

Measures of central tendency

Central tendency is a statistical measure that determines a single value that accurately describes the center of the distribution. The objective of central tendency is to identify the single value that is the best representative for the entire set of data.

Different measures of central tendency are:

- Mean
    - Arithmetic mean
    - Geometric mean
    - Harmonic mean
- Median
- Mode
- Quartiles
- Deciles
- Percentiles

Mean (Arithmetic mean: A.M.):

The mean is the most commonly used measure of central tendency. For computation of the mean data should be numerical values measured on an interval or ratio scale. To compute the mean, we add the observation of data sets and then divide by the number of observation.

$$Mean = \frac{Sum\ of\ all\ observation}{Total\ number\ of\ observationa}$$

Simple mean: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Mean for frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The arithmetic mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

Properties of mean:

- It depends on change of origin as well as the change of scale.

$$U = a + hX$$

Origin — Scale

Then $\bar{U} = a + h\bar{X}$.

- If are $\bar{X}_1$and $\bar{X}_2$ the means of two sets of values with $n_1$and n_2 observations respectively, then their combined mean is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

- Algebraic sum of deviations of set of values from their mean is zero.

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- The sum of squares of deviation of set of values about its mean is minimum

$\sum_{i=1}^{n}(X_i - A)^2$ is minimum when $A = \bar{X}$

Merits of mean:

- Easy to understand
- Easy to calculate.
- It is rigidly defined.
- It is based on all observations.
- It is least affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of mean:

- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.
- It is not suitable for highly skewed distribution.

Geometric mean (G.M.):

For n observations, Geometric mean is the n^{th} root of their product.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The geometric mean is defined as

$$G = (X_1 * X_2 * \dots * X_n)^{1/n}$$

For frequency distribution: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The geometric mean is defined as

$$G = ({X_1}^{f_1} * {X_2}^{f_2} * \dots * {X_n}^{f_n})^{1/N}$$

Use of geometric mean:

- Measure average relative changes, averaging ratios and percentages
- Best average for construction of index number

Merits of geometric mean:

- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of geometric mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristic.
- It cannot be calculated if any observations are missing in the data series.

Harmonic mean (H.M.):

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the observations of the sets.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The harmonic mean is defined as

$$H = \frac{n}{\sum_{i=1}^{n} 1/X_i}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The harmonic mean is defined as

$$H = \frac{N}{\sum_{i=1}^{n} f_i/X_i}$$

Use of harmonic mean:

- Measure the change where the values of a variable are compared with a constant quantity of another variable like time, distance traveled within a given time, quantities purchased or sold over a unit.

Merits of harmonic mean:

- It gives more weight to the small item and less weight to large values.
- It is based on all observations.
- It is not affected by sampling fluctuations.
- It is capable of further mathematical treatment.

Demerits of harmonic mean:

- If any of the values is zero, it cannot be calculated.
- It is affected by extreme values.
- It cannot be calculated for open end class frequency distribution.
- It cannot be located graphically.
- It cannot be calculated for qualitative characteristics.
- It cannot be calculated if any observations are missing in the data series.

Relation between A.M., G.M. and H.M.:

- For given two observations, $A.M. \geq G.M. \geq H.M.$
- $G.M. = \sqrt{A.M. * H.M.}$
- $A.M. = \frac{G.M.^2}{H.M.}$

- $H.M. = \frac{G.M.^2}{A.M.}$

Median:

Median is the value situated in the middle position when all the observations are arranged in an ascending/descending order. The median is the central value of an ordered data series. It divides the data sets exactly into two parts. Fifty percent of observations are below the median and 50% are above the median. Median is also known as 'positional average'. The Median is the 50^{th} percentiles, 10^{th} deciles, and 2^{nd} quartiles. Median is also the intersect point of less than and more than ogive curve.

Median for non-frequency data:

Step 1 Order the data from smallest to largest.

Step 2 If the number of observations is odd, then $(n + 1)/2^{th}$ observation (in the ordered set) is the median. When the total number of observations is even, the median is given by the mean of n/2th and $(n/2 + 1)^{th}$ observation.

Median for group frequency data:

Step 1 Obtain the cumulative frequencies for the data.

Step 2 Mark the class corresponding to which a cumulative frequency is greater than N/2. That class is the median class.

Step 3 Then median is evaluated by an interpolation formula

$$Median = l + \frac{h}{f}(\frac{N}{2} - C)$$

Where, l = lower limit of the median class

N= Number of observations

C = cumulative frequency of the class proceeding to the median class

f = frequency of the median class

h= magnitude of the median class

Note: Graphically, we can find the median by histogram.

Use of median:

- Qualitative data can be arranged in ascending or descending order of magnitude.
- Find average intelligence, honesty, etc.

Merits of median:

- It is rigidly defined.
- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on an ordinal scale.

Demerits of median:

- It is not based on all observations.

- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compared to the mean, it is much affected by sampling fluctuations.

Mode:

Mode is defined as the value that occurs most frequently in the data. If in the data sets each observation occurs only once, then it does not have mode. When the data set has two or more values equal to the highest frequency than two or more mode are present in the datasets.

Mode for ungroup frequency data: The observation which has the highest frequency in the data sets.

Mode for group (equal width) frequency data:

Step 1 Identify the modal class. Modal class is the class with the largest frequency.

Step 2 Find mode by using interpolated formula.

$$mode = l + \frac{h(f_0 - f_{-1})}{(f_0 - f_{-1}) - (f_1 - f_0)}$$

Where, l = lower limit of the modal class

f_0 = frequency of the modal class

f_{-1}= frequency of the preceding modal class

f_1= frequency of the succeeding modal class

h= magnitude of the modal class

Note: Graphically, we can find mode by histogram.

Use of mode:

- To find ideal consumer preferences for different kinds of products.
- The best measure for the average size of shoes or shirts.

Merits of mode:

- It is not affected by extreme values.
- It can be located graphically.
- It can be calculated for open end class frequency distribution.
- It can be calculated for data based on a nominal scale.

Demerits of mode:

- It is ill-defined.
- It is not based on all observations.
- The calculation is more complex than the mean.
- It is not capable of further mathematical treatment.
- As compare to the mean, it is much affected by sampling fluctuations.

Quartiles: Quartiles are the three points that divide the whole data into four equal parts.

$$Q_i = l + \frac{h}{f}(\frac{iN}{4} - C)$$

Deciles: Deciles are the nine points that divide the whole data into ten equal parts.

$$D_i = l + \frac{h}{f}(\frac{iN}{10} - C)$$

Percentiles: Percentiles are the ninety-nine point that divides the whole data into hundreds of equal parts.

$$P_i = l + \frac{h}{f}(\frac{iN}{100} - C)$$

Note: $\boldsymbol{Median = 2nd\ Quartles = 5th\ Deciles = 50th\ Percentiles}$

Empirical formula between mean median and mode: If the data sets are asymmetric in nature, then

$$\boldsymbol{Mean - Mode = 3(Mean - Median)}$$

The best measure of central tendency:

According to prof. Yule, Mean is the best measure of central tendency. But there are some situations where the other measures of central tendency are preferred.

Scale	Use measure	Best measure
Interval	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ratio	Mean, Median, Mode	Symmetrical data: Mean Asymmetrical data: Median
Ordinal	Median, Mode	Median
Nominal	Mode	Mode

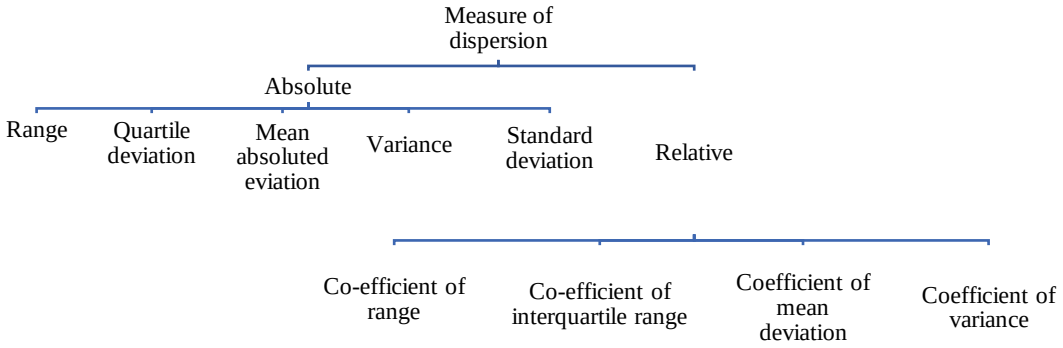
Measures of Dispersion

The measure of central tendency such as mean, median, and mode only locate the center of the data. It does not infer anything about the spread of the data. Two data sets can have the same mean but they can be entirely different.

Data 1	38	42	41	44	45
Data 2	50	53	41	35	31

In the above example, two datasets have the same mean. So measures of central tendency are not adequate to describe data. Thus to describe data, one needs to know the measure of scatterness of observations. Dispersion is defined as deviation or scatterness of observations from their central values.

Various measures of dispersion are:



Range (R):

Range is the simplest measure of dispersion. It is defined as the difference between the highest value and lowest value of the variable. It is a crude measure of dispersion.

$$\boldsymbol{Range = highest\ value\ (H) - lowest\ value\ (L)}$$

Merits of range:

- It is easy to understand and calculate.
- It is not affected by frequency of the data.

Demerits of range:

- It does not depend on all observations.
- It is very much affected by the extreme items.
- It cannot be calculated from open-end class intervals.
- It is not suitable for further mathematical treatment.
- It is the most unreliable measure of dispersion.

Quartile deviation (Q.D.):

Interquartile range is the difference between the first and third quartile. Hence the interquartile range describes the middle 50% of observations.

$$\boldsymbol{Inter\ quartile\ range = Q3 - Q1}$$

Where,

Q^3= first quartile of the data

Q^1= third quartile of the data

Quartile deviation (Q.D.) is the half of the interquartile range.

$$\boldsymbol{Quartile\ deviation\ (Q.D.) = \frac{Q3 - Q1}{2}}$$

Merits of Quartile deviation:

- It is easy to understand and calculate.
- It is not affected by extreme values
- It can be calculated for open end frequency data

Demerits of Quartile deviation:

- It does not depend on all observations.
- It is not suitable for further mathematical treatment.
- It is very much affected by sampling fluctuations.

Mean absolute deviation (MAD):

The absolute deviation of each value from the central value (mean is preferable) is calculated and the arithmetic mean of these deviations is called mean absolute deviation.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observations of a data set. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} |X_i - A|}{n}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

For frequency data: Let $X_1, X_2, \dots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \dots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The mean absolute deviation (MAD) about A is given by

$$MAD_A = \frac{\sum_{i=1}^{n} f_i |X_i - A|}{N}$$

The mean absolute deviation (MAD) about mean is given by

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} f_i |X_i - \bar{X}|}{N}$$

Merits of mean absolute deviation about mean:

- It is easy to understand and calculate.
- It is based on all observations.

Demerits of mean absolute deviation about mean:

- It is not suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.

Standard deviation (S.D.):

It is the best measure and the most commonly used measure of dispersion. It is defined as the positive square-root of the arithmetic mean of the square of the deviations of the given observation from their arithmetic mean. It takes into consideration the magnitude of all the observations and gives the minimum value of dispersion possible. It is also known as Root Mean Square Deviation about mean.

For non-frequency data: Let $X_1, X_2, \dots, X_n$ are the n observation of a data set. The standard deviation A is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

For frequency data: Let $X_1, X_2, \ldots, X_n$ are observations with corresponding frequencies are $f_1, f_2, \ldots, f_n$ and $\sum_{i=1}^{n} f_i = N$. The standard deviation is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}{N}}$$

Properties of standard deviation:

- It is the independent of the change of origin but dependent on the change of scale

Let $U = a + hX$, then $sd(U) = |h| * sd(x)$

- If all observations are equal standard deviation is zero.
- It is never less than the quartile deviation and mean absolute deviation.

Merits of standard deviation:

- It is based on all observations.
- It is less affected by extreme values.
- It is suitable for further mathematical treatment.

Demerits of standard deviation:

- It is suitable for further mathematical treatment.
- It does not take the sign of deviation under consideration.
- It is affected by extreme values.
- It cannot be computed for open-end class data.

Variance

It is defined as the square of the standard deviation. Unit of the variance is the square of the actual observations, whereas unit of the standard deviation is same as actual observations.

Relations between R, Q.D., M.D. and S.D.

$$\mathbf{9QD = \frac{15}{2}MD = 6SD = R}$$

Coefficient of Variation (CV):

The Coefficient of variation for a data set defined as the ratio of the standard deviation to the mean and expressed in percentage.

$$\boldsymbol{CV = \frac{SD}{mean} * 100\%}$$

C.V is the relative measure of dispersion. It is the best measure among all the relative measure of dispersion. C.V is used to compare variability or consistency between two or more data series. If C.V. is greater indicate that the group is more variable, less stable, less uniform and less

consistent. If the C.V. is less, it indicates that the group is less variable or more stable or more uniform and more consistent.

Example: Consider the data on score of Kohli and Smith in ODI cricket. The mean and standard deviation for Kohli are 55 and 5 respectively. The mean and standard deviation for Smith are 50 and 10 respectively. Find C.V. value for both the data and make compare them.

Solution:

For Kohli, $CV = \frac{5}{55} * 100 = 9\%$

For Smith, $CV = \frac{10}{50} * 100 = 20\%$

The Smith is subject to more variation in score than Kohli. So Kohli is more consistent than Smith.

$$\textbf{Coefficient of range} = \frac{H-L}{H+L} * 100\%$$

$$\textbf{Coefficient of inter quartile range} = \frac{Q3-Q1}{Q3+Q1} * 100\%$$

$$\textbf{Coefficient of mean deviation} = \frac{MAD}{averave\ from\ which\ it\ is\ calculated} * 100\%$$

Numerical Examples: The marks of 10 students in statistics examination are as follows:

10,12,15,12,16, 20, 13,17,15,15

Find mean, median, mode, range and standard deviation.

Solution:

X_i	f_i	$f_i X_i$	$f_i (X_i - \bar{X})$	$(X_i - \bar{X})^2$	$f_i (X_i - \bar{X})^2$
10	1	10	-4.5	20.25	20.25
12	2	24	-5	6.25	12.5
13	1	13	-1.5	2.25	2.25
15	3	45	1.5	0.25	0.75
16	1	16	1.5	2.25	2.25
17	1	17	2.5	6.25	6.25
20	1	20	5.5	30.25	30.25
Total	10	145		67.75	74.5

$mean = \frac{145}{10} = 14.5$

$median = 15$

$mode = 15$

$range = 20 - 10 = 10$

$SD = \frac{74.5}{10} = 7.45$

Skewness and kurtosis:

We have discussed measures of central tendency and measure of dispersion which describe the location and scale parameter of the data sets. They do not give any idea about the shape of the data structure. The measure of skewness and kurtosis illustrate the shape of the data sets. The measure of skewness gives the direction and the magnitude of the lack of symmetry and the measure of kurtosis gives the idea of the flatness of the curve.

Skewness

Skewness measures the degree of asymmetry of the data. Skewness refers to the lack of symmetry. Skewness is mainly three types: Positive skewness, Negative skewness, and Symmetric data.

Positive Skewness:

A data is said to be positive skew if the long tail is on the right side of the peak. The mean is on the right of the peak value. Here Mean > Median > Mode.

Negative Skewness:

A data is said to be negative skew if the long tail is on the left side of the peak. The mean is on the left of the peak value. Here Mean < Median < Mode.

Symmetric

The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle. When data is symmetrically distributed, the left-hand side, and right-hand side, contains the same number of observations. Here Mean = Median = Mode.

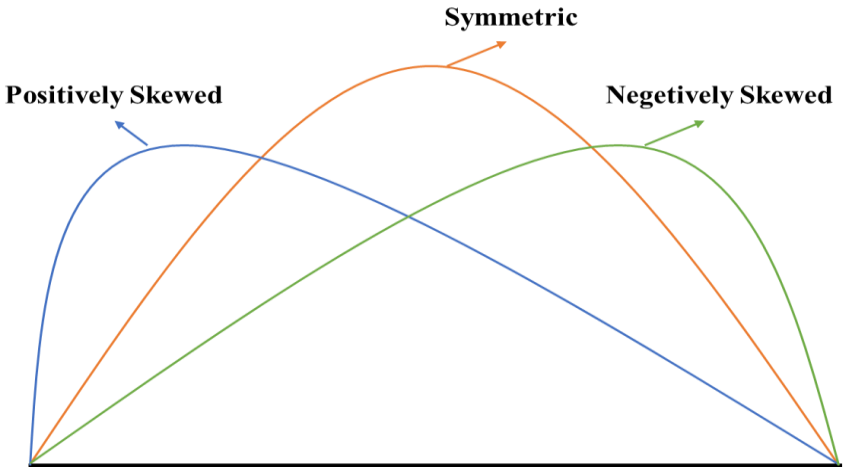


Figure 1. Skewness

The measure of Skewness:

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Interpretation:

1. If $S_k = 0$, then the frequency distribution is normal and symmetrical.
2. If $S_k > 0$, then the frequency distribution is positively skewed.
3. If $S_k < 0$, then the frequency distribution is negatively skewed.

Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails or outliers. Data sets with low kurtosis tend to have light tails or lack of outliers. A uniform distribution would be the extreme case.

Types of kurtosis: Leptokurtic or heavy-tailed distribution, Mesokurtic, Platykurtic or short-tailed distribution

Leptokurtic

Leptokurtic indicates that distribution is peaked and possesses thick tails.

Platykurtic

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is a flatter (less peaked) when compared with the normal distribution.

Mesokurtic

Mesokurtic is the same as the normal distribution. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

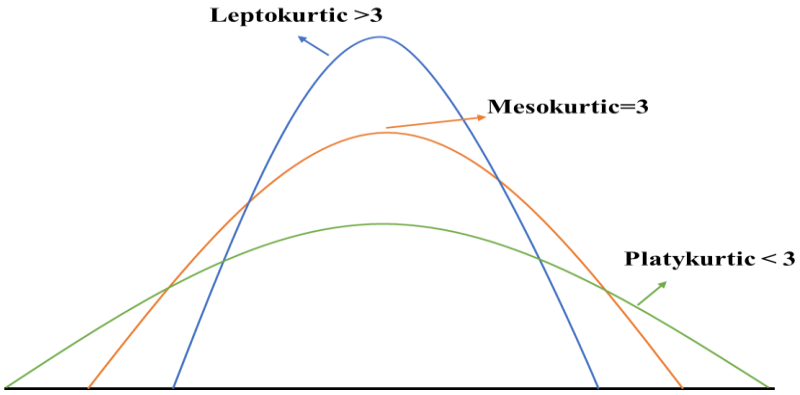


Figure 2. Kurtosis

Measurement of Kurtosis $(\beta_2) = \frac{1}{N-1}\frac{\sum(y_i-\bar{y})^4}{s^4}$

$$\gamma_2 = \beta_2 - 3$$

Data presentation

There are three broad ways of presenting data. These are Textual presentation, Tabular presentation, and Graphic or diagrammatic presentation. We discussed only a few important diagrammatic presentations of data.

Non dimensional diagram	Pictograms
Two dimensional diagram	Bar diagrams, Pie diagrams, Histograms, Box Plot
Three dimensional diagram	Cubes, Cylinders diagrams

Bar Diagram

Simple Bar Diagram

If the classification is based on attributes and if the attributes are to be compared with respect to a single character we use a simple bar diagram. Simple bar diagrams consist of vertical bars of equal width. The heights of these bars are proportional to the volume or magnitude of the attribute. All bars stand on the same baseline. The bars are separated from each other by equal intervals. The bars may be colored or marked.

Multiple bar diagram

If the data is classified by attributes and if two or more characters or groups are to be compared within each attribute we use multiple bar diagrams. If only two characters are to be compared within each attribute, then the resultant bar diagram used is known as the double bar diagram. The multiple bar diagram is simply the extension of a simple bar diagram. For each attribute, two or more bars representing separate characters or groups are to be placed side by side. Each bar within an attribute will be marked or colored differently in order to distinguish them. The same type of marking or coloring should be done under each attribute. A footnote has to be given explaining the markings or colorings.

Component bar diagram

This is also called a subdivided bar diagram. Instead of placing the bars for each component side by side, we may place this one on top of the other. This will result in a component bar diagram.

Histogram

Histograms is suitable for continuous class frequency distribution. We mark off class intervals along the x-axis and frequencies (frequency density for unequal frequency data) along the y-axis.

- For equal class intervals, the heights of the rectangles will be proportional to the frequencies, while for unequal class intervals, the heights will be equal (or proportional) to the frequency densities.
- A frequency polygon is a line graph obtained by connecting the midpoints of the tops of the rectangles in the histogram.

Table 1. Differences between bar diagrams and histograms

Characteristics	**Bar Diagrams**	**Histograms**
Frequency is measured by	Height of the bar	Area of the bar
Gaps between the bars	Yes	No
Width of the bar	Equal	May not be equal
Data types	Discrete and Continuous	Continuous only

Pie diagrams

When we are interested in the relative importance of the different components of a single factor, we use pie diagrams. For the pie diagram, one circle is used and the area enclosed by it being taken as 100. It is then divided into a number of sectors by drawing angles at the center, the area of each sector representing the corresponding percentage.

Box Plot

Minimum, maximum, and quartiles (Q_1, Median, Q_3) together provide information on the center and variation of the variable in a nice compact way. Written in increasing order, they comprise what is called the five-number summary of the variable. A boxplot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of the variable in a data set. It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

REGRESSION ANALYSIS

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

Introduction

Regression analysis is a statistical approach that makes use of the relation between two or more variables so as to predict the value of one variable from another. This methodology is widely used among researchers in the field of business, biology and social and behavioral sciences. For example if one wishes to draw the relationship between how much one eats and how much they weigh. Here the concept of regression can be easily applied

The multiple regression model depends on the estimates of the individual regression coefficients. Some inferences that are frequently made include

1. Identifying the comparative effects of the regressor variables,
2. Prediction and/or estimation, and
3. Selection of appropriate set of variables for the model.

An operative relation between two variables is expressed by a formula. If X represents the independent variable and Y represents the dependent variable, an operative relation is of the form

$$Y = f(X)$$

For a particular value of X, the function f shows the corresponding value of Y. A statistical relation is not always a perfect one. The observations for any statistical relation in general do not lie directly on the curve of the relationship (Bhar, 2015; Paul and Bhar, 2019).

A regression model that has more than one regressor variables is called a multiple regression model. In other words, it is a linear relationship between a dependent variable and a group of independent variables. Multiple regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables. Multiple linear regression models are often used as approximating functions. That is, true functional relationship between y and x is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate estimation to the true unknown function. If the model fits the data well, the overall R^2 value will be high, and the corresponding P value will be low (P value is the observed significance level at which the null hypothesis is rejected). In addition to the overall P value, multiple regressions also report an individual P for each independent variable. A low P value here means that this particular independent variable significantly improves the fit of the model. It is computed by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent

variable is omitted. If the fit is much worse when that variable is removed from the model, the P value will be low, indicating that the variable has a significant impact on the model.

Depending on the nature of the relationships between *X* and *Y*, regression approach may be grouped into two categories, linear regression models and nonlinear regression models. The response variable in general is related to other causal variables via some parameters. The models that are linear along these parameters are known as linear models, and in nonlinear models parameters appear nonlinearly. Linear models in general give satisfactory estimations for most regression applications. At times there are occasions, when an empirically indicated or a theoretically justified nonlinear model is more appropriate (Barnett and Lewis, 1984).

Linear Regression Models

We take into consideration one of the simple linear model in which one is a predictor variable and second is the regression function and both are linear in nature. Further, the model with more than one predictor variable is straight forward in their nature. The model can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

Here Y_i is corresponds to the value of the response variable in the i^{th} trial, β_0 and β_1 both are parameters, X_i is the value of the predictor variable in the i^{th} trial, ε_i is a random error term with mean zero and variance σ^2 and ε_i and ε_j ($i \neq j$) are uncorrelated in nature such that their covariance will be zero.

A regression model (1) is defined to be simple and linear in terms of parameters, and linear in the predictor variable. It is "simple" which means there is only one predictor variable, "linear in the parameters" as there are no parameters that is present in form of an exponent or its multiple or divided by another parameter, and "linear in predictor variable" since its variable present only in the order of first power. A model that is linear in the parameters and in the predictor variable is also called first order model.

Regression parameters

The parameters β_0 and β_1 in regression model (1) are called as regression coefficients, β_1 is identified as the slope of the regression line. It simply implies the change in the mean of the *Y* per unit change in *X*. The parameter β_0 is the intercept of the regression line. When the scope of the model gives $X = 0$, β_0 gives the mean of the probability distribution of *Y* at $X = 0$. When the scope of the model does not cover this $X = 0$, β_0 does not show any special meaning as a separate term in the regression model (Belsley *et al.*, 2004).

Method of Ordinary Least Squares

To evaluate the regression parameters β_0 and β_1, we make use of the method of least squares for the estimation. Using every observation (X_i, Y_i) for each case, we calculate the deviation of Y from its expected value using the method of least squares, $Y_i - \beta_0 - \beta_1 X_i$. To be specific, the main requirement is to take into consideration sum of the n squared deviations. This criterion can be denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2$$

Taking into consideration the definition of method of least squares, β_0 and β_1 are the estimators whose values are b_0 and b_1, respectively, that can minimize the criterion Q for the observations provided which are solutions to the following normal equations:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 .$$

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \frac{1}{n} \left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i \right) = \overline{Y} - b_1 \overline{X} ,$$

where $\overline{X}$ and $\overline{Y}$ are the means of the X_i and the Y_i observations

Inferences in Linear Models

Statisticians want to draw an inference about β_1, the slope of the regression line and, tests concerning β_1 are of interest, especially one of the form:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

The main basis for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there lies no linear relationship between Y and X. Considering the normal error regression model, the condition $\beta_1 = 0$ implies even more than no linear relationship between Y and X. Value of $\beta_1 = 0$ for the normal error regression model means there is no linear relationship between X and Y and as well as there does not exists any form of relationship between X and Y, since the probability distribution of Y are then same at all levels of X.

An explicit test of the alternatives is based on the test statistic:

$$t = \frac{b_1}{s(b_1)},$$

where $s(b_1)$ is standard error of b_1 and can be calculated as $s(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$, where MSE stands for mean square error.

The decision rule with this test statistic when controlling level of significance at α is

if $|t| \le t(1-\alpha/2; n-p)$, do not reject H_0,

if $|t| > t(1-\alpha/2; n-p)$, reject H_0.

Similarly testing for other parameters can be carried out.

Measure of fitting i.e. R^2

Some researchers are interested in estimating the degree of linear association. Here, one descriptive measure has been discussed that is used in practice for describing the degree of linear relationship between *Y* and *X*.

Represented by SSTO = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, total sum of squares which computes the variation in the observation Y_i, or the unpredictability in predicting the value of *Y*, when no account of the predictor variable *X* is taken in consideration. Thus, SSTO is a way of measuring of unpredictability in predicting *Y* when *X* is not considered. Similarly, SSE (Error sum of squares) gives the variation in the *Y*i when a regression model utilizing the predictor variable *X* is employed. A natural measure of the effect of *X* in reducing the variation in *Y*, i.e., in reducing the uncertainty in predicting Y, is to express the reduction in variation [SSTO-SSE=SSR (Regression sum of squares)] as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

The measure R^2 is called coefficient of determination, $0 \le R^2 \le 1$. In general the value of R^2 is not expected to be 0 or 1 but it lies somewhere between the given limits. The more closer it is to 1, the greater is said to be the degree of linear association between *X* and *Y*.

Variables selection techniques

Forward selection procedure

Forward selection procedure makes use of a subset of predictor variables for the final model. Steps to be followed are:

- Begin with a null model. The null model has no predictors, and it has only the intercept.
- Now, start fitting p simple linear regression models, each with one of the given variables (with no of predictor variables as p) and intercept. Here basically, we just search through all the single-variable models to find the best one.
- Now we search through the remaining p minus 1 variables and find out the variable that should be added to the current model in order to improve the residual sum of squares.
- Continue until some stopping rule is satisfied, say when all remaining variables have a P-value above some threshold.

Backward elimination procedure

In order to perform backward selection, we should be in a position where we have more observations than the variables because we can apply least squares regression when n is greater than p. If the value of p is greater than n, we cannot fit a least squares model and it's not even defined. To begin with:

- We start with all the given variables in the model.
- The partial F-test value is calculated for every predictor variable treated as though it was the last variable to come in the regression equation.
- The lowest partial F-test value, say, F_L, is compared with a preselected or default significance level, say, F_0.
    a. If $F_L < F_0$, remove the variable Z_L, which gave rise to F_L, from consideration and recalculate the regression equation for the remaining variables.
    b. If $F_L > F_0$, use the regression equation as calculated.

Stepwise selection procedure

The step wise regression procedure starts by choosing the equation containing the single best X variable and then attempts to build up with subsequent additions of X's one at a time as long as these additions are worthwhile. The order of addition is decided by using the partial F-test values to select which variable should enter next. The highest partial F-value is compared to a (selected or default) F-to-enter value. When variable has been added, the equation is examined to see if any variable should be deleted.

The basic procedure is as follows. First, we select the predictor variable most correlated with Y (suppose it is Z_1) and find the first-order, linear regression equation $\hat{Y} = f(Z_1)$. we check if this variable is significant. If it is not, we quit and adopt the model $Y = \bar{Y}$ as best; otherwise we search for the second predictor variable to enter the regression.

We examine the partial F-values for all the predictor variables not in regression. The predictor variable with the highest such value (suppose this is Z_2) is now selected and a second regression

equation $\hat{Y} = f(Z_1, Z_2)$ is fitted. The overall regression is checked for significance, the improvement in the R^2 value is noted, and the partial F-values for both variables now in the equation are examined. The lower of these two partial F's is then compared with appropriate F percentage point, F-to-remove, and the corresponding predictor variable is kept in the equation or rejected according to whether the test is significant or not significant.

Diagnostics and alternative measures

When a regression model is selected to be applied to an application, we cannot ascertain in advance that the model is appropriate for that particular application, any one, or several other, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to determine the aptness of the model for the dataset before inferences based on that particular model is undertaken. We should check for the errors whether there are departures from:

(i) The linearity of the regression function.
(ii) The constancy of the error variance.
(iii) The independency of the error terms.
(iv) Presence of one or a few outlier observations.
(v) The normal distribution of the error terms.
(vi) One or several important predictor variables have been removed from the model.
(vii) Presence of Multicollinearity.

Nonlinear Regression Models

It is not always possible to make use of linear regression model. For example, an engineer or a scientist might assume the form of the relationship between the response variable and the regressors from his direct knowledge, perhaps from the theory underlying the phenomena. In actuality, relationship between the response and the regressors may be a differential equation, or the solution to a differential equation. Often, this will lead to a model of nonlinear form.
Any model that is not linear in the unknown parameters is a nonlinear regression model. For example, the model

$$y = \theta_1 e^{\theta_2 x} + \varepsilon$$

is not linear in the unknown parameters θ_1 and θ_2. In general, the nonlinear regression model is written as

$$y = f(x, \theta) + \varepsilon$$

where θ is a p x 1 vector of unknown parameters, and ε is an uncorrelated random error term with $E(\varepsilon) = \theta$ and $Var(\varepsilon) = \sigma^2$. It is assumed that the errors are normally distributed, as in linear regression. The function $f(x, \theta)$ is called the expectation function for the nonlinear regression model. This is very similar to the linear regression case, except that now the expectation function is a nonlinear function of the parameters.

For the nonlinear regression model, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. To illustrate these points, consider a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$ under usual notations

With expectation function $f(x,\beta) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$

Now

$$\frac{\delta f(x,\beta)}{\delta \beta_j} = x_j, j = 0,1, \ldots k$$

where $x_0 \equiv 1$. Consider that in the linear case the derivatives are not functions of the $\beta's$.
Now consider this nonlinear model

$$y = f(x,\beta) + \varepsilon$$
$$= \theta_1 e^{\theta_2 x} + \varepsilon$$

The derivatives of the expectation function with respect to θ_1 and θ_2 are

$$\frac{\delta f(x,\theta)}{\delta \theta_1} = e^{\theta_2 x}$$

and

$$\frac{\delta f(x,\theta)}{\delta \theta_2} = \theta_1 x e^{\theta_2 x}$$

Since the derivatives are functions of the unknown parameters θ_1 *and* θ_2, the model is nonlinear in nature.

Fitting of Nonlinear models

Like in the case of linear regression model, in non-linear case also, parameter estimates can be obtained as a result from the 'Method of least squares'. However, minimization of residual sum of squares yield normal equations which are nonlinear in the parameters. Since it is not feasible to solve nonlinear equations exactly, the next option is to obtain approximate analytic solutions by employing iterative procedures. Three main methods of this kind are:

i) Linearization (or Taylor series) method
ii) Steepest Descent method
iii) Levenberg-Marquardt's method

The details of these methods can be found in Draper and Smith (1998); Chatterjee and Price (1977); Kleinbaum and Kupper (1978); Montgomery *et al.* (2003). The results of linear least square theory in a succession of stages uses in the linearization method. Although, neither this method nor the steepest descent method, is ideal. The latter method is able to converge on true

parameter values even though initial trial values are far from the true parameter values, but this convergence tends to be very slow at the later stages of the iterative process. On the other side of linearization method will converge very swiftly provided the vicinity of the true parameter values has been reached, but if initial trial values are too far their convergence may not occur at all.
The most frequently used method of computing nonlinear least squares estimators is the Levenberg-Marquardt's method. This method illustrates a compromise between the other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not 'slow down' at the latter part of the iterative process. The procedure is available in standard statistical software packages.

REFERENCES

- Barnett, V. and Lewis, T. (1984). Outliers in Statistical Data, New York: Wiley.

- Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). Regression diagnostics – Identifying influential data and sources of collinearity, New York.: Wiley

- Bhar, L. (2015). Regression Analysis: Diagnostics and Remedial Measures, pp 245-270. In: Rajni Jain and S S Raju Edition. Decision support system in agriculture using quantitative analysis, Agrotech Publishing Academy, ISBN: 978-81-8321-395-0

- Chatterjee, S. and Price, B (1977). Regression analysis by example, New York: John Wiley & sons

- Draper, N.R. and Smith, H. (1998). Applied Regression analysis, New York: Wiley Eastern Ltd.

- Kleinbaum, D.G. and Kupper, L.L. (1978). Applied Regression analysis and other multivariate methods, Massachusetts: Duxbury Press

- Paul, R. K. and Bhar, L.M. (2019). Linear and Nonlinear Regression Analysis, in Nikam, V., Jhajhria, A. and Pal, S (ed) Quantitative Methods for social sciences, 59-69. ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi. ISBN: 978-81-940080-2-6.

REGRESSION DIAGNOSTICS

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this section we discuss some simple graphic methods for studying the appropriateness of a model, as well as some remedial measures that can be helpful when the data are not in accordance with the conditions of the regression model.

Departures from Model to be studied

We shall consider following six important types of departures from linear regression model with normal errors:

(i) The linearity of regression function.
(ii) The constancy of error variance.
(iii) The independency of error terms.
(iv) Presence of one or a few outlier observations.
(v) The normal distribution of error terms.
(vi) One or several important predictor variables have been omitted from the model.
(vii) Presence of multicollinearity.

Graphical Tests for Model Departures

Nonlinearity of Regression Model

Whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the predictor variable or equivalently from a residual plot against the fitted values.

Figure 1(a) shows a prototype situation of the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative.

Figure 1(b) shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative.

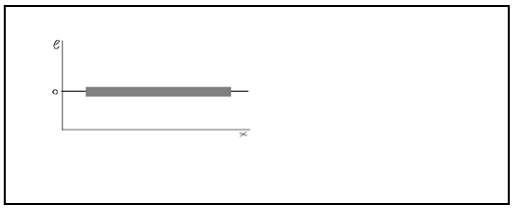


Fig. 1(a)

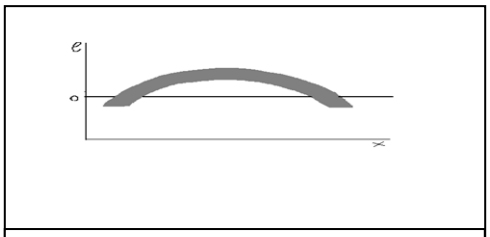


Fig. 1(b)

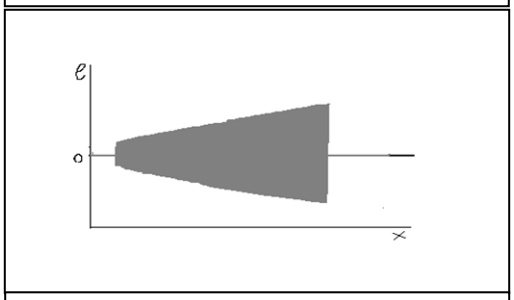


Fig. 1(c)

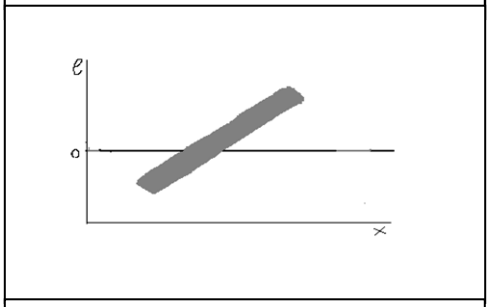


Fig. 1(d)

Non-constancy of Error Variance

Plots of residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. The prototype plot in Figure 1(a) exemplifies residual plots when error term variance is constant. Figure 1(c) shows a prototype picture of residual plot when the error variance increases with *X*. In many biological science applications, departures from constancy of the error variance tend to be of the "meghaphone" type.

Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from residual plots against *X* or $\hat{Y}$.

Nonindependence of Error Terms

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographical areas, it is good idea to prepare a sequence plot of the residuals. The purpose of plotting the residuals against time or some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence. A prototype residual plot showing a time related trend effect is presented in Figure 1(d), which portrays a linear time related trend effect. When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0.

Non-normality of Error Terms

Small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

Comparison of frequencies: when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 90% of the residuals fall between $\pm 1.645 \sqrt{MSE}$.

Normal probability plot: Still another possibility is to prepare a normal probability plot of the residuals. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The purpose of this additional analysis is to determine whether there are any key variables that could provide important additional descriptive and predictive power to the model. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

Statistical Tests for Model departures

Graphical analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residuals plots will frequently reveal difficulties with the model more clearly than particular formal tests.

Tests for Randomness

A run test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specially designed for lack of randomness in least squares residuals, is the

Durbin-Watson test:

The Durbin-Watson test assumes the first order autoregressive error models. The test consists of determining whether or not the autocorrelation coefficient (ρ , say) is zero. The usual test alternatives considered are:

$H_0 : \rho = 0$

$H_0 : \rho > 0$

The Durbin-Watson test statistic D is obtained by using ordinary least squares to fit the regression function, calculating the ordinary residuals: $e_t = Y_t - \hat{Y}_t$, and then calculating the statistic:

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{n \sum_{t=1} e_t^2} \qquad (1)$$

Exact critical values are difficult to obtain, but Durbin-Watson have obtained lower and upper bound d_L and d_U such that a value of D outside these bounds leads to a definite decision. The decision rule for testing between the alternatives is:

if $D > d_U$, conclude H_0

if $D < d_L$, conclude H_1

if $d_L \le D \le d_U$, test is inconclusive.

Small value of D lead to the conclusion that $\rho > 0$.

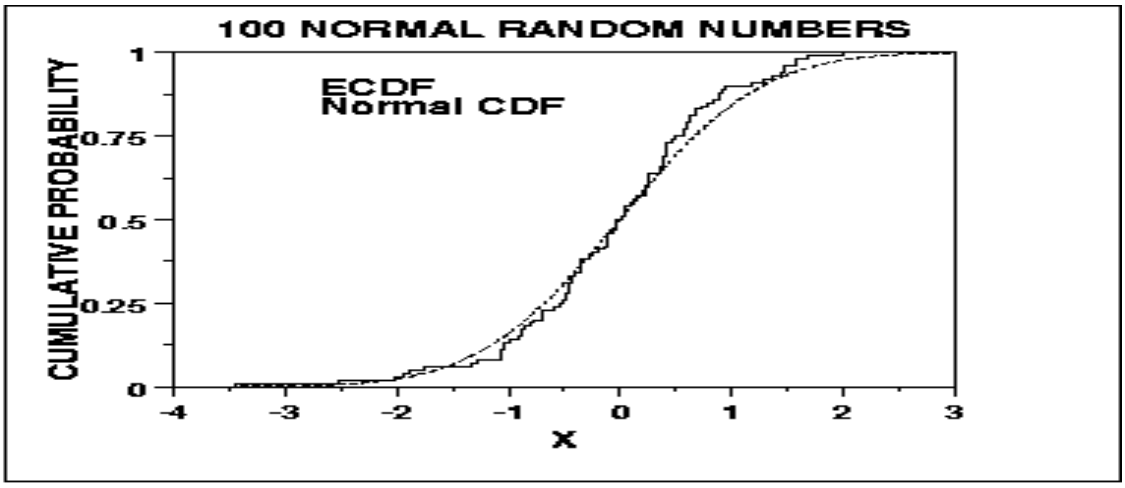
Tests for Normality

Correlation Test for Normality: In addition to visually assessing the appropriate linearity of the points plotted in a normal probability plot, a formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between residuals e_i and their expected values under normality. A high value of the correlation coefficient is indicative of normality.

Kolmogorov-Smirnov test: The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF). Given *N ordered* data points $Y_1, Y_2, ..., Y_N$, the ECDF is defined as

$E_N = n(i) / N$,

where $n(i)$ is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by 1/*N* at the value of each ordered data point. The graph below is a plot of the empirical distribution function with a normal cumulative distribution function for 100 normal random numbers. The K-S test is based on the maximum distance between these two curves.



An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the K-S test has several important drawbacks:

1. It only applies to continuous distributions.
2. It tends to be more sensitive near the center of the distribution than at the tails.
3. Perhaps the most serious limitation is that the distribution must be fully specified. That is, if location, scale, and shape parameters are estimated from the data, the critical region of the K-S test is no longer valid. It typically must be determined by simulation.

Due to limitations 2 and 3 above, many analysts prefer to use the **Anderson-Darling goodness-of-fit test.**

The Kolmogorov-Smirnov test is defined by:

H_0: The data follow a specified distribution

H_1: The data do not follow the specified distribution

The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right) \qquad (2)$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data). The hypothesis regarding the distributional form is rejected if the test statistic, ***D***, is greater than the critical value obtained from a table. There are several variations of these tables in the literature that use somewhat different scalings for the K-S test statistic and critical regions. These alternative formulations should be equivalent, but it is necessary to ensure that the test statistic is calculated in a way that is consistent with how the critical values were tabulated.

Anderson-Darling Test

The Anderson-Darling test is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Currently, tables of critical values are available for the normal, lognormal, exponential, Weibull, extreme value type I, and logistic distributions.
The Anderson-Darling test is defined as:

H_0: The data follow a specified distribution.
H_1: The data do not follow the specified distribution

The Anderson-Darling test statistic is defined as $A^2 = -N - S$,

where, $S = \sum_{i=1}^{N} \frac{(2i-1)}{N} [\ln F(Y_i) + \ln(1 - F(Y_{N+1-i})]$ (3)

F is the cumulative distribution function of the specified distribution. Note that the Y_i are the *ordered* data. The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas are available in literature for a few specific distributions (normal, lognormal, exponential, Weibull, logistic, extreme value type 1). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A, is greater than the critical value.

Tests for Constancy of Error Variance

Modified Levene Test : The test is based on the variability of the residuals. Let e_{i1} denotes the i^{th} residual for group 1 and e_{i2} denotes the i^{th} residual for group 2. Also we denote n_1 and n_2 to denote the sample sizes of the two groups, where: $n_1 + n_2 = n$.
Further, we shall use $\tilde{e}_1$ and $\tilde{e}_2$ to denote the medians of the residuals in the two groups. The modified Levene test uses the absolute deviations of the residuals around their median, to be denoted by d_{i1} and d_{i2}:

$$\bar{d}_{i1} = |e_{i1} - \tilde{e}_1|, \quad \bar{d}_{i2} = |e_{i2} - \tilde{e}_2|$$

With this notation, the two-sample t test statistic becomes:

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (4)$$

Where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 is:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2} .$$

If the error terms have constant variance and n_1 and n_2 are not too small, t_L^* follows approximately the t distribution with n-2 degrees of freedom. Large absolute values of t_L^* indicate that the error terms do not have constant variance.

White Test

In statistics, the White test is a statistical test that establishes whether the residual variance of a variable in a regression model is constant: that is for homoscedasticity. This test, and an estimator for heteroscedasticity-consistent standard errors, were proposed by Halbert White in 1980. These methods have become extremely widely used, making this paper one of the most cited articles in economics. To test for constant variance one undertakes an auxiliary regression analysis. This regresses the squared residuals from the original regression model onto a new set of regressors, which contains the original regressors, the cross-products of the regressors and the squared regressors. One then inspects the R^2. The LM test statistic is the product of the R^2 value and sample size:

$$LM = n.R^2 \qquad (5)$$

This follows a chi-square distribution, with degrees of freedom equal to the number of estimated parameters (in the auxiliary regression) minus one.

Tests for Outlying Observations

(i) *Elements of **Hat Matrix*** **:** The Hat matrix is defined as $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{X}$ is the matrix for explanatory variables. The larger values reflect data points are outliers.

(ii) ***WSSD$_i$***: $WSSD_i$ is an important statistic to locate points that are remote in *x*-space. $WSSD_i$ measures the weighted sum of squared distance of the i^{th} point from the center of the data. Generally if the $WSSD_i$ values progress smoothly from small to large, there are probably no extremely remote points. However, if there is a sudden jump in the magnitude of $WSSD_i$, this often indicates that one or more extreme points are present.

(iii) **Cook's D_i**

Cook's D_i is designed to measure the shift in $\hat{y}$ when i^{th} obsevation is not used in the estimation of parameters. D_i follows approximately $F_{(p,n-p-1)}(1\text{-}\alpha)$. Lower 10% point of this distribution

is taken as a reasonable cut off (more conservative users suggest the 50% point). The cut off for D_i can be taken as $\frac{4}{n}$.

(iv) **$DFFITS_i$**

DFFIT is used to measure difference in i[th] component of $\left(\hat{y} - \hat{y}_{(i)}\right)$. It is suggested that $DFFITS_i \geq 2\left(\frac{p+1}{n}\right)^{1/2}$ may be used to flag off influential observations.

(v) ***DFBETAS*** $_{j(i)}$

Cook's D_i reveals the impact of i[th] observation on the entire vector of the estimated regression coefficients. The influential observations for individual regression coefficient are identified by $DFBETAS_{j(i)}, j = 1,2,...,p+1$, where each $DFBETAS_{j(i)}$ is the standardized change in b_j when the i[th] observation is deleted.

(vi) ***COVRATIO*** $_i$

The impact of the i[th] observation on variance-covariance matrix of the estimated regression coefficients is measured by the ratio of the determinants of the two variance-covariance matrices. Thus, COVRATIO reflects the impact of the i[th] observation on the precision of the estimates of the regression coefficients. Values near 1 indicate that the i[th] observation has little effect on the precision of the estimates. A value of COVRATIO greater than 1 indicates that the deletion of the i[th] observation decreases the precision of the estimates; a ratio less than 1 indicates that the deletion of the observation increases the precision of the estimates. Influential points are indicated by $\left|COVRATIO_i - 1\right| > \frac{3(p+1)}{n}$.

(vii) ***FVARATIO*** $_i$ ***:***

The statistic detects change in variance of $\hat{y}_i$ when an observation is deleted. A value near 1 indicates that the i[th] observation has negligible effect on variance of y_i. A value greater than 1 indicates that deletion of the i[th] observation decreases the precision of the estimates, a value less than one increases the precision of the estimates.

Tests for Multicollinearity

The use and interpretation of a multiple regression model depends implicitly on the assumption that the explanatory variables are not strongly interrelated. In most regression applications the explanatory variables are not orthogonal. Usually the lack of orthogonality is not serious enough to affect the analysis. However, in some situations the explanatory variables are so strongly

interrelated that the regression results are ambiguous. Typically, it is impossible to estimate the unique effects of individual variables in the regression equation. The estimated values of the coefficients are very sensitive to slight changes in the data and to the addition or deletion of variables in the equation. The regression coefficients have large sampling errors which affect both inference and forecasting that is based on the regression model. The condition of severe non-orthogonality is also referred to as the problem of multicollinearity.

The presence of multicollinearity has a number of potentially serious effects on the least squares estimates of regression coefficients as mentioned above. Some of the effects may be easily demonstrated. Multicollinearity also tends to produce least squares estimates b_j that are too large in absolute value.

Detection of Multicollinearity

Let $R = (r_{ij})$ and $R^{-1} = (r^{ij})$ denote simple correlation matrix and its inverse. Let $\lambda_i, i = 1,2,...,p$ $(\lambda_p \leq \lambda_{p-1} \leq ... \lambda_1)$ denote the eigen values of *R*. The following are common indicators of relationships among independent variables.

1. Simple pair-wise correlations $|r_{ij}| = 1$
2. The squared multiple correlation coefficients

$R_i^2 = 1 - \frac{1}{r^{ii}} > 0.9$, where R_i^2 denote the squared multiple correlation coefficients for the regression of x_1 on the remaining *x* variables.

3. The variance inflation factors, $VIF_i = r^{ii} > 10$ and
4. eigen values, $\lambda_i = 0$.

The first of these indicators, the simple correlation coefficients between pairs of independent variables r_{ij}, may detect a simple relationship between x_i and x_j. Thus $|r_{ij}| = 1$ implies that the ith and jth variables are nearly proportional.

The second set of indicators, R_i^2, the squared multiple correlation coefficient for the regression of x_i on the remaining *x* variables indicates the degree to which x_i is explained by a linear combination of all of the other input variables.

The third set of indicators, the diagonal elements of the inverse matrix, which have been labeled as the Variance Inflation Factors, VIF_i. The term arises by noting that with standardized data (mean zero and unit sum of squares), the variance of the least squares estimate of the ith

coefficient is proportional to r^{ii}, $VIF_i > 10$ is probably based on the simple relation between R_i and VIF_i. That is $VIF_i > 10$ corresponds to $R_i^2 > 0.9$.

Overview of Remedial Measures

If the simple regression model (1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model and develop and use a more appropriate model.

2. Employ some transformation on the data so that regression model (1) is appropriate for the transformed data.

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedure for estimating the parameters. Successful use of transformations, on the other hand, leads to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformation may obscure the fundamental interconnections between the variables, though at times they may illuminate them.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model (1) by altering the nature of the regression function. For instance, a quadratic regression function might be used.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

or an exponential regression function:

$$Y_i = \gamma_0 \gamma_1^{X_i} + \varepsilon_i .$$

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful.

Non-constancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the method to allow for this and use the method of weighted least squares to obtain the estimates of the parameters. Transformations are another way in stabilizing the variance. We first consider transformation for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformation on X should be attempted. The

reason why transformation on *Y* may not be desirable here is that a transformation on *Y*, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution and may lead to substantially differing error term variance.

Following transformations are generally applied for stabilizing variance.

(1) When the error variance is rapidly increasing $Y' = \log_{10} Y$ or $Y' = \sqrt{Y}$

(2) When the error variance is slowly increasing, $Y' = Y^2$ or $Y' = Exp(Y)$

(3) When the error variance is decreasing, $Y' = 1/Y$ or $Y' = Exp(-Y)$.

Box - Cox Transformations

It is difficult to determine, which transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variance, and nonlinearity of the regression function. The Box-Cox transformation automatically identifies a transformation from the family of power transformations on Y. The family of power transformations is of the form: $Y' = Y^{\lambda}$, where is a parameter to be determined from the data. Using standard computer programme it can be determined easily.

Non-independence of Error Terms

When the error terms are correlated, a direct approach is to work with a model that calls for error terms. A simple remedial transformation that is often helpful is to work with first differences.

Non-normality of Error terms

Lack of normality and non-constant error variance frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore, desirable that the transformation for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present.

Omission of Important Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model.

Outlying Observations

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents in error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstances. When outlying observations are present, use of the least squares and maximum likelihood estimates for regression model (1) may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. Robust Regression falls under such methods.

Multicollinearity

i) **Collection of additional data:** Collecting additional data has been suggested as one of the methods of combating multicollinearity. The additional data should be collected in a manner designed to break up the multicollinearity in the existing data.

ii) **Model respecification:** Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to respecification is to redefine the regressors. For example, if x_1, x_2 and x_3 are nearly linearly dependent it may be possible to find some function such as $x = (x_1+x_2)/x_3$ or $x = x_1x_2x_3$ that preserves the information content in the original regressors but reduces the multicollinearity.

iii) **Ridge Regression**: When method of least squares is used, parameter estimates are unbiased. A number of procedures have been developed for obtaining biased estimators of regression coefficients to tackle the problem of multicollinearity. One of these procedures is ridge regression. The ridge estimators are found by solving a slightly modified version of the normal equations. Each of the diagonal elements of $\mathbf{X'X}$ matrix are added a small quantity.

Example

Table 1

Case	X_{11}	X_{21}	X_{31}	Y_i
1	12.980	0.317	9.998	57.702
2	14.295	2.028	6.776	59.296
3	15.531	5.305	2.947	56.166
4	15.133	4.738	4.201	55.767
5	15.342	7.038	2.053	51.722
6	17.149	5.982	-0.055	60.446
7	15.462	2.737	4.657	60.715
8	12.801	10.663	3.048	37.447
9	17.039	5.132	0.257	60.974
10	13.172	2.039	8.738	55.270
11	16.125	2.271	2.101	59.289
12	14.340	4.077	5.545	54.027
13	12.923	2.643	9.331	53.199
14	14.231	10.401	1.041	41.896
15	15.222	1.220	6.149	63.264
16	15.740	10.612	-1.691	45.798
17	14.958	4.815	4.111	58.699
18	14.125	3.153	8.453	50.086
19	16.391	9.698	-1.714	48.890
20	16.452	3.912	2.145	62.213
21	13.535	7.625	3.851	45.625
22	14.199	4.474	5.112	53.923
23	15.837	5.753	2.087	55.799
24	16.565	8.546	8.974	56.741
25	13.322	8.589	4.011	43.145
26	15.949	8.290	-0.248	50.706

Table 2: Indicators of Influential Observations

Case	r_i	t_i	$t_I^*=s.t/s_i$	h_{ii}	D_i	$WSSD_i$
1	0.460	0.289	0.281	0.215	0.005	39*
2	1.253	0.732	0.724	0.093	0.013	12
3	0.377	0.215	0.210	0.048	0.001	1
4	0.044	0.025	0.026	0.042	0.000	1

5	-0.256	-0.146	-0.141	0.053	0.000	3
6	1.010	0.611	0.602	0.155	0.017	20
7	0.389	0.226	0.221	0.081	0.001	7
8	0.132	0.088	0.086	0.301	0.001	41
9	0.432	0.262	0.256	0.155	0.003	18
10	0.589	0.355	0.347	0.147	0.005	23
11	-3.302	-2.021	-2.193	0.173	0.214	14
12	-0.406	-0.232	-0.226	0.053	0.001	3
13	0.194	0.118	0.117	0.163	0.001	24
14	-0.268	-0.164	-0.161	0.175	0.001	23
15	0.802	0.476	0.469	0.122	0.007	15
16	-0.482	-0.295	-0.289	0.177	0.005	26
17	3.756	2.134	2.343	0.041	0.048	0
18	-6.072	-3.589	-5.436	0.114	0.412	8
19	-1.198	-0.727	-0.719	0.160	0.025	24
20	1.126	0.666	0.658	0.114	0.014	11
21	0.449	0.266	0.259	0.119	0.003	12
22	0.791	0.453	0.444	0.055	0.003	3
23	-0.060	-0.035	-0.032	0.059	0.000	3
24	0.574	1.181	1.188	0.927	4.409	19
25	0.268	0.163	0.158	0.159	0.001	19
26	-0.606	-0.356	-0.350	0.101	0.004	11

Table 3: Indicators of Influential Observations

Case	Cov Ratio	Dffits	Intercept	X1	X2	X3
				DFBETAS		
1	1.512	0.148	0.056	-0.053	-0.006	0.006
2	1.203	0.232	0.062	-0.042	-0.042	-0.050
3	1.254	0.047	-0.005	0.010	-0.008	-0.007
4	1.257	0.005	0.000	0.000	-0.001	0.000
5	1.267	-0.033	-0.001	-0.001	-0.006	0.006
6	1.331	0.258	-0.095	0.132	-0.042	-0.050
7	1.299	0.068	-0.005	0.015	-0.036	-0.005
8	1.721	0.057	0.027	-0.034	0.026	-0.006
9	1.408	0.109	-0.030	0.048	-0.035	-0.031
10	1.380	0.144	0.058	-0.058	-0.041	0.016
11	0.639	-1.004	-0.154	-0.045	0.776	0.525

12	1.260	-0.054	-0.017	0.014	0.014	0.000
13	1.435	0.051	0.017	-0.19	-0.004	0.013
14	1.452	-0.074	-0.026	0.031	-0.35	0.015
15	1.315	0.175	-0.008	0.033	-0.105	0.002
16	1.441	-0.134	-0.014	0.014	-0.044	0.047
17	0.496	0.482	0.061	-0.17	-0.107	-0.046
18	0.410	-1.945	0.362	-0.308	-0.220	-1.177
19	1.301	-0.341	0.031	-0.045	-0.080	0.094
20	1.252	0.236	-0.055	0.097	-0.105	-0.051
21	1.350	0.095	0.054	-0.061	0.024	-0.018
22	1.228	0.108	0.052	-0.048	-0.028	-0.020
23	1.279	-0.008	0.001	-0.002	0.001	0.002
24	12.715	4.230	-3.642	3.276	3.180	3.934
25	1.426	0.069	0.031	-0.039	0.029	-0.003
26	1.309	-0.117	0.000	-0.007	-0.016	0.043

Table 4: Regression Coefficients and Summary Statistics

Description	**b_0**	**b_1**	**b_2**	**b_3**	**s**	**R^2**	**Max VIF**	**Min e.v.**	**Max R_i^2**
All Data (n=26)	8.11	3.56	-1.63	0.34	1.80	0.94	2.82	0.210	0.65
Delete (11, 17, 18)	7.17	3.66	-1.79	0.40	0.51	0.99	2.85	0.210	0.65
Delete (24)	30.91	2.39	-2.14	-0.36	1.78	0.94	30.64	0.017	0.97
Delete (11, 17, 18, 24)	24.27	2.79	-2.11	-0.16	0.50	0.99	171.90	0.003	0.99
Ridge k=0.05 (n=22)	14.28	3.22	-1.73	0.25	0.66	0.99	10.20	0.053	0.90
Delete X3 (n=22)	19.50	3.03	-2.00		0.49	0.99	1.02	0.863	0.02

REFERENCES

- Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). Regression diagnostics – Identifying influential data and sources of collinearity, New York.: Wiley
- Barnett, V. and Lewis, T. (1984). Outliers in Statistical Data, New York: Wiley Ltd.
- Chatterjee, S. and Price, B (1977). Regression analysis by example, New York: John Wiley & sons
- Draper, N.R. and Smith, H. (1998). Applied Regression analysis, New York: Wiley Eastern Ltd.
- Kleinbaum, D.G. & Kupper, L.L. (1978). Applied Regression analysis and other multivariate methods, Massachusetts: Duxbury Press
- Montgomery, D.C., Peck, E. and Vining, G. (2003). Introduction to linear regression analysis, 3rd Edition, New York: John Wiley and Sons Inc.

LINEAR TIME SERIES MODELLING

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

Introduction

A data set containing observations on a single phenomenon observed over multiple time periods is called 'time-series'. In time-series data, both the *values* and the *ordering* of the data points have meaning. For many agricultural products, data are usually collected over time. Analysis of time series has been a part of statistics for long. Some methods have also been developed for its analysis to suit the distinct features of time series data, which differ both from cross section and panel or pooled data. Various approaches are available for time series modeling. Some of the tools and models which can be used for time series analysis, modeling and forecasting are briefly discussed. Various statistical approaches viz. regression, time series, stochastic and, of late, machine learning approaches are in vogue for statistical modeling. However, the same cannot be claimed to be complete and exhaustive. Every approach has its own advantages and limitations. These models typically utilize a host of empirical data and attempt to forecast market behavior and estimate future values of key variables by using past values of core economic indicators.

Forecasting plays a crucial role in business, Industry, government and institutional planning because many important decisions depend on the anticipated future values of certain variables. Forecast can be made in many different ways, the choice of the method depending on the purpose and importance of the forecasts as well as the costs of alternative methods. The most widely used technique for analysis of time-series data is; undoubtedly, the Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology. In this presentation, we shall talk about 'Univariate' Box-Jenkins models, also referred to as ARIMA models. Univariate or single series means that forecasts are based only on past values of the variable being forecast, they are not based on any other data series. The time-series data refer to observations on a variable that occur in a time sequence. One characteristic of such data is that the successive observations are dependent. Each observation of the observed data series, Y_t, may be considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t, t \in T\}$, where $T = \{0, \pm 1, \pm 2, \ldots\}$, and apply standard time-series approach to develop an ideal model which will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner.

We denote by Y_t, the observation made at time t (t = 1, 2, …. , n). Thus, a time-series involving n points may be represented as sequence of n observations Y_1, Y_2, ... , Y_n. The statistical analysis of time series data differs from the classical regression analysis. Time series

data typically violates the assumption that the error terms/successive observations are uncorrelated with each other. This effect, known as autocorrelation, biases the standard error associated with regression slope parameters estimates and makes the relevant t-test invalid. Contrary to statistically independence of observations, in the Box-Jenkins method, we suppose that the time sequenced observations (Y_1, Y_2,. . . . ., Y_{t-1}, Y_t, Y_{t+1}, . . .) may be statistically related to others in the same series. Our goal is to find a good way of stating that statistical relationship. That is, we want to find a good model that describes how the observations in a single time-series are related to each other.

The Box-Jenkins models are especially suited to short term forecasting because most ARIMA models place greater emphasis on the recent past rather than the distant past. The Box-Jenkins method applies to both discrete data as well as to continuous data. However, the data should be available at equally spaced discrete time intervals. Also, building of a ARIMA model requires a minimum of about 40-50 observations.

Time series models and components

Time series (TS) data refers to observations on a variable that occurs in a time sequence. Mostly these observations are collected at equally spaced, discrete time intervals. The TS movements of such chronological data can be decomposed into trend, periodic (say, seasonal), cyclical and irregular variations. One or two of these components may overshadow the others in some series. A basic assumption in any TS analysis/modeling is that some aspects of the past pattern will continue to remain in the future.

TS models have advantages over other statistical models in certain situations. They can be used more easily for forecasting purposes because historical sequences of observations upon study variables are readily available from published secondary sources. These successive observations are statistically dependent and TS modeling is concerned with techniques for the analysis of such dependencies. Thus in TS modeling, the prediction of values for the future periods is based on the pattern of past values of the variable under study, but not generally on explanatory variables which may affect the system. There are two main reasons for resorting to such TS models. First, the system may not be understood, and even if it is understood it may be extremely difficult to measure the cause and effect relationship, second, the main concern may be only to predict what will happen and not to know why it happens. Many a time, collection of information on causal factors (explanatory variables) affecting the study variable(s) may be cumbersome /impossible and hence availability of long series data on explanatory variables is a problem. In such situations, the TS models are a boon for forecasters. Hence, if TS models are put to use, say, for instance, for forecasting purposes, then they are especially applicable only in the ‘short term’.

A detailed discussion regarding various TS components has been done by Croxton *et al.* (1979). A good account on exponential smoothing methods is given in Makridakis *et al.* (1998). A practical treatment on ARIMA modeling along with several case studies can be found in Pankratz

(1983). A reference book on ARIMA and related topics with a more rigorous theoretical flavour is by Box *et al.* (1994). Paul (2010), Paul and Das (2010, 2013), Paul *et al*. (2013, 2014) applied ARIMA model in the field of agriculture as well as livestock's and fisheries.

An important step in analyzing TS data is to consider the types of data patterns, so that the models most appropriate to those patterns can be utilized. Four types of TS components can be distinguished.

They are

I. **Horizontal** – when data values fluctuate around a constant value
II. **Trend** – when there is long term increase or decrease in the data
III. **Seasonal** – when a series is influenced by seasonal factor and recurs on a regular periodic basis
IV. **Cyclical** – when the data exhibit rises and falls that are not of a fixed period

Note that many data series include combinations of the preceding patterns. After separating out the existing patterns in any TS data, the pattern that remains unidentifiable form the 'random' or 'error' component. Time plot (data plotted over time) and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data.

Trend analysis of TS data is usually done to analyse a variable over time to detect or investigate long-term changes. Trend is 'long-term' behaviour of a TS process usually in relation to the mean level. The trend of TS may be studied because the interest lies in the trend itself, or may be to eliminate the trend statistically in order to have insight into other components such as periodic variations in the series. A periodic movement is one which recurs with some degree of regularity, within a definite period. The most frequently studied periodic movement is that which occurs within a year and which is known as seasonal variation. Sometimes the TS data are de-seasonalized for the purpose of making the other movements (particularly trend) more readily discernible. Climatic conditions directly affect the production system in agriculture and hence in turn their patterns of prices and thus are primarily responsible for most of the seasonal variations exhibited in such series.

Over a period of time, a TS is very likely to show a tendency to increase or to decrease otherwise termed as an upward or downward trend respectively. One should not lose sight of the underlying factors that sometimes may cause such trend like growth in population, price changes etc. Technological developments and adoption patterns have been affecting agriculture so as to increase output enormously. Not always keeping pace with them, but induced by them, have been changes in the main variable (read here price of agricultural commodities) under study. Not all historical series show upward trends. Some, say, plant disease incidence, exhibit a generally

downward trend. This particular declining trend is attributable to better and more widely available advisory and extension services or due to good government policies. An economic series may have a downward trend because a better or cheaper substitute may be available.

Many techniques such as time plots, auto-correlation functions, box plots and scatter plots abound for suggesting relationships with possibly influential factors. For long and erratic series, time plots may not be helpful. Alternatives could be to go for smoothing or averaging methods like moving averages exponential smoothing methods etc. In fact, if the data contains considerable error, then the first step in the process of trend identification is smoothing.

Stationarity of a TS process

A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time. Thus, if we consider different subsets of a realization (TS 'sample') the different subsets will typically have means, variances and autocorrelation functions that do not differ significantly.
A statistical test for stationarity or test for unit root has been proposed by Dickey and Fuller (1979). The test is applied for the parameter ρ in the auxiliary regression

$$\Delta_1 y_t = \rho\, y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t$$

where Δ_1 denotes the differencing operator i.e. $\Delta_1 y_t = y_t - y_{t-1}$.

The relevant null hypothesis is $\rho = 0$ i.e. the original series is non-stationary and the alternative is $\rho < 0$ i.e. the original series is stationary. Usually, differencing is applied until the acf shows an interpretable pattern with only a few significant autocorrelations.

Autocorrelation functions

Autocorrelation

Autocorrelation refers to the way the observations in a TS are related to each other and is measured by the simple correlation between current observation (Y_t) and observation from p periods before the current one (Y_{t-p}). That is for a given series Y_t, autocorrelation at *lag p* is the correlation between the pair (Y_t , Y_{t-p}) and is given by

$$r_p = \frac{\sum_{t=1}^{n-p}\left(Y_t - \overline{Y}\right)\left(Y_{t+p} - \overline{Y}\right)}{\sum_{t=1}^{n}\left(Y_t - \overline{Y}\right)^2}$$

It ranges from –1 to +1. Box and Jenkins has suggested that maximum number of useful r_p are roughly N/4 where N is the number of periods upon which information on y_t is available.

Partial autocorrelation

Partial autocorrelations are used to measure the degree of association between y_t and y_{t-p} when the y-effects at other time lags 1,2,3,…,p-1 are removed.

Autocorrelation function (ACF) and partial autocorrelation function (PACF)

Theoretical ACFs and PACFs (Autocorrelations versus lags) are available for the various models chosen (say, see Pankratz, 1983) for various values of orders of autoregressive and moving average components i.e. p and q. Thus compare the correlograms (plot of sample ACFs versus lags) obtained from the given TS data with these theoretical ACF/PACFs, to find a reasonably good match and tentatively select one or more ARIMA models. The general characteristics of theoretical ACFs and PACFs are as follows:- (here 'spike' represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

Model	**ACF**	**PACF**
AR	Spikes decay towards zero	Spikes cutoff to zero
MA	Spikes cutoff to zero	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

Description of ARIMA models

Autoregressive (AR) Model

A stochastic model that can be extremely useful in the representation of certain practically occurring series is the autoregressive model. In this model, the current value of the process is expressed as a finite, linear aggregate of previous values of the process and a shock ε_t. Let us denote the values of a process at equally spaced time epochs *t*, *t-1*, *t-2*, . . . by y_t, y_{t-1}, y_{t-2}, . . ., then y_t can be described by the following expression:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

If we define an autoregressive operator of order *p* by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p,$$

where *B* is the backshift operator such that $By_t = y_{t-1}$, the autoregressive model can be written as $\varphi(B) y_t = \varepsilon_t$.

Moving Average (MA) Model

Another kind of model of great practical importance in the representation of observed time-series is the finite moving average process. MA (q) model is defined as

$$y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}.$$

If we define a moving average operator of order q by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q,$$

where B is the backshift operator such that $By_t = y_{t-1}$, the moving average model can be written as $y_t = \theta(B)\varepsilon_t$.

Autoregressive Moving Average (ARMA) Model

To achieve greater flexibility in fitting of actual time-series data, it is sometimes advantageous to include both autoregressive and moving average processes. This leads to the mixed autoregressive-moving average model

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}$$

or

$$\varphi(B)y_t = \theta(B)\varepsilon_t.$$

This is written as ARMA(p, q) model. In practice, it is frequently true that adequate representation of actually occurring stationary time-series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than *2* and often less than *2*.

Autoregressive Integrated Moving Average (ARIMA) Model

A generalization of ARMA models which incorporates a wide class of non- stationary time-series is obtained by introducing the differencing into the model. The simplest example of a non-stationary process which reduces to a stationary one after differencing is Random Walk. A process $\{y_t\}$ is said to follow an Integrated ARMA model, denoted by ARIMA (p, d, q), if $\nabla^d y_t = (1-B)^d \varepsilon_t$ is ARMA (p, q). The model is written as

$$\varphi(B)(1-B)^d y_t = \theta(B)\varepsilon_t$$

where $\varepsilon_t \sim WN(0, \sigma^2)$, *WN* indicating White Noise. The integration parameter d is a nonnegative integer. When $d = 0$, ARIMA (p, d, q) ≡ ARMA (p, q).

The ARIMA methodology is carried out in three stages, viz. identification, estimation and diagnostic checking. Parameters of the tentatively selected ARIMA model at the identification stage are estimated at the estimation stage and adequacy of tentatively selected model is tested at the diagnostic checking stage. If the model is found to be inadequate, the three stages are repeated until satisfactory ARIMA model is selected for the time-series under consideration. An

excellent discussion of various aspects of this approach is given in Box *et al*. (2007). Most of the standard software packages, like SAS, SPSS, R and EViews contain programs for fitting of ARIMA models.

Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The fundamental fact about seasonal time-series with period *S* is that observations, which are *S* intervals apart, are similar. Therefore, the operation $L(y_t) = y_{t-1}$ plays a particularly important role in the analysis of seasonal time-series. In general, the order of SARIMA model is denoted by $(p, d, q) \times (P, D, Q)_S$, and the model is represented as follows:

$$\phi_p(L)\Phi_P(L^S)\nabla^d\nabla_S^D y_t = \theta_q(L)\Theta_Q(L^S)\varepsilon_t$$

where $\phi_p(L)$, $\theta_q(L)$ are polynomials in L of degrees p and q respectively and $\Phi_P(L^S)$, $\Theta_Q(L^S)$ are polynomials in L^S of degrees P and Q respectively. For estimation of parameters, iterative least squares method is used.

Model building

Identification

The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for stationary series. If the original series is non stationary then first of all it should be made stationary.

The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters, p, q, and P, Q. They could be obtained by looking for significant autocorrelation and partial autocorrelation coefficients (see section 5 (iii)). Say, if second order auto correlation coefficient is significant, then an AR (2), or MA (2) or ARMA (2) model could be tried to start with. This is not a hard and fast rule, as sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients. Still they can be used as initial values while the final models are achieved after going through the stages repeatedly. Note that usually up to order 2 for p, d, or q are sufficient for developing a good model in practice.

Estimation

At the identification stage one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then we attempt to obtained precise estimates of parameters of the model by least squares as advocated by Box and Jenkins. Standard computer packages like SAS, SPSS etc. are available for finding the estimates of relevant

parameters using iterative procedures. The methods of estimation are not discussed here for brevity.

Diagnostics

Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained with following diagnostics.

Low Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC)/ Schwarz-Bayesian Information Criteria (SBC)

AIC is given by (-2 log L + 2 m) where m=p+ q+ P+ Q and L is the likelihood function. Since -2 log L is approximately equal to $\{n (1+\log 2\pi) + n \log \sigma^2\}$ where σ^2 is the model MSE, Thus AIC can be written as AIC=$\{n (1+\log 2\pi) + n \log \sigma^2 + 2 m\}$
and because first term in this equation is a constant, it is usually omitted while comparing between models. As an alternative to AIC, sometimes SBC is also used which is given by SBC = $\log \sigma^2 + (m \log n) /n$.

Plot of residual ACF

Once the appropriate ARIMA model has been fitted, one can examine the goodness of fit by means of plotting the ACF of residuals of the fitted model. If most of the sample autocorrelation coefficients of the residuals are within the limits $\pm 1.96 / \sqrt{N}$ where N is the number of observations upon which the model is based then the residuals are white noise indicating that the model is a good fit.

Non-significance of auto correlations of residuals via Portmonteau tests (Q-tests based on Chisquare statistics)-Box-Pierce or Ljung-Box texts

After tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and, if need be, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity Q known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and is computed as follows:

$$Q = n \Sigma r^2 (j)$$

where summation extends from 1 to k with k as the maximum lag considered, n is the number of observations in the series, r (j) is the estimated autocorrelation at lag j; k can be any positive integer and is usually around 20. Q follows Chi-square with (k-m1) degrees of freedom where

m1 is the number of parameters estimated in the model. A modified Q statistic is the Ljung-box statistic which is given by

$$Q = n(n+2) \Sigma r^2 (j) / (n-j)$$

The Q Statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and Q should be small (the probability value should be large). A significant value indicates that the chosen model does not fit well.

All these stages require considerable care and work and they themselves are not exhaustive.

An Illustration

All-India data of inland fish production during the period 1951 to 2008 are obtained from handbook of fishery, ministry of agriculture, Govt. of India and the website www.indiastat.com and the same are exhibited in Fig. 1. From the total 56 data points, first 50 data points corresponding to the period 1951 to 2000 are used for building the model and remaining are used for validation purpose. A perusal of the data shows that, there is a linear trend in the inland fish production.

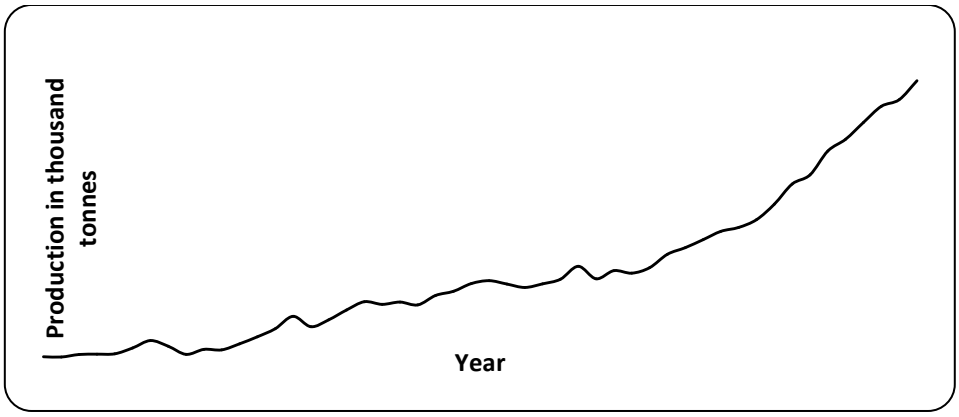


Fig. 1 Inland fish production

Fitting of ARIMA Model

From the estimated autocorrelation function (ACF), reported in Table 1, it is found that it decays very slowly thereby requires to be differenced so that the resulting series depicts a pattern for a possible ARMA modelling. Further, in this situation it becomes difficult for selection of order of ARIMA model. The test for unit root proposed by Dickey and Fuller (1979) is applied for the parameter ρ in the auxiliary regression

$$\Delta_1 y_t = \rho\, y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t$$

The relevant null hypothesis is $\rho = 0$ and the alternative is $\rho < 0$. In the present situation the estimate of ρ is 0.061 with calculated *t*-statistic is 5.55 which is greater than the critical value of *t* at 5% level of significance i.e. -1.95 (Franses, 1998) resulting the acceptance of null hypothesis. Thus there is presence of unit root and so differencing is required. Usually, differencing is

applied until the ACF shows an interpretable pattern with only a few significant autocorrelations. On taking the second difference of the original series, it is seen that only a few ACFs, reported in Table 1, are high making it easier to select the order of the model.

Table 1. Sample autocorrelation functions (ACF) and partial autocorrelation functions (PACF) of the original and differenced series

Lag	ACF of the series	PACF of the series	ACF of the differenced series	PACF of the differenced series	ACF of the double differenced series	PACF of the double differenced series
1	0.912	0.912	0.227	0.227	-0.564	-0.564
2	0.829	-0.013	0.346	0.31	0.131	-0.275
3	0.743	-0.065	0.221	0.112	-0.077	-0.219
4	0.661	-0.026	0.203	0.058	-0.075	-0.336
5	0.584	-0.018	0.328	0.231	0.106	-0.243
6	0.509	-0.035	0.242	0.107	0	-0.134
7	0.447	0.024	0.196	-0.017	-0.048	-0.194
8	0.383	-0.048	0.157	-0.02	0.182	0.122
9	0.325	-0.013	-0.101	-0.299	-0.203	0.081
10	0.277	0.018	-0.037	-0.206	0.038	-0.013

The appropriate model is chosen on the basis of minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. Using eqs.(3) and (4), the AIC and BIC values are respectively computed and listed in table 2. A perusal of table 2 shows that the AIC and BIC values are minimum for ARIMA (1,2,2) but the corresponding values for ARIMA (1,2,1) model do not differ much from that of ARIMA(1,2,2). As because ARIMA (1,2,1) is more parsimonious than ARIMA (1,2,2), the ARIMA(*1*,2,*1*) model is selected for modelling and forecasting of the inland fish production in India. The estimates of parameters of above model are reported in Table 3.

Table 2. AIC and BIC values for different ARIMA models

Criteria	**ARIMA (1,2,0)**	**ARIMA (1,2,1)**	**ARIMA (2,2,0)**	**ARIMA (2,2,1)**	**ARIMA (2,2,2)**	**ARIMA (1,2,2)**
AIC	425.87	413.22	422.97	414.32	412.93	414.27
BIC	443.34	430.69	440.44	431.79	430.41	431.75

Table 3. Estimates of parameters along with their SE for fitted ARIMA (1,*2,1)* model

Parameter	Estimate	Standard error
AR1	-0.141	0.171
MA1	0.823	0.107
Constant	2.623	1.469

The graph of fitted model along with data points is exhibited in Fig. 2. A perusal of fig. 2 indicates that the fitted ARIMA(*1,2,1)* model is able to capture the trend present in the inland fish production in India very well.

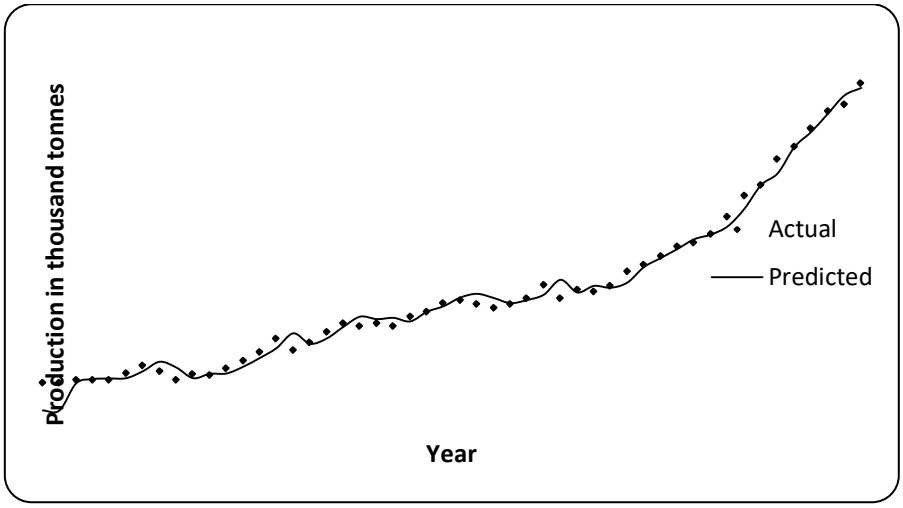


Fig.2 Fitted ARIMA (1, 2, 1) model along with the data points

One-step ahead forecasts of inland fish production along with their corresponding standard errors, upper confidence interval and lower confidence interval for the year, 2001 to 2008 in respect of above fitted model are reported in Table 4. The attractive feature for fitted ARIMA model is that all the forecast values except for 2008, lie within one standard error of forecasts.

Table 4. Forecasts of inland fish production (in tonnes) for fitted models

Years	Actual	Forecasts by ARIMA(*1*,2,*1*)	SE of Forecast	Lower Confidence Limit	Upper Confidence Limit
2001	2823.0	2727.09	59.037	2609.3	2844.87
2002	2845.0	2860.68	84.010	2691.15	3030.2
2003	3126.0	2996.05	108.299	2774.81	3217.3
2004	3210.0	3134.17	132.228	2861.08	3407.27
2005	3458.0	3274.9	156.400	2948.71	3601.09
2006	3525.0	3418.25	181.033	3037.4	3799.11
2007	3755.0	3564.23	206.247	3126.96	4001.5
2008	4200.0	3712.83	232.103	3217.32	4208.33

The out of sample forecast of inland fish production in India for the year 2009 and 2010 have been found out as 4360 and 4610 thousand tonnes. For measuring the accuracy in fitted time series model, Mean absolute error (MAE), Mean absolute percentage error (MAPE) and Relative mean absolute prediction error (RMAPE) are computed by using the formulae given in eqs. 5, 6 and 7. The MAE, MAPE and RMAPE values for fitted ARIMA(1,2,1) model are respectively computed as 160.64, 0.044 and 4.43.

$$\text{MAE} = 1/8\sum_{i=1}^{8}\left|y_{t+i} - \hat{y}_{t+i}\right| \tag{5}$$

$$\text{MAPE} = 1/8\sum_{i=1}^{8}\left\{\left|y_{t+i} - \hat{y}_{t+i}\right| / y_{t+i}\right\} \tag{6}$$

$$\text{RMAPE} = 1/8\sum_{i=1}^{8}\left\{\left|y_{t+i} - \hat{y}_{t+i}\right| / y_{t+i}\right\}\times 100 \tag{7}$$

Conclusion

The ARIMA models being stochastic in nature emphasized variations in data using empirically based methods to determine the proper form of the model that is best suited for short-term forecasting. The more realistic forecast intervals for India's inland fish production data obtained through ARIMA approach could be of immense help to planners in formulating appropriate strategies. These in turn would also benefit the farmers in production of optimum quantities of fish. All this would ultimately result in efficient management of India's inland fish production scenario through sound statistical technique.

REFERENCES

- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time series analysis: Forecasting and control*, Pearson Education, Delhi.
- Croxton, F.E., Cowden, D.J. and Klein, S. (1979). Applied General Statistics, New Delhi: Prentice Hall of India Pvt. Ltd.
- Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). *Forecasting Methods and Applications*, 3rd Edition, John Wiley, New York.
- Pankratz, A. (1983). Forecasting with univariate Box – Jenkins models: concepts and cases, John Wiley, New York.
- Paul, R. K. and Das, M. K. (2010). Statistical modelling of inland fish production in India. *Journal of the Inland Fisheries Society of India*, 42, 1-7.
- Paul, R. K. (2010). Stochastic Modeling of Wholesale Price of Rohu in West Bengal, India. *Interstat*.
- Paul, R. K. and Das, M. K. (2013). Forecasting of average annual fish landing in Ganga Basin. *Fishing chimes*, 33 (3), 51-54
- Paul, R. K., Panwar, S., Sarkar, S. K., Kumar, A. Singh, K. N., Farooqi, S. and Chaudhary, V. K. (2013). Modelling and Forecasting of Meat Exports from India. *Agricultural Economics Research Review*, 26 (2), 249-256.
- Paul, R. K., Alam, W. and Paul, A. K. (2014). Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences*, 84(4), 130-134.

ARTIFICIAL NEURAL NETWORKS

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

Introduction

Machine learning (ML) is a branch of Artificial Intelligence. In ML, computer-based algorithms are developed to make the system learn the complex pattern from the data and based on the learned pattern predictions are done for the new individuals. Artificial neural network (ANN) is one of the most important machine learning techniques. The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. In its simplest form, an artificial neural network (ANN) is an imitation of the human brain. The term "**Artificial Neural Network**" is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one another; artificial neural networks also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.

Human Brain and ANN

Brain is made of cells called neurons. Interconnection of such cells (neurons) makes up the neural network or the brain. ANN is an imitation of the natural neural network where the artificial neurons are connected in a similar fashion as the brain network. A biological neuron is made up of cell body, axon and dendrite. Dendrite receives electro-chemical signals from other neurons into the cell body. Cell body, called Soma contains nucleus and other chemical structures required to support the cell. Axon carries the signal from the neuron to other neurons. Connection between dendrites of two neurons, or neuron to muscle cells is called synapse.

An artificial neural network consists of processing units called neurons. An artificial neuron tries to replicate the structure and behaviour of the natural neuron. A neuron consists of inputs (dendrites), and one output (synapse via axon). The neuron has a function that determines the activation of the neuron.

Biological Neural Network	Artificial Neural Network
Dendrites	**Inputs**
Cell nucleus	**Nodes**
Synapse	**Weights**
Axon	**Output**

Brief History

The history of **neural networking** arguably began in the late **1800s** with scientific endeavours to study the activity of the human brain. In **1890**, **William James** published the first work about brain activity patterns. In **1943**, **McCulloch** and **Pitts** created a model of the neuron that is still used today in an **artificial neural network**.

In **1949**, **Donald Hebb** published "**The Organization of Behavior,"** which illustrated a law for synaptic neuron learning. This law, later known as **Hebbian Learning** in honor of Donald Hebb, is one of the most straight-forward and simple learning rules for artificial neural networks.

In **1951**, **Narvin Minsky** made the **first Artificial Neural Network (ANN)** while working at Princeton.

In **1958**, "**The Computer and the Brain**" were published, a year after **Jhon von Neumann's** death. In that book, von Neumann proposed numerous extreme changes to how analysts had been modelling the brain.

Perceptron was created in **1958**, at **Cornell University** by **Frank Rosenblatt**. The perceptron was an endeavour to use neural network procedures for character recognition. Perceptron was a linear system and was valuable for solving issues where the input classes were linearly separable in the input space. In **1960**, **Rosenblatt** published the book principles of neurodynamic, containing a bit of his research and ideas about modelling the brain.

The **backpropagation** algorithm, initially found by **Werbos** in **1974,** was rediscovered in **1986** with the book Learning Internal Representation by Error Propagation by Rumelhart, Hinton, and Williams. Backpropagation is a type of gradient descent algorithm used with artificial neural networks for reduction and curve-fitting.

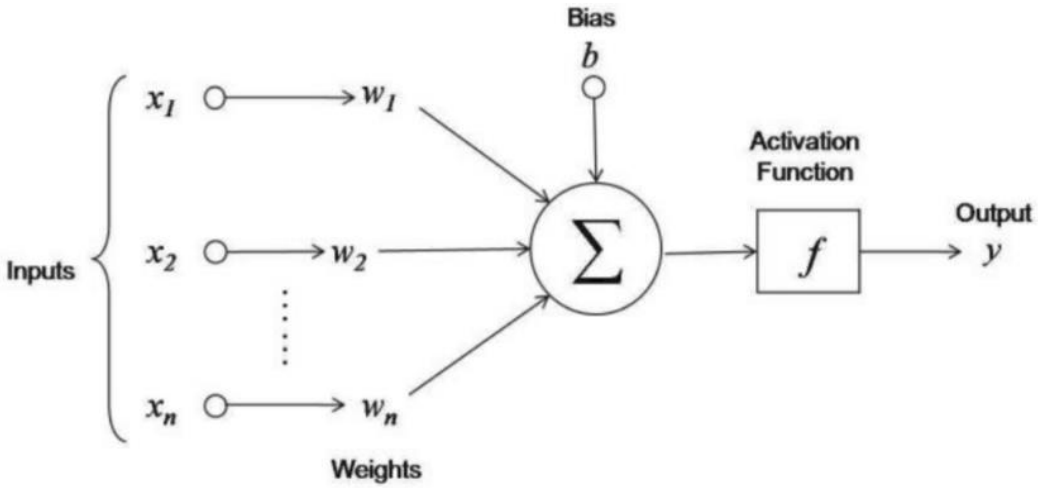
Architecture

Artificial neural networks are composed of elementary computational units called neurons combined according to different architectures. They can be arranged in layers (multi-layer network). Layered networks consist of:

Input layer: made of n neurons (one for each network input)

Hidden layer: composed of one or more hidden (or intermediate) layers consisting of m neurons;

Output layer: consisting of p neurons (one for each network output).



$$Y = \sum (weight * input) + bias$$

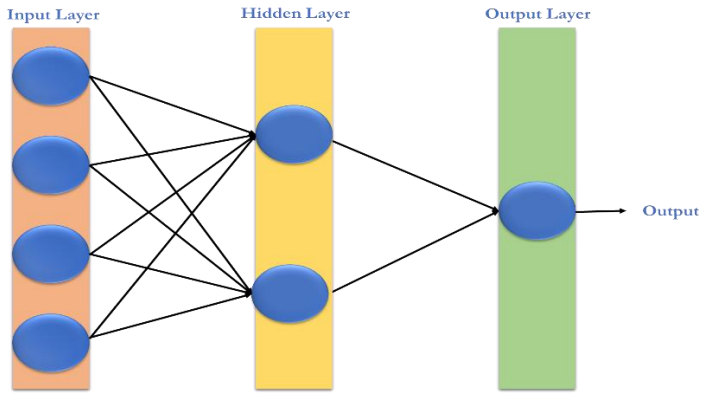


Figure 1. ANN architecture

Training algorithm (Back Propagation Algorithms)

Back propagation algorithm is the most commonly used method in training the neural network. Here the difference in targeted output, and the output obtained, is propagated back to the layers and the weights adjusted. A back propagation neural network (BPNN) uses a supervised learning method and feed-forward architecture. It is one of the most frequently utilized neural network techniques for classification and prediction. In BP algorithm, the outputs of hidden layers are propagated to the output layer where the output is calculated. This output is compared with the desired output for the given input. Based on this difference, the error is propagated back from the output layer to hidden layer, and from hidden layer to input layer. As the flow moves back, it changes the weights between the neurons. This cycle of going forward from input and output, and from output to input is called an epoch. A neural network is first given a set of known input data and asked to obtain a known output. This is called training the network. The network undergoes many such epochs till the error (difference between actual output and desired output is within a certain tolerance). Now the network is said to be trained. This process of training sets the weights

between all the neurons in all the layers. The weights so obtained from a trained network are used in calculating the response of the network to an unknown data.

Application

- **Artificial Neural Network for Classification**
- **Artificial Neural Network for Regression**
- **Artificial Neural Network for Time series analysis**

Implementing ANN model in R

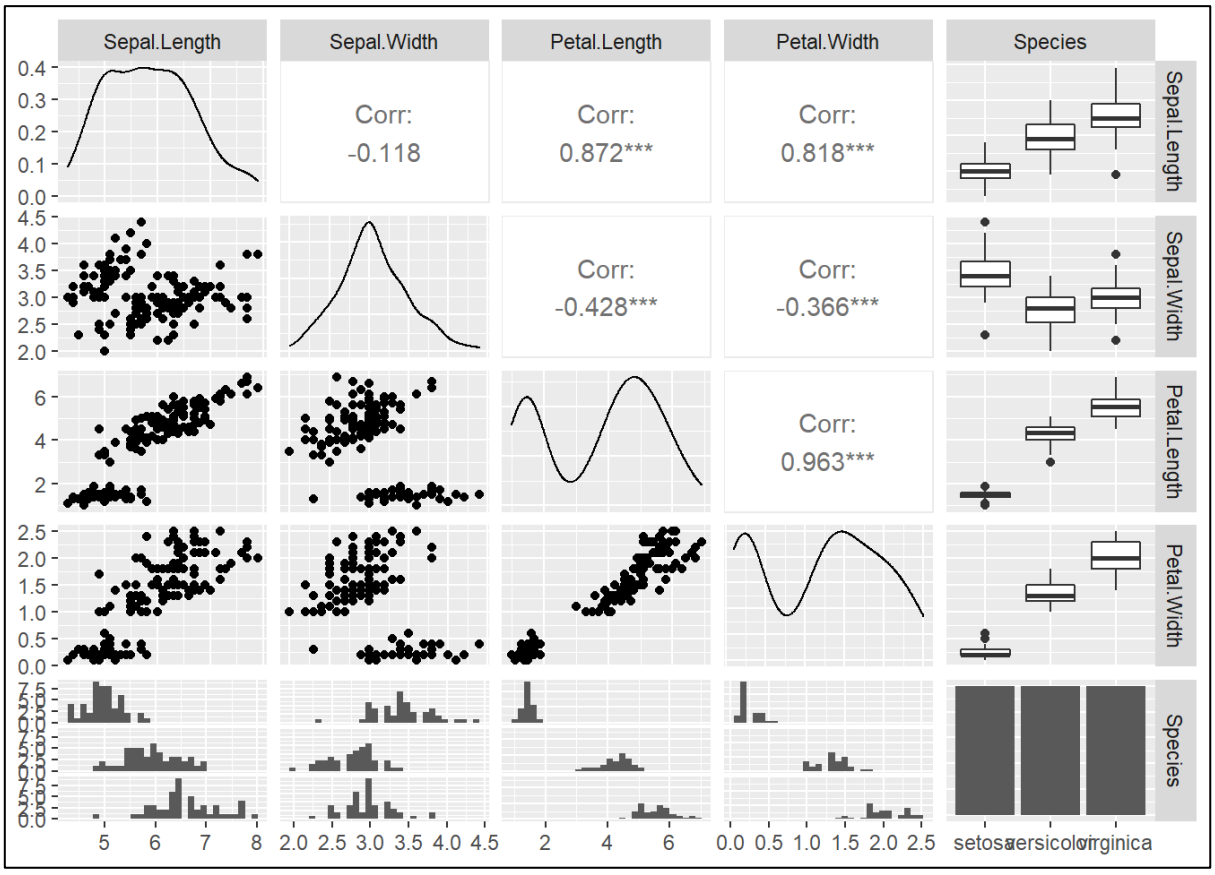
Artificial Neural Network for Classification

```
library(caTools)
library(nnet)
library(GGally)
```

```
## Loading required package: ggplot2
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
data<-iris
ggpairs(data)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
View(data)

part <- caTools::sample.split(data, SplitRatio = 0.80)

train_set <- subset(data, part == TRUE)

test_set <- subset(data, part == FALSE)

dim(train_set)

## [1] 120   5

dim(test_set)

## [1] 30  5

ann_cl<-nnet(Species~.,data = data,size = 3)

## # weights:  27

## initial  value 189.892267

## iter  10 value 104.371716

## iter  20 value 20.718849

## iter  30 value 6.884383

## iter  40 value 5.124974
```

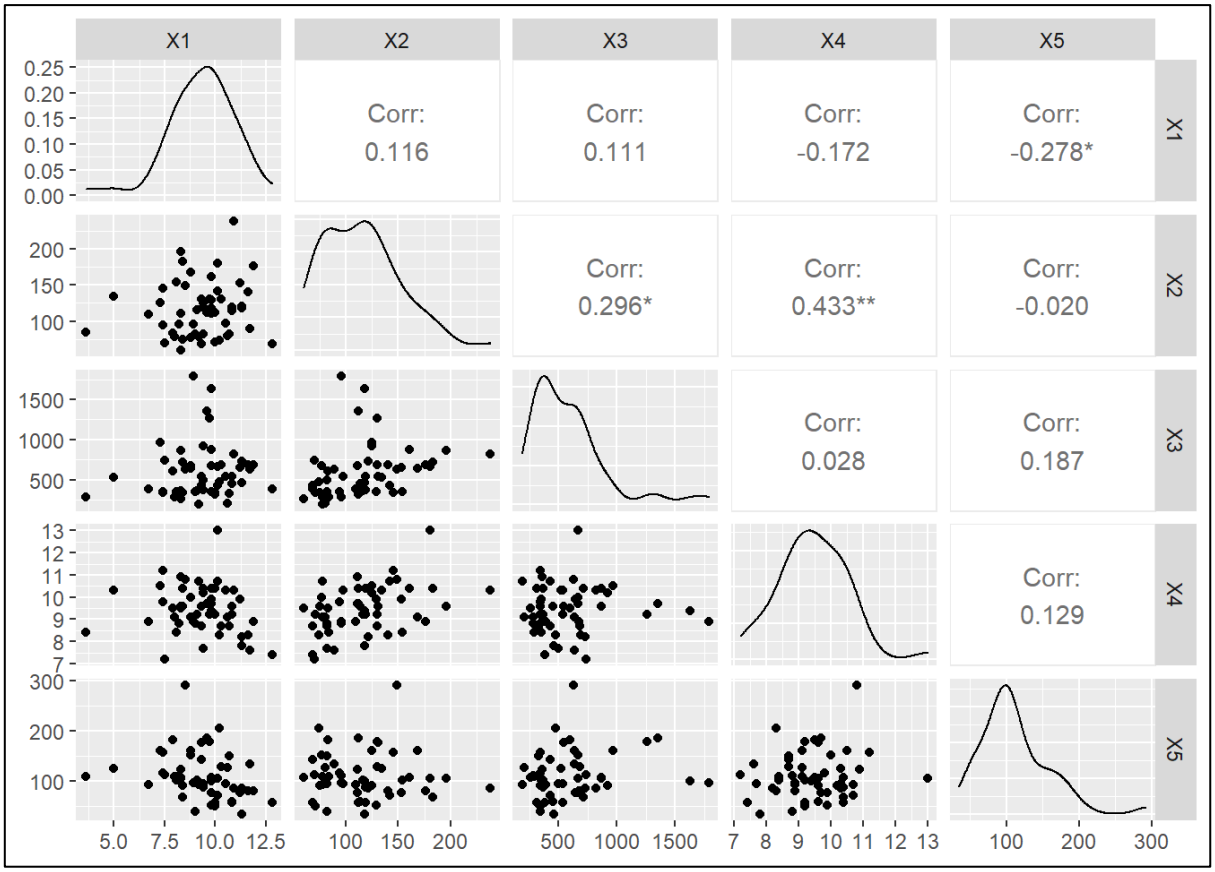
```
## iter  50 value 4.972721
## iter  60 value 4.927445
## iter  70 value 4.923758
## iter  80 value 4.922467
## iter  90 value 4.922451
## iter 100 value 4.922183
## final  value 4.922183
## stopped after 100 iterations
library(caret)
## Loading required package: lattice
varImp(ann_cl)
##                Overall    setosa versicolor virginica
## Sepal.Length  9.998404  9.998404   9.998404  9.998404
## Sepal.Width  19.021514 19.021514  19.021514 19.021514
## Petal.Length 43.686689 43.686689  43.686689 43.686689
## Petal.Width  27.293393 27.293393  27.293393 27.293393
test_pred <- predict(ann_cl,newdata=test_set, type = "class")


x<-as.factor(test_set$Species)
y<-as.factor(test_pred)
# Performance metrics


val<-confusionMatrix(x,y)
val
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         10          0         0
##   versicolor      0         10         0
##   virginica       0          0        10
```

```
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.8843, 1)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : 4.857e-15
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            1.0000           1.0000
## Specificity                 1.0000            1.0000           1.0000
## Pos Pred Value              1.0000            1.0000           1.0000
## Neg Pred Value              1.0000            1.0000           1.0000
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3333           0.3333
## Detection Prevalence        0.3333            0.3333           0.3333
## Balanced Accuracy           1.0000            1.0000           1.0000
```

Artificial Neural Network for Regression

```
library(neuralnet)
library(GGally)
## Loading required package: ggplot2
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(MLmetrics)
##
## Attaching package: 'MLmetrics'
## The following object is masked from 'package:base':
##
##     Recall
library(readxl)


data<- read_excel("E:/Training/Winter School ANN/Data_health.xlsx")
View(data)
dim(data)
## [1] 53  5
ggpairs(data)
```



```
train_set<-data[1:40,]
test_set<-data[41:53,]
dim(train_set)
## [1] 40  5
dim(test_set)
## [1] 13  5
ann_reg<-neuralnet::neuralnet(X1~.,data = train_set,hidden = 5)
plot(ann_reg)
test_pred <- predict(ann_reg,newdata=test_set)


#Performance metrics
mse <- MLmetrics::MSE(y_true=test_set$X1, y_pred=test_pred)
mae <- MLmetrics::MAE(y_true=test_set$X1, y_pred=test_pred)
mape <- MLmetrics::MAPE (y_true=test_set$X1, y_pred=test_pred)
rmse <- MLmetrics::RMSE(y_true=test_set$X1, y_pred=test_pred)


Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse)
Accuracy
```

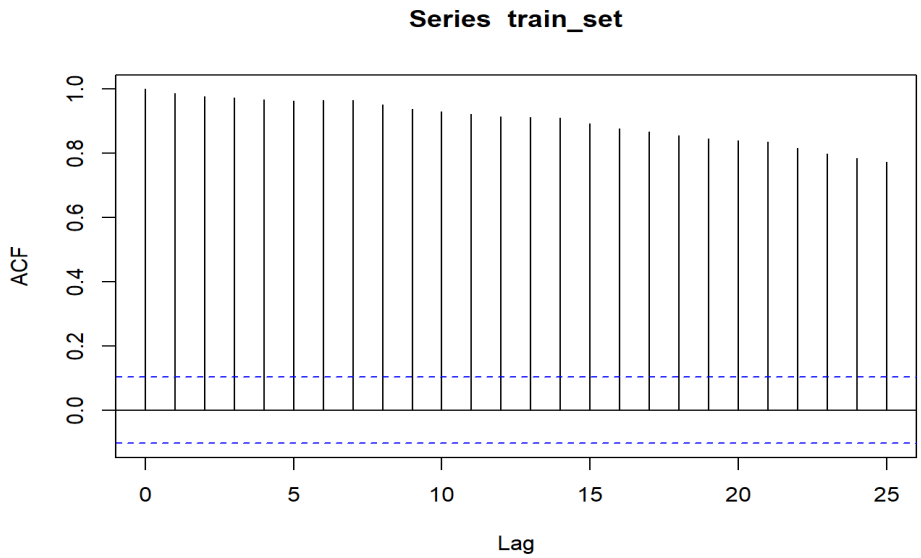
MSE	MAE	MAPE	RMSE
<dbl>	<dbl>	<dbl>	<dbl>
4.702489	1.692567	0.2400792	2.168522

1 row

```
results<-cbind(test_set$X1,test_pred)


matplot(results, type = "l",pch=1,col = 1:2)
legend("topright", legend = c("actual", "predicted"), col=1:2, pch=1)
```

Artificial Neural Network for Time series analysis

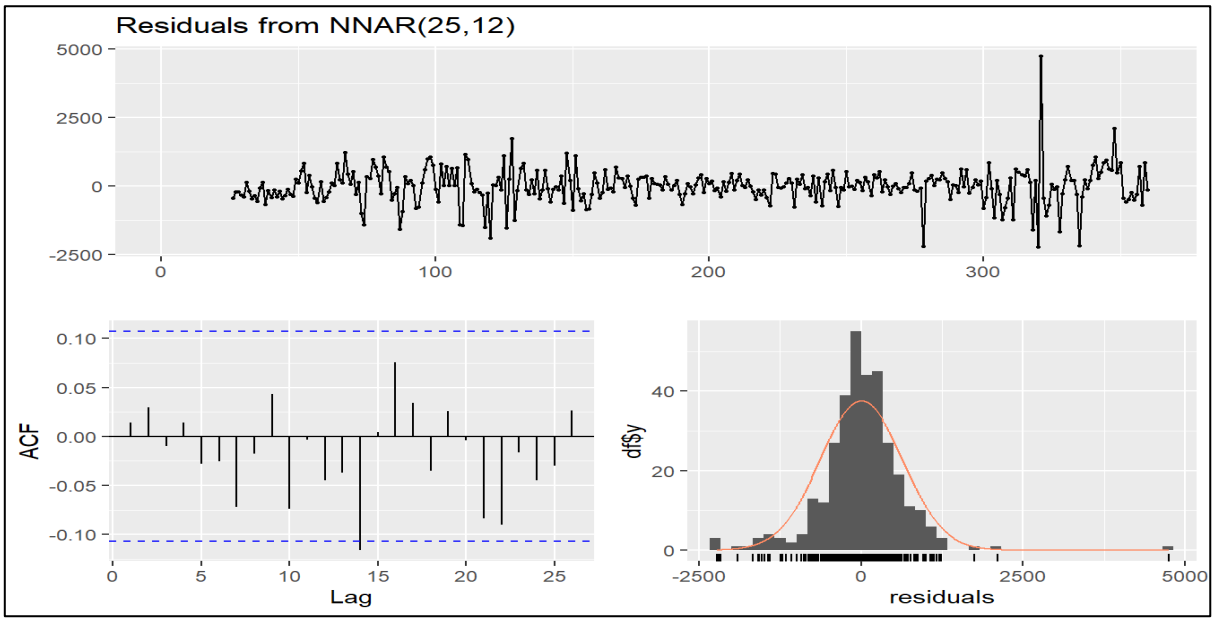
```
library(forecast)
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
library(readxl)
library(TSANN)
covid <- read_excel("E:/Training/Winter School ANN/covid.xlsx")


data<-covid$All_india
length(data)
## [1] 450
train_set <- data[c(1:(length(data)*.8))]
test_set<- data[-c(1:(length(data)*.8))]
length(train_set)
## [1] 360
length(test_set)
## [1] 90
acf(train_set)
```



```
ann<- forecast::nnetar(train_set, p=25, size = 12)
```

```
residual_analysis<-checkresiduals(ann)
```

```
## Warning in modeldf.default(object): Could not find appropriate degrees of
## freedom for this model.
```




```
test_pred<-forecast::forecast(ann,h=length(test_set),data=test_set)
str(test_pred)
```

```
## List of 16
```

```
pred<-test_pred$mean
# Performance metrics
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
## The following object is masked from 'package:base':
##
##     Recall
```

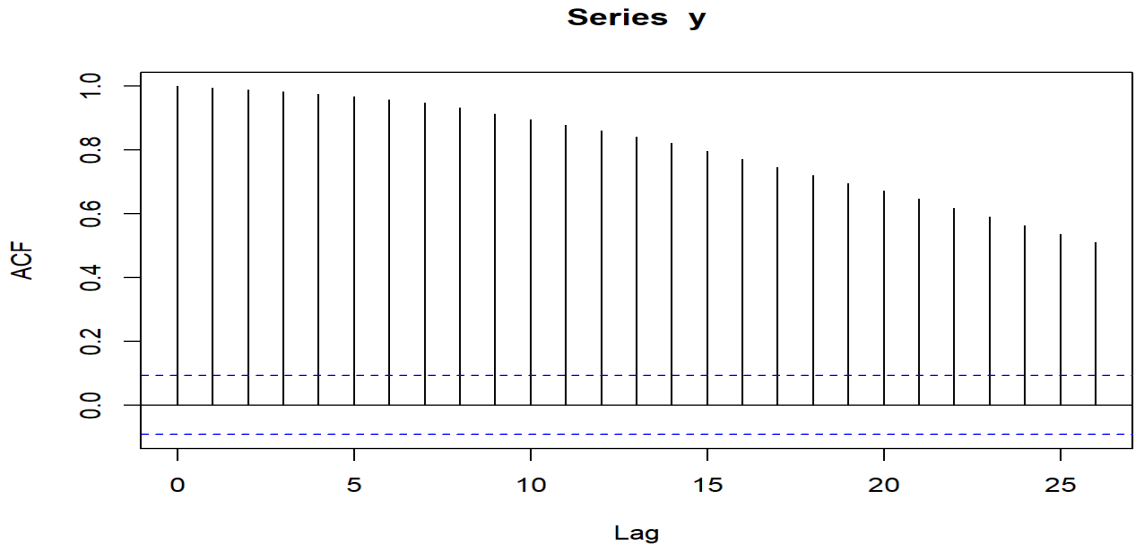
```
mse <- MLmetrics::MSE(y_true=test_set, y_pred=pred)
mae <- MLmetrics::MAE(y_true=test_set, y_pred=pred)
```

```
mape <- MLmetrics::MAPE (y_true=test_set, y_pred=pred)

rmse <- MLmetrics::RMSE(y_true=test_set, y_pred=pred)


Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse)

Accuracy
```

MSE	MAE	MAPE	RMSE
<dbl>	<dbl>	<dbl>	<dbl>
34056753345	144028.1	0.6072153	184544.7

1 row

```
tsann<-TSANN::Auto.TSANN(data, min.size=2, max.size=4,split.ratio=0.8)
```




```
tsann$Train.RMSE
```

```
## [1] 1326.419
```

```
tsann$FinalModel
```

```
##
## Average of 20 networks, each of which is
## a 21-4-1 network with 93 weights
## options were - linear output units
```

REFERENCES

- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.
- James, W. (1890). *The principles of psychology* (Vols. 1–2). New York, NY: Henry Holt and Company.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4), 115–133. https://doi.org/10.1007/BF02478259
- Minsky, M. (1954). *Neural nets and the brain model problem* (Doctoral dissertation). Princeton University, Princeton, NJ.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. https://doi.org/10.1037/h0042519
- Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan Books.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.
- Von Neumann, J. (1958). *The computer and the brain*. New Haven, CT: Yale University Press.
- Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Doctoral dissertation). Harvard University, Cambridge, MA.

OVERVIEW ON R SOFTWARE

Dr. Md Yeasin[1] and Dr. Ranjit Kumar Paul[2]
[1]Scientist and [2]National Fellow
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
yeasin.iasri@gmail.com, ranjitstat@gmail.com

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R is a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

R environment

The R environment provides an integrated suite of software facilities for data manipulation, calculation and graphical display. It has

- a data handling and storage facility,
- a suite of operators for calculations on arrays and matrices,
- a large, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display, and
- a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined functions and input and output facilities.

Origin

R can be regarded as an implementation of the S language which was developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland started the project on R in 1995 and hence the name software has been named as 'R'.

R was introduced as an environment within which many classical and modern statistical techniques can be implemented. A few of these are built into the base R environment, but many are supplied as packages. There are a number of packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via http://cran.r-project.org) and elsewhere.

Availability

Since R is an open source project, it can be obtained freely from the website www.**r**-project.org. One can download R from any CRAN mirror out of several CRAN (Comprehensive R Archive Network) mirrors.

Installation

To install R in windows operating system, simply double click on the setup file. It will automatically install the software in the system.

Usage

R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windows set up only.

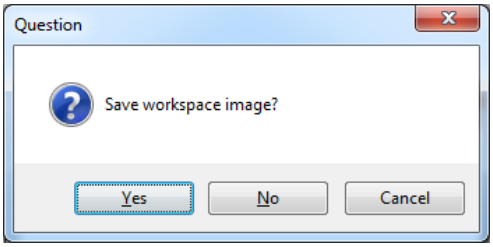
Difference with other packages

There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output and store the results in an object for subsequent interrogation by further R functions.

Invoking R

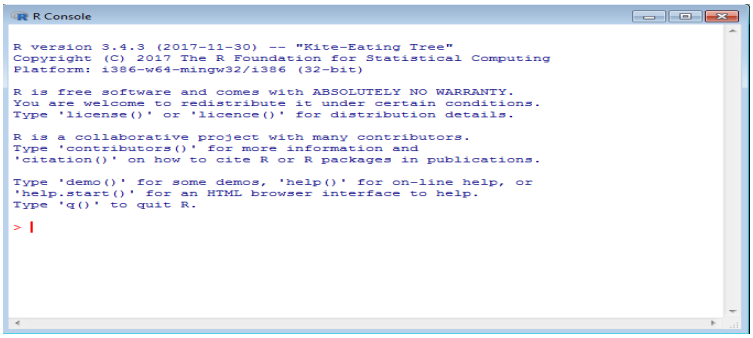
If properly installed, usually R has a shortcut icon on the desktop screen and/or we can find it under Start|All Programs|R menu.

To quit R, type *q()* at the R prompt (>) and press Enter key. A dialog box will ask whether to save the objects we have created during the session so that they will become available next time when R will be invoked.



Windows of R

R has only one window and when R is started it looks like



R commands

i. R commands are case sensitive, so X and x are different symbols and would refer to different variables.
ii. Elementary commands consist of either expressions or assignments.
iii. If an expression is given as a command, it is evaluated, printed and the value is lost.
iv. An assignment also evaluates an expression and passes the value to a variable but the result is not automatically printed.
v. Commands are separated either by a semi-colon (';'), or by a newline.
vi. Elementary commands can be grouped together into one compound expression by braces '{' and '}'.
vii. Comments can be put almost anywhere, starting with a hashmark ('#'). Anything written after # marks to the end of the line is considered as a comment.
viii. Window can be cleared of lines by pressing Ctrl + L keys.

Data Types

R has a wide variety of data types including scalars, vectors (numerical, character, logical), matrices, dataframes, and lists.

Vectors

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector
```

Refer to elements of a vector using subscripts.

```
a[c(2,4)] # 2nd and 4th elements of vector
```

Matrices

All columns in a matrix must have the same mode(numeric, character, etc.) and the same length. The general format is

```
mymatrix <- matrix(vector, nrow=r, ncol=c, byrow=FALSE,
  dimnames=list(char_vector_rownames, char_vector_colnames))
```

byrow=TRUE indicates that the matrix should be filled by rows. **byrow=FALSE** indicates that the matrix should be filled by columns (the default). **dimnames** provides optional labels for the columns and rows.

```
# generates 5 x 4 numeric matrix
y<-matrix(1:20, nrow=5,ncol=4)

# another example
cells <- c(1,26,24,68)
rnames <- c("R1", "R2")
cnames <- c("C1", "C2")
mymatrix <- matrix(cells, nrow=2, ncol=2, byrow=TRUE,
  dimnames=list(rnames, cnames))
```

Identify rows, columns or elements using subscripts.

```
x[,4] # 4th column of matrix
x[3,] # 3rd row of matrix
x[2:4,1:3] # rows 2,3,4 of columns 1,2,3
```

Arrays

Arrays are similar to matrices but can have more than two dimensions. See **help(array)** for details.

Dataframes

A dataframe is more general than a matrix, in that different columns can have different modes (numeric, character, factor, etc.). This is similar to SAS and SPSS datasets.

```
d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)
mydata <- data.frame(d,e,f)
names(mydata) <- c("ID","Color","Passed") # variable names
```

There are a variety of ways to identify the elements of a dataframe .

```
myframe[3:5] # columns 3,4,5 of dataframe
myframe[c("ID","Age")] # columns ID and Age from dataframe
myframe$X1 # variable x1 in the dataframe
```

Lists

An ordered collection of objects (components). A list allows us to gather a variety of (possibly unrelated) objects under one name.

```
# example of a list with 4 components -
# a string, a numeric vector, a matrix, and a scaler
w <- list(name="Fred", mynumbers=a, mymatrix=y, age=5.3)

# example of a list containing two lists
v <- c(list1,list2)
```

Intentify elements of a list using the [[]] convention.

```
mylist[[2]] # 2nd component of the list
mylist[["mynumbers"]] # component named mynumbers in list
```

Factors

Tell **R** that a variable is **nominal** by making it a factor. The factor stores the nominal values as a vector of integers in the range [1... k] (where k is the number of unique values in the nominal variable), and an internal vector of character strings (the original values) mapped to these integers.

```
# variable gender with 20 "male" entries and
# 30 "female" entries
gender <- c(rep("male",20), rep("female", 30))
gender <- factor(gender)
# stores gender as 20 1s and 30 2s and associates
# 1=female, 2=male internally (alphabetically)
# R now treats gender as a nominal variable
summary(gender)
```

An ordered factor is used to represent an **ordinal variable**.

```
# variable rating coded as "large", "medium", "small'
rating <- ordered(rating)
# recodes rating to 1,2,3 and associates
# 1=large, 2=medium, 3=small internally
# R now treats rating as ordinal
```

R will treat factors as nominal variables and ordered factors as ordinal variables in statistical proceedures and graphical analyses. We can use options in the **factor( )** and **ordered( )** functions

to control the mapping of integers to strings (overiding the alphabetical ordering). We can also use factors to create value labels.

Executing commands from or diverting output to a file

If commands are stored in an external file, say 'D:/commands.txt' they may be executed at any time in an R session with the command

```
> source("d:/commands.txt")
```

For Windows Source is also available on the File menu.

The function *sink()*,

```
> sink("d:/record.txt")
```

will divert all subsequent output from the console to an external file, 'record.txt' in D drive. The command

```
> sink()
```

restores it to the console once again.

Simple manipulations of numbers and vectors

R operates on named data structures. The simplest such structure is the numeric vector, which is a single entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

The function *c()* assigns the five numbers to the vector x. The assignment operator (<-) 'points' to the object receiving the value of the expression. Once can use the '=' operator as an alternative. A single number is taken as a vector of length one.

Assignments can also be made in the other direction, using the obvious change in the assignment operator. So the same assignment could be made using

```
> c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed. So now if we were to use the command given below. The reciprocals of the five values would be printed at the terminal.

```
> 1/x
```

Operators

R's binary and logical operators will look very familiar to programmers. Note that binary operators work on vectors and matrices as well as scalars.

Arithmetic Operators

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
^ or **	exponentiation
x %% y	modulus (x mod y) 5%%2 is 1
x %/% y	integer division 5%/%2 is 2

Logical Operators

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
!=	not equal to

!x	Not x
x | y	x OR y
x & y	x AND y
isTRUE(x)	test if X is TRUE

```
 # An example
x <- c(1:10)
x[(x>8) | (x<5)]
# yeilds 1 2 3 4 9 10

# How it works
x <- c(1:10)
x
1 2 3 4 5 6 7 8 9 10
x > 8
F F F F F F F F T T
x < 5
T T T T F F F F F F
x > 8 | x < 5
T T T T F F F F T T
x[c(T,T,T,T,F,F,F,F,T,T)]
1 2 3 4 9 10
```

Built-in Functions

Almost everything in **R** is done through functions. Here I'm only refering to numeric and character functions that are commonly used in creating or recoding variables.

Numeric Functions

Function	**Description**
abs(*x*)	absolute value
sqrt(*x*)	square root
ceiling(*x*)	ceiling(3.475) is 4
floor(*x*)	floor(3.475) is 3

trunc(*x***)**	trunc(5.99) is 5
round(*x***, digits=***n***)**	round(3.475, digits=2) is 3.48
signif(*x***, digits=***n***)**	signif(3.475, digits=2) is 3.5
cos(*x***), sin(***x***), tan(***x***)**	also acos(*x*), cosh(*x*), acosh(*x*), etc.
log(*x***)**	natural logarithm
log10(*x***)**	common logarithm
exp(*x***)**	e^*x*

Character Functions

Function	**Description**
substr(*x***, start=***n1***, stop=***n2***)**	Extract or replace substrings in a character vector. x <- "abcdef" substr(x, 2, 4) is "bcd" substr(x, 2, 4) <- "22222" is "a222ef"
grep(*pattern***, *x* ,** **ignore.case=**FALSE**,** **fixed=**FALSE**)**	Search for *pattern* in *x*. If fixed =FALSE then *pattern* is a regular expression. If fixed=TRUE then *pattern* is a text string. Returns matching indices. grep("A", c("b","A","c"), fixed=TRUE) returns 2
sub(*pattern***, *replacement*, *x*,** **ignore.case =**FALSE**,** **fixed=**FALSE**)**	Find *pattern* in *x* and replace with *replacement* text. If fixed=FALSE then *pattern* is a regular expression. If fixed = T then *pattern* is a text string. sub("\\s",".","Hello There") returns "Hello.There"
strsplit(*x***, *split*)**	Split the elements of character vector *x* at *split*. strsplit("abc", "") returns 3 element vector "a","b","c"
paste(..., sep="")	Concatenate strings after using *sep* string to seperate them. paste("x",1:3,sep="") returns c("x1","x2" "x3") paste("x",1:3,sep="M") returns c("xM1","xM2" "xM3") paste("Today is", date())

toupper(*x***)**	Uppercase
tolower(*x***)**	Lowercase

Statistical Probability Functions

The following table describes functions related to probaility distributions. For random number generators below, we can use set.seed(1234) or some other integer to create reproducible pseudo-random numbers.

Function	**Description**

dnorm(*x***)**	normal density function (by default m=0 sd=1)
	# plot standard normal curve
	x <- pretty(c(-3,3), 30)
	y <- dnorm(x)
	plot(x, y, type='l', xlab="Normal Deviate", ylab="Density", yaxs="i")

pnorm(*q***)**	cumulative normal probability for q (area under the normal curve to the right of q) pnorm(1.96) is 0.975
qnorm(*p***)**	normal quantile. value at the p percentile of normal distribution qnorm(.9) is 1.28 # 90th percentile
rnorm(*n***, m**=0**,sd**=1**)**	n random normal deviates with mean m and standard deviation sd. #50 random normal variates with mean=50, sd=10 x <- rnorm(50, m=50, sd=10)

dbinom(*x***,** *size***,** *prob***)**
pbinom(*q***,** *size***,** *prob***)**
qbinom(*p***,** *size***,** *prob***)**
rbinom(*n***,** *size***,** *prob***)**
	binomial distribution where size is the sample size
	and prob is the probability of a heads (pi)
	# prob of 0 to 5 heads of fair coin out of 10 flips
	dbinom(0:5, 10, .5)
	# prob of 5 or less heads of fair coin out of 10 flips
	pbinom(5, 10, .5)

dpois(*x***,** *lamda***)**
ppois(*q***,** *lamda***)**
qpois(*p***,** *lamda***)**
	poisson distribution with m=std=lamda
	#probability of 0,1, or 2 events with lamda=4
	dpois(0:2, 4)

rpois(*n***,** *lamda***)**	# probability of at least 3 events with lamda=4 1- ppois(2,4)
dunif(*x***, min**=0**, max**=1**)** **punif(***q***, min=0, max**=1**)** **qunif(***p***, min**=0**, max**=1**)** **runif(***n***, min**=0**, max**=1**)**	uniform distribution, follows the same pattern as the normal distribution above. #10 uniform random variates x <- runif(10)

Other Statistical Functions

Other useful statistical functions are provided in the following table. Each has the option na.rm to strip missing values before calculations. Otherwise the presence of missing values will lead to a missing result. Object can be a numeric vector or dataframe.

Function	**Description**
mean(*x***, trim**=0**,** **na.rm=**FALSE**)**	mean of object x # trimmed mean, removing any missing values and # 5 percent of highest and lowest scores mx <- mean(x,trim=.05,na.rm=TRUE)
sd(*x***)**	standard deviation of object(x). also look at var(x) for variance and mad(x) for median absolute deviation.
median(*x***)**	median
quantile(*x***,** *probs***)**	quantiles where x is the numeric vector whose quantiles are desired and probs is a numeric vector with probabilities in [0,1]. # 30th and 84th percentiles of x y <- quantile(x, c(.3,.84))
range(*x***)**	range
sum(*x***)**	sum
diff(*x***, lag=***1***)**	lagged differences, with lag indicating which lag to use
min(*x***)**	minimum
max(*x***)**	maximum

scale(*x***, center**=TRUE**, scale**=TRUE**)**	column center or standardize a matrix.

Other Useful Functions

Function	**Description**
seq(*from* **,** *to***,** *by***)**	generate a sequence indices <- seq(1,10,2) #indices is c(1, 3, 5, 7, 9)
rep(*x***,** *ntimes***)**	repeat *x n* times y <- rep(1:3, 2) # y is c(1, 2, 3, 1, 2, 3)
cut(*x***,** *n***)**	divide continuous variable in factor with *n* levels y <- cut(x, 5)

Note that while the examples on this page apply functions to individual variables, many can be applied to vectors and matrices as well.

Descriptive Statistics

R provides a wide range of functions for obtaining summary statistics. One method of obtaining descriptive statistics is to use the **sapply( )** function with a specified summary statistic.

```
# get means for variables in dataframe mydata
# excluding missing values
sapply(mydata, mean, na.rm=TRUE)
```

Possible functions used in sapply include **mean, sd, var, min, max, med, range, and quantile**.

There are also numerous **R** functions designed to provide a range of descriptive statistics at once. For example

```
# mean,median,25th and 75th quartiles,min,max
summary(mydata)

# Tukey min,lower-hinge, median,upper-hinge,max
fivenum(x)
```

Using the Hmisc package

```
library(Hmisc)
describe(mydata)
# n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles
# 5 lowest and 5 highest scores
```

Using the pastecs package

```
library(pastecs)
stat.desc(mydata)
# nbr.val, nbr.null, nbr.na, min max, range, sum,
# median, mean, SE.mean, CI.mean, var, std.dev, coef.var
```

Using the psych package

```
library(psych)
describe(mydata)
# item name ,item number, nvalid, mean, sd,
# median, mad, min, max, skew, kurtosis, se
```

Summary Statistics by Group

A simple way of generating summary statistics by grouping variable is available in the psych package.

```
library(psych)
describe.by(mydata, group,...)
```

The **doBy** package provides much of the functionality of SAS PROC SUMMARY. It defines the desired table using a model formula and a function. Here is a simple example.

```
library(doBy)
summaryBy(mpg + wt ~ cyl + vs, data = mtcars,
  FUN = function(x) { c(m = mean(x), s = sd(x)) } )
# produces mpg.m wt.m mpg.s wt.s for each
# combination of the levels of cyl and vs
```

See also: aggregating data.

Frequencies and Crosstabs

This section describes the creation of frequency and contingency tables from categorical variables, along with tests of independence, measures of association, and methods for graphically displaying results.

Generating Frequency Tables

R provides many methods for creating frequency and contingency tables. Three are described below. In the following examples, assume that A, B, and C represent categorical variables.

We can generate frequency tables using the **table( )** function, tables of proportions using the **prop.table( )** function, and marginal frequencies using **margin.table( )**.

```
# 2-Way Frequency Table
attach(mydata)
mytable <- table(A,B) # A will be rows, B will be columns
mytable # print table

margin.table(mytable, 1) # A frequencies (summed over B)
margin.table(mytable, 2) # B frequencies (summed over A)

prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages
prop.table(mytable, 2) # column percentages
```

table( ) can also generate multidimensional tables based on 3 or more categorical variables. In this case, use the **ftable( )** function to print the results more attractively.

```
# 3-Way Frequency Table
mytable <- table(A, B, C)
ftable(mytable)
```

Table ignores missing values. To include **NA** as a category in counts, include the table option exclude=NULL if the variable is a vector. If the variable is a factor we have to create a new factor using

```
newfactor <- factor(oldfactor, exclude=NULL)
```

xtabs

The **xtabs( )** function allows us to create crosstabulations using formula style input.

```
# 3-Way Frequency Table
mytable <- xtabs(~A+B+c, data=mydata)
ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence
```

If a variable is included on the left side of the formula, it is assumed to be a vector of frequencies (useful if the data have already been tabulated).

Crosstable

The **CrossTable( )** function in the gmodels package produces crosstabulations modeled after PROC FREQ in **SAS** or CROSSTABS in **SPSS**. It has a wealth of options.

```
# 2-Way Cross Tabulation
library(gmodels)
CrossTable(mydata$myrowvar, mydata$mycolvar)
```

There are options to report percentages (row, column, cell), specify decimal places, produce Chi-square, Fisher, and McNemar tests of independence, report expected and residual values (pearson, standardized, adjusted standardized), include missing values as valid, annotate with row and column titles, and format as **SAS** or **SPSS** style output!
See **help(CrossTable)** for details.

Tests of Independence

Chi-Square Test

For 2-way tables we can use **chisq.test(***mytable***)** to test independence of the row and column variable. By default, the p-value is calculated from the asymptotic chi-squared distribution of the test statistic. Optionally, the p-value can be derived via Monte Carlo simultation.

Fisher Exact Test

fisher.test(*x***)** provides an exact test of independence. *x* is a two dimensional contingency table in matrix form.

Mantel-Haenszel test

Use the **mantelhaen.test(***x***)** function to perform a Cochran-Mantel-Haenszel chi-squared test of the null hypothesis that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction. *x* is a 3 dimensional contingency table, where the last dimension refers to the strata.

Loglinear Models

We can use the **loglm( )** function in the **MASS** package to produce log-linear models. For example, let's assume we have a 3-way contingency table based on variables A, B, and C.

```
library(MASS)
mytable <- xtabs(~A+B+C, data=mydata)
```

We can perform the following tests:

Mutual Independence: A, B, and C are pairwise independent. loglm(~A+B+C, mytable)

Partial Independence: A is partially independent of B and C (i.e., A is independent of the composite variable BC).

```
loglin(~A+B+C+B*C, mytable)
```

Conditional Independence: A is independent of B, given C.

```
loglm(~A+B+C+A*C+B*C, mytable) #No Three-Way Interaction

loglm(~A+B+C+A*B+A*C+B*C, mytable)
```

Martin Theus and Stephan Lauer have written an excellent article on Visualizing Loglinear Models, using mosaic plots. There is also great tutorial example by Kevin Quinn on analyzing loglinear models via glm.

Measures of Association

The **assocstats(***mytable***)** function in the **vcd** package calculates the phi coefficient, contingency coefficient, and Cramer's V for an rxc table. The **kappa(***mytable***)** function in the **vcd** package calculates Cohen's kappa and weighted kappa for a confusion matrix. See Richard Darlington's article on Measures of Association in Crosstab Tables for an excellent review of these statistics.

Visualizing results

Use bar and pie charts for visualizing frequencies in one dimension.

Use the vcd package for visualizing relationships among categorical data (e.g. mosaic and association plots).

Use the **ca** package for correspondence analysis (visually exploring relationships between rows and columns in contingency tables).

Converting Frequency Tables to an "Original" Flat file

Finally, there may be times that we wil need the original "flat file" dataframe rather than the frequency table. Marc Schwartz has provided code on the Rhelp mailing list for converting a table back into a dataframe.

Correlations

We can use the **cor( )** function to produce correlations and the **cov( )** function to produces covariances.

A simplified format is **cor(x, use=, method=)** where

Option	Description
x	Matrix or data frame
use	Specifies the handling of missing data. Options are **all.obs** (assumes no missing data - missing data will produce an error), **complete.obs** (listwise deletion), and **pairwise.complete.obs** (pairwise deletion)
method	Specifies the type of correlation. Options are **Pearson**, **Spearman** or **Kendall**.

```
# Correlations/covariances among numeric variables in
# dataframe mtcars. Use listwise deletion of missing data.
cor(mtcars, use="complete.obs", method="kendall")
cov(mtcars, use="complete.obs")
```

Unfortunately, neither **cor( )** or **cov( )** produce tests of significance, although we can use the **cor.test( )** function to test a single correlation coefficient.

The **rcorr( )** function in the **Hmisc** package produces correlations/covariances and significance levels for pearson and spearman correlations. However, input must be a matrix and pairwise deletion is used.

```
# Correlations with significance levels
library(Hmisc)
rcorr(x, type="pearson") # type can be pearson or spearman

#mtcars is a dataframe
rcorr(as.matrix(mtcars))
```

We can use the format **cor(X, Y)** or **rcorr(X, Y)** to generate correlations between the columns of X and the columns of Y. This similar to the VAR and WITH commands in **SAS** PROC CORR.

```
# Correlation matrix from mtcars
# with mpg, cyl, and disp as rows
```

```
# and hp, drat, and wt as columns
x <- mtcars[1:3}
y <- mtcars[4:6]
cor(x, y)
```

Other Types of Correlations

```
# polychoric correlation
# x is a contingency table of counts
library(polychor)
polychor(x)

# heterogeneous correlations in one matrix
# pearson (numeric-numeric),
# polyserial (numeric-ordinal),
# and polychoric (ordinal-ordinal)
# x is a dataframe with ordered factors
# and numeric variables
library(polychor)
hetcor(x)


# partial correlations
library(ggm)
data(mydata)
pcor(c("a", "b", "x", "y", "z"), var(mydata))
# partial corr between a and b controlling for x, y, z
```

Visualizing Correlations

Use **corrgram( )** to plot correlograms .

Use the pairs() or splom( ) to create scatterplot matrices.

A great example of a plotted **correlation matrix** can be found in the R Graph Gallery.

t-tests

The **t.test( )** function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance and applies the Welsh df modification.# independent 2-group t-test

```
t.test(y~x) # where y is numeric and x is a binary factor

# independent 2-group t-test
t.test(y1,y2) # where y1 and y2 are numeric
```

```
# paired t-test
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric

# one samle t-test
t.test(y,mu=3) # Ho: mu=3
```

We can use the **var.equal = TRUE** option to specify equal variances and a pooled variance estimate. We can use the **alternative="less"** or **alternative="greater"** option to specify a one tailed test.

Nonparametric and resampling alternatives to t-tests are available.

Visualizing Results

Use box plots or density plots to visualize group differences.

Nonparametric Tests of Group Differences

R provides functions for carrying out Mann-Whitney U, Wilcoxon Signed Rank, Kruskal Wallis, and Friedman tests.

```
# independent 2-group Mann-Whitney U Test
wilcox.test(y~A)
# where y is numeric and A is A binary factor

# independent 2-group Mann-Whitney U Test
wilcox.test(y,x) # where y and x are numeric

# dependent 2-group Wilcoxon Signed Rank Test
wilcox.test(y1,y2,paired=TRUE) # where y1 and y2 are numeric

# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test(y~A) # where y1 is numeric and A is a factor

# Randomized Block Design - Friedman Test
friedman.test(y~A|B)
# where y are the data values, A is a grouping factor
# and B is a blocking factor
```

For the wilcox.test we can use the **alternative="less"** or **alternative="greater"** option to specify a one tailed test.

Parametric and resampling alternatives are available.

The package npmc provides nonparametric multiple comparisons.

```
library(npmc)
npmc(x)
```

```
# where x is a dataframe containing variable 'var'
# (response variable) and 'class' (grouping variable)
```

Visualizing Results

Use box plots or density plots to visual group differences.

Multiple (Linear) Regression

R provides comprehensive support for multiple linear regression. The topics below are provided in order of increasing complexity.

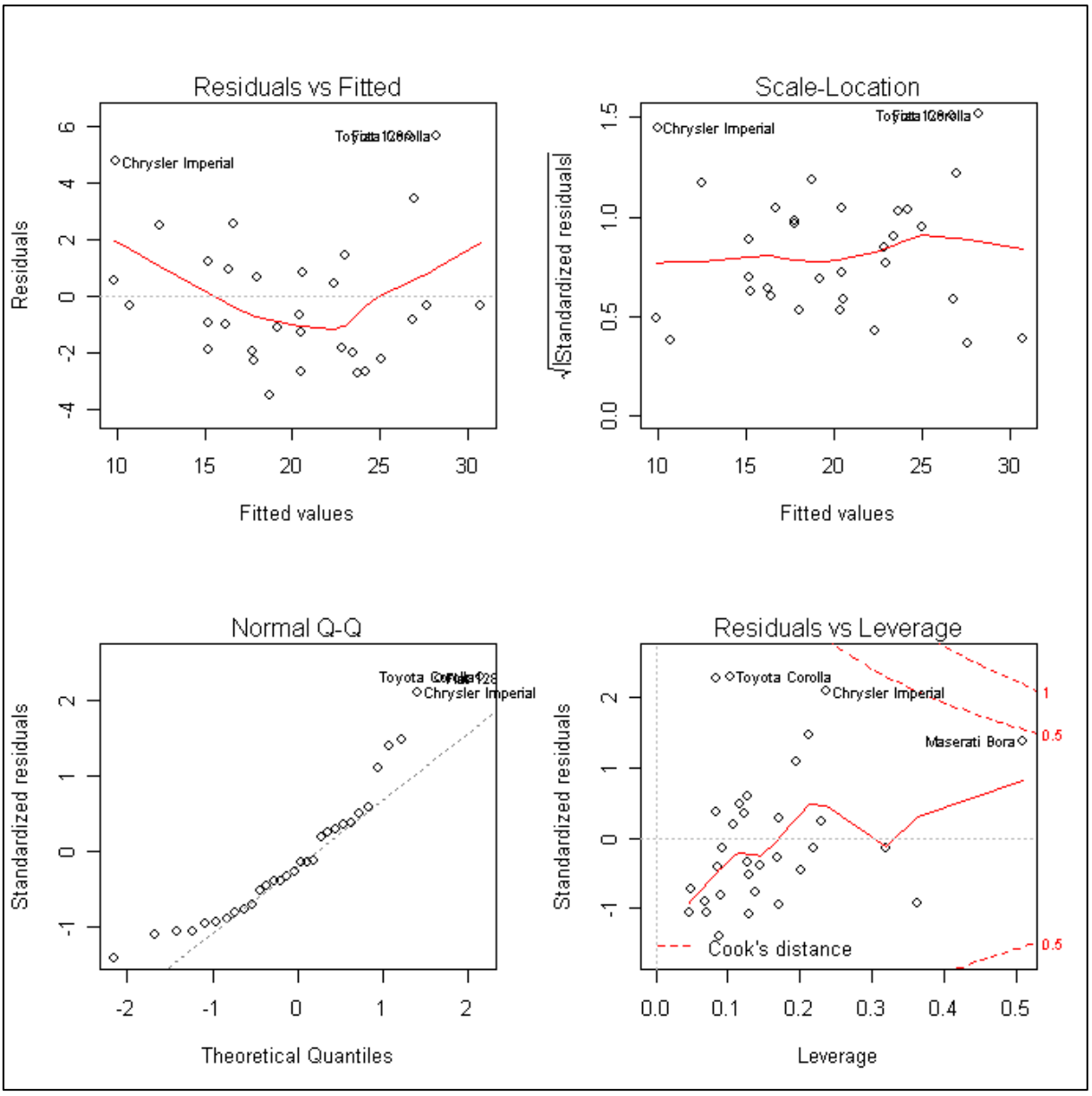
Fitting the Model

```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results

# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

Diagnostic Plots

Diagnostic plots provide checks for heteroscedasticity, normality, and influential observerations.

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)
```



Comparing Models

We can compare nested models with the anova( ) function. The following code provides a simultaneous test that x3 and x4 add to linear prediction above and beyond x1 and x2.

```
# compare models
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)
```

Cross Validation

We can do K-Fold cross-validation using the **cv.lm( )** function in the DAAG package.

```
# K-fold cross-validation
library(DAAG)
cv.lm(df=mydata, fit, m=3) # 3 fold cross-validation
```

Sum the MSE for each fold, divide by the number of observations, and take the square root to get the cross-validated standard error of estimate.

We can assess **R2 shrinkage** via K-fold cross-validation. Using the **crossval()** function from the bootstrap package, do the following:

```
# Assessing R2 shrinkage using 10-Fold Cross-Validation

fit <- lm(y~x1+x2+x3,data=mydata)

library(bootstrap)
# define functions
theta.fit <- function(x,y){lsfit(x,y)}
theta.predict <- function(fit,x){cbind(1,x)%*%fit$coef}
# matrix of predictors
X <- as.matrix(mydata[c("x1","x2","x3")])
# vector of predicted values
y <- as.matrix(mydata[c("y")])

results <- crossval(X,y,theta.fit,theta.predict,ngroup=10)
cor(y, fit$fitted.values)**2 # raw R2
cor(y,results$cv.fit)**2 # cross-validated R2
```

Variable Selection

Selecting a subset of predictor variables from a larger set (e.g., stepwise selection) is a controversial topic. We can perform stepwise selection (forward, backward, both) using the **stepAIC( )** function from the **MASS** package. **stepAIC( )** performs stepwise model selection by exact AIC.

```
# Stepwise Regression
library(MASS)
fit <- lm(y~x1+x2+x3,data=mydata)
step <- stepAIC(fit, direction="both")
step$anova # display results
```

Alternatively, we can perform all-subsets regression using the **leaps( )** function from the leaps package. In the following code nbest indicates the number of subsets of each size to report. Here, the ten best models will be reported for each subset size (1 predictor, 2 predictors, etc.).

```
# All Subsets Regression
library(leaps)
attach(mydata)
leaps<-regsubsets(y~x1+x2+x3+x4,data=mydata,nbest=10)

# view results
summary(leaps)
# plot a table of models showing variables in each model.
# models are ordered by the selection statistic.
plot(leaps,scale="r2")
# plot statistic by subset size
library(car)
subsets(leaps, statistic="rsq")
```

Creating/editing data objects

```
> y<-c(1,2,3,4,5);y
[1] 1 2 3 4 5
```

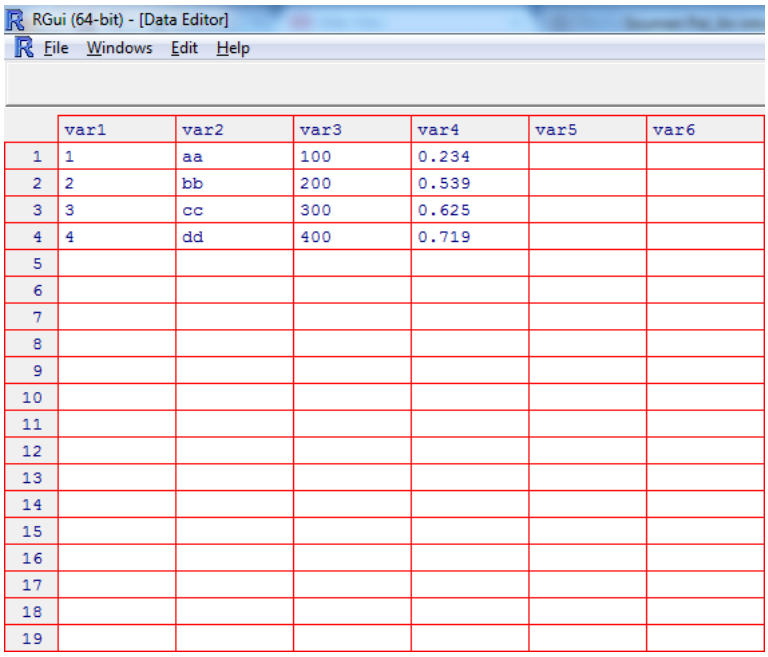
If we want to modify the data object, use *edit()* function and assign it to an object. For example, the following command opens R Editor for editing.

```
> y<-edit(y)
```

If we prefer entering the data.frame in a spreadsheet style data editor, the following command invokes the built-in editor with an empty spreadsheet.

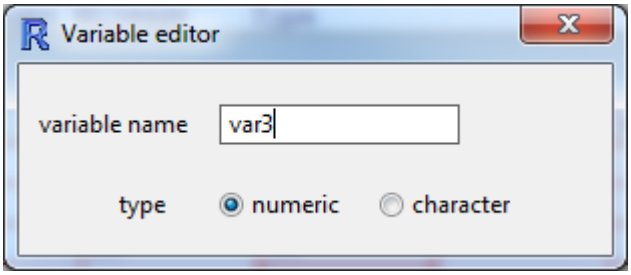
```
> data1<-edit(data.frame())
```

After entering a few data points, it looks like this:



	var1	var2	var3	var4	var5	var6
1	1	aa	100	0.234		
2	2	bb	200	0.539		
3	3	cc	300	0.625		
4	4	dd	400	0.719		
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

We can also change the variable name by clicking once on the cell containing it. Doing so opens a dialog box:



When finished, click  in the upper right corner of the dialog box to return to the Data Editor window. Close the Data Editor to return to the R command window (R Console). Check the result by typing:

```
> data1
```

Reading data from files

When data files are large, it is better to read data from external files rather than entering data through the keyboard. To read data from an external file directly, the external file should be arranged properly.

The first line of the file should have a name for each variable. Each additional line of the file has the values for each variable.

Input file form with names and row labels:

Price	Floor	Area	Rooms	Age	isNew
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	yes
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes

...

By default numeric items (except row labels) are read as numeric variables and non-numeric variables, such as isNew in the example, as factors. This can be changed if necessary.

The function *read.table()* can then be used to read the data frame directly

```
> HousePrice <- read.table("d:/houses.data", header = TRUE)
```

Reading comma delimited data

The following commands can be used for reading comma delimited data into R.

read.csv(filename) — This command reads a .CSV file into R. We need to specify the exact filename with path.

read.csv(file.choose()) — This command reads a .CSV file but the *file.choose()* part opens up an explorer type window that allows us to select a file from our computer. By default, R will take the first row as the variable names.

read.csv(file.choose(), header=T) — This reads a .CSV file, allowing us to select the file, the header is set explicitly. If we change to header=F then the first row will be treated like the rest of the data and not as a label.

Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,

```
>attach(HousePrice)
```

causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice to use the *attach(datafile)* function immediately after reading the *datafile* into R.

Packages

All R functions and datasets are stored in packages. The contents of a package are available only when the package is loaded. This is done to run the codes efficiently without much memory usage. To see which packages are installed at our machine, use the command

```
> library()
```

To load a particular package, use a command like

```
> library(forecast)
```

Users connected to the Internet can use the *install.packages()* and *update.packages()* functions to install and update packages. Use *search()* to display the list of packages that are loaded.

Standard packages

The standard (or base) packages are considered part of the R source code. They contain the basic functions those allow R to work with the datasets and standard statistical and graphical functions. They should be automatically available in any R installation.

Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (https://cran.r-project.org/web/packages/), and other repositories such as Bioconductor (http://www.bioconductor.org/). The collection of available packages changes frequently. As on December 11, 2018, the CRAN package repository contains 13528 available packages.

Getting Help

Complete help files in HTML and PDF forms are available in R. To get help on a particular command/function etc., type *help (command name)*. For example, to get help on function 'mean', type *help(mean)* as shown below

```
> help(mean)
```

This will open the help file with the page containing the description of the function mean. Another way to get help is to use "?" followed by function name. For example,

```
>?mean
```

will open the same window again.

In this lecture note, all R commands and corresponding outputs are given in Courier New font to differentiate from the normal texts. Since R is case-sensitive, i.e. typing *Help(mean),* would generate an error message,

```
> Help(mean)
```

Error in Help(mean) : could not find function "Help"

Further Readings

Various documents are available in https://cran.r-project.org/manuals.html from beginners' level to most advanced level. The following manuals are available in pdf form:

1. An Introduction to R
2. R Data Import/Export
3. R Installation and Administration
4. Writing R Extensions
5. The R language definition
6. R Internals
7. The R Reference Index

REFERENCES

- https://www.cran.r-project.org/
- http://www.gardenersown.co.uk/Education/Lectures/R/index.htm
- Matloff, N. (2009). *The Art of R Programming*.
- Quick R. http://www.statmethods.net/index.html
- W. N. Venables, D. M. Smith and the R Development Core Team. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics Version 2.9.1 (2009-06-26).*

LONG MEMORY TIME SERIES MODELS

Dr. Muhammed Irshad M
Research Associate
Agro-Economic Research Centre
Visva Bharati
Email: 07737312101@visva-bharati.ac.in

Long Memory Time Series Models

Long-memory or long-range dependence in a time series, refers to the slow rate at which correlation between distant time points decay or simply hyperbolic. This concept can be understood through the Autocorrelation Function (ACF), which shows a persistence structure of correlations across distant time lags, which can be explained by the Hurst exponent (*H*), where *H* ∈ (0, 1). This persistence suggests that values in the series are influenced by distant past values. A series is said to have long memory in it when *H* lies between 0.5 and 1. In the case of well-known stationary processes like ARMA, their ACF typically exhibits exponential decay, denoted as $\rho_k \approx |m|^k$ where $|m|$ is less than 1. This property is the characteristic of stationary ARMA (*p, q*) processes. In contrast, long memory processes showcase a notably slower decay rate in their autocorrelation function. This slower decay aligns with $\rho_k \approx Hk^{2d-1}$, where k approaches infinity, and *H* is a constant while *d* represents the long memory parameter.

For a short-memory process

$$\sum_{k=0}^{\infty}|\rho_k| < \infty \ ,$$

For a long-memory process

$$\sum_{k=0}^{\infty}|\rho_k| = \infty \ ,$$

To model such processes, the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model is used. Unlike ARIMA, ARFIMA allows for a fractional differencing parameter, which helps the model to capture the long-memory characteristics by using non-integer values for the differencing order.

Autoregressive Fractionally Integrated Moving Average Model

Long-memory time-series models, known as Autoregressive Fractionally Integrated Moving Average (ARFIMA) models, offer the flexibility to employ non-integer values for the differencing parameter, (Hosking, 1981). These models prove valuable in encapsulating a fundamental trait of time series with long memory: the gradual decay in dependence between two data points, contrasting with the exponential decay observed in the standard ARIMA process. In the integrated section of an ARIMA model, the differencing operator $(1 - L)$ where *L* represents the backshift operator, is traditionally raised to an integer power. The ARFIMA model, on the

other hand, provides the advantage of allowing non-integer differencing (Granger and Joyeux, 1980)

$$(1 - L)^2 = 1 - 2L + L^2$$

Whereas, in ARFIMA process, the power of the model is allowed to take fractional values, with the expression of the term illustrated using the following formal binomial series expansion (Helms, 1984),

$$(1 - L)^d = 1 - dL + \frac{d(d-1)}{2!}L^2 - \cdots = \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j L^j \quad ,$$

$$\text{where } \binom{d}{j} = \frac{d!}{j!(d-j)!}$$

An ARFIMA (p , d , q) process of a time series y_t is defined as

$$\rho(L)(1 - L)^d = \theta(L)e_t \, ,$$

where $\rho(L) = 1 - \rho_1 L - \cdots - \rho_1 L^p$, $\theta(L) = 1 - \theta_1 L - \cdots - \theta q L^q$ are the AR and MA operators sharing no common roots. $(1 - L)^d$ is the fractional differencing operator and e_t are assumed to be independent and identically distributed (i.i.d) with zero mean and constant variance σ^2. For d $\in$ (0, ½), the autocovariance function of this solution satisfies $\lim_{k\to\infty} \rho(k)/[ck^{1-2d}] \to 1$, where c is a constant.

Maximum Likelihood Estimation (MLE), Local Whittle (LW) estimate, Geweke and Porter-Hudak (GPH) test, Sperio test and wavelet technique are among the many approaches for approximating the long-memory parameter d. According to the value of d long memory process can be sub-divided into four groups and they are:

Value of *d*	**Description**
$d \in \left(-\frac{1}{2}, 0\right)$	Intermediate Memory and Anti-persistence
$d = 0$	White noise (Short-Memory)
$d \in \left(0\, ,\, \frac{1}{2}\right)$	Stationary and Persistent Long Memory
$d \in \left(\frac{1}{2}\, ,\, 1\right)$	Non-stationary and Persistent Long Memory

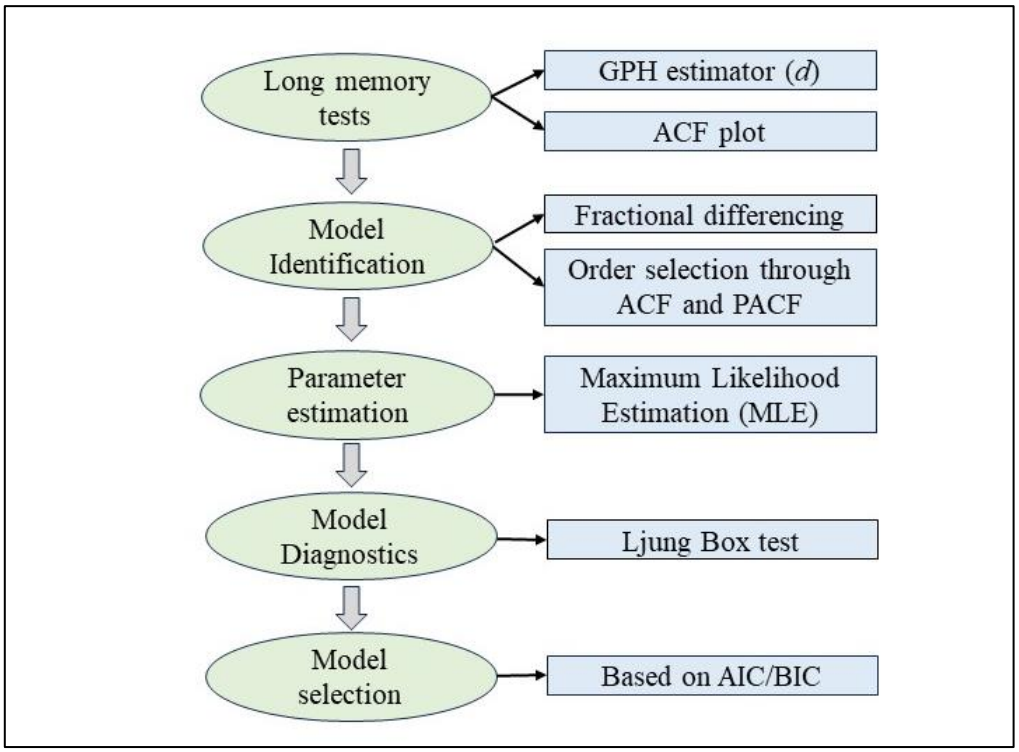


Figure: Steps in ARFIMA model fitting

R Code for Fitting ARFIMA Model

1. Install and Load Required Packages

```
install.packages("forecast")
library(forecast)
```

2. Load Time Series Data

Example using built-in dataset "AirPassengers"

```
data(AirPassengers)

ts_data <- log(AirPassengers) # Log transformation stabilizes variance.

plot(ts_data, main="Log Transformed AirPassengers Data") #Time series plot
```

3. Fit ARFIMA Model Automatically

```
model <- arfima(ts_data, drange = c(0,0.5))

summary(model)
```

4. Diagnostic Checking

```
checkresiduals(model) # Ensure the residuals are white noise
```

REFERENCES

- Bhardwaj, G., and Swanson, N. R. (2006). An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series. *Journal of Econometrics*, *131*(1), 539–578. https://doi.org/10.1016/j.jeconom.2005.01.016

- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (4th ed.). John Wiley & Sons.

- Brock, W.A., Dechert, W. and Scheinkman, J. (1996). A Test for Independence Basedon the Correlation Dimension. *Econometric Reviews*, *15*: 197–235.

- Geweke, J., and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis, 4*, 221-238.

- Irshad, M. M., Sarkar, K. A., Dhakre, D. S., and Bhattacharaya, D. (2025). *fracARMA: Fractionally integrated ARMA model* (R package version 0.1.0). https://CRAN.R-project.org/package=fracARMA

- Irshad, M. M., Sarkar, K. A., Dhakre, D. S., and Bhattacharya, D. (2024). Forecasting of monthly cardamom price using long memory time series modelling technique. *International Journal of Statistics and Applied Mathematics*, 9*(6)*, 121–126. https://doi.org/10.22271/maths.2024.v9.i6b.1906

HYBRID TIME SERIES MODELS

Dr. Muhammed Irshad M
Research Associate
Agro-Economic Research Centre
Visva Bharati
Email: 07737312101@visva-bharati.ac.in

Hybrid Time Series Models

By combining linear and nonlinear components simultaneously, hybrid forecasting methods solve the constraints of both approaches and enhance time series analysis. The difficulty of deciding whether a time series is linear or nonlinear as well as the reality that real-world data can show both patterns drive these techniques. Combining linear and nonlinear models helps hybrid methods more successfully capture various patterns. Usually, the procedure consists of two phases: first, linear models are used to identify linear patterns; subsequently, nonlinear models on the residuals from the linear models are used to handle any residual nonlinearity.

Traditional models, despite not considering the second-order moment or conditional variance, provide an excellent starting point for studying time series data. They operate under the assumption of stationarity is a key assumption in many time series models, meaning that the statistical properties (mean, variance, autocorrelation, etc.) of the series do not change over time. The autocorrelation structure of the residuals will be assessed using the Ljung-Box test and BDS test. If the auxiliary tests provide substantial evidence of nonlinearity and autocorrelation in the residuals, it is become essential to integrate one nonlinear component into the model so that all information in the data is extracted.

For example: **Hybrid ARFIMA-GARCH Model**

Suppose an ARFIMA model is built using a long-memory time series dataset, which captures the linear component and the long-range dependence present in the data. After fitting the ARFIMA model, the residuals are analysed to verify whether the model has adequately captured the underlying structure of the series. However, in many economic and financial time series, the variance of the residuals may not remain constant over time. Instead, the residuals may exhibit periods of high volatility followed by periods of low volatility, a phenomenon known as volatility clustering. This behaviour indicates the presence of Autoregressive Conditional Heteroskedasticity, where the variance of the error terms depends on past squared residuals.

To detect these effects, the residuals obtained from the initial ARFIMA model are examined and their squared residuals are analysed using statistical tests such as the ARCH–LM test. If the null hypothesis of no ARCH effect is rejected, it implies that the residuals exhibit heteroskedasticity, meaning the variance changes over time rather than remaining constant. In such situations, it becomes necessary to model the conditional variance explicitly using volatility models such as the Generalized Autoregressive Conditional Heteroskedasticity model. By

combining ARFIMA for the mean equation and GARCH for the variance equation, the resulting ARFIMA–GARCH framework is able to capture both long-memory in the mean and time-varying volatility in the variance, thereby improving the overall modelling and forecasting performance of the time series.

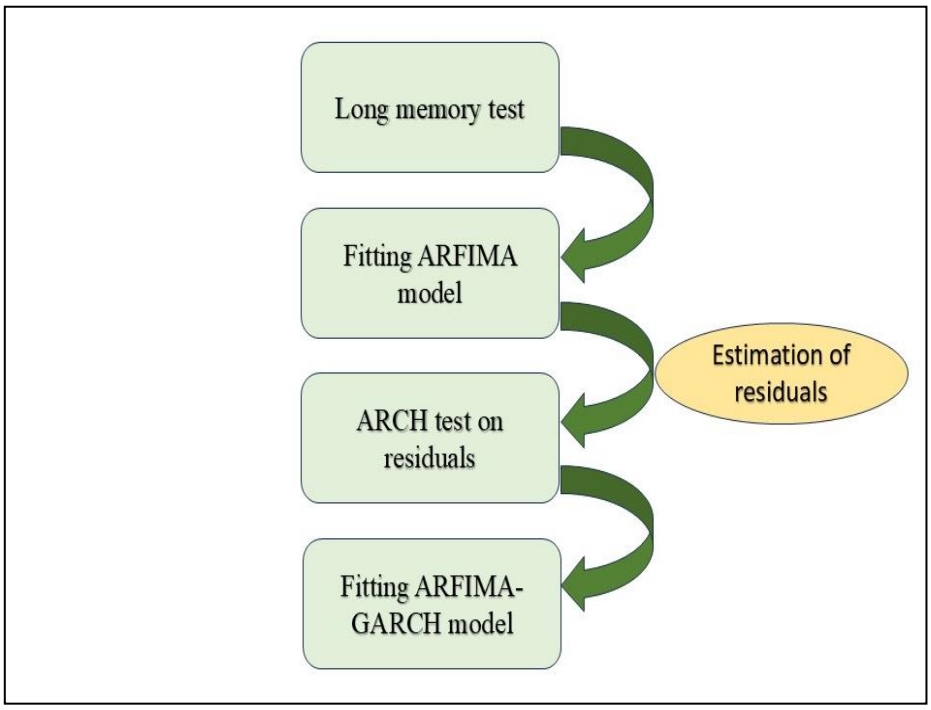


Figure: Steps in building an ARFIMA-GARCH model

R Code for Fitting Hybrid ARFIMA-GARCH Model

1. Install and Load Packages

```
install.packages("rugarch")

library(rugarch)
```

2. Generate Example Time Series Data

```
set.seed(123)

# simulate 500 observations

data <- rnorm(500)

# convert to time series

ts_data <- ts(data)
```

```
plot(ts_data)
```

3. Specify ARFIMA–GARCH Model

```
spec <- ugarchspec(
        variance.model = list(model = "sGARCH", garchOrder = c(1,1) ),      mean.model = list(armaOrder = c(1,1), include.mean = TRUE, arfima = TRUE ),distribution.model = "norm")
```

4. Fit the Model

```
fit <- ugarchfit(spec = spec, data = ts_data)

show(fit)
```

Hybrid ARIMA-ANN Model

The ARIMA–ANN Hybrid Model is used to capture both linear and nonlinear patterns present in a time series. In this approach, the ARIMA model is first applied to the data to model the linear structure of the series. After fitting the ARIMA model, the residuals are obtained, which are tested using diagnostic tests (Box test and BDS test) may still contain nonlinear relationships that the ARIMA model cannot capture. These residuals are then modeled using an Artificial Neural Network to identify the nonlinear patterns in the data. Finally, the forecasts from the ARIMA model (linear component) and the ANN model (nonlinear component) are combined to produce the final forecast. This hybrid approach improves forecasting accuracy by integrating the strengths of both statistical and machine learning techniques.

```
library(forecast)

library(tseries

set.seed(123)

# simulate 500 observations

data <- rnorm(500)

# convert to time series

ts_data <- ts(data)
```

ARIMA model

```
acf(ts)  # ACF plot
```

```
pacf(ts) # PACF plot
adf_test<-adf.test(ts) # stationary plot
train<-ts[c(1:(length(ts)*0.9))] #train data
test<-ts[-c(1:(length(ts)*0.9))] #test data
model<-auto.arima(train) #model
Fitted<-model$fitted  ## fitted value
Forecast<-forecast(model, h=length(test)) # future forecast
accuracy(Forecast, x=test) ## accuarcy measure
plot(Forecast) ##plot
```

Residual Check

```
residual_arima<-model$residuals
Box.test(residual_arima)
checkresiduals(model)
bds.test(residual_arima)
```

ANN model

```
ANN<-nnetar(residual_arima, p=4)
ANN_fitted<-ANN$fitted
ANN_forecast<-forecast(ANN, h=length(test))
```

Residual Check

```
residual_ANN<-ANN$residuals
Box.test(residual_ANN)
checkresiduals(ANN)
```

##Final Forecast

```
Final_fitted<-Fitted+ANN_fitted
Final_forecast<-Forecast$mean+ANN_forecast$mean
```

Accuracy

```
MAPE_train<-MLmetrics::MAPE(Final_fitted[-c(1:4)], train[-c(1:4)])

MAPE_test<-MLmetrics::MAPE(Final_forecast, test)

RMSE_train<-MLmetrics::RMSE(Final_fitted[-c(1:4)], train[-c(1:4)])

RMSE_test<-MLmetrics::RMSE(Final_forecast, test)
```

REFERENCES

- Ali, S. A., and Ahmad, A. (2019). Forecasting MSW generation using artificial neural network time series model: A study from metropolitan city. *SN Applied Sciences*, *1*(11), 1338. https://doi.org/10.1007/s42452-019-1382-7.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, *31*, 307-327.
- Douc, R., Roueff, F. and Soulier, P. (2008). On the existence of some ARCH (∞) processes. *Stochastic Processes and Applications*, *118*, 755-761.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*, 987-1007.
- Irshad, M. M., Sarkar, K. A., Dhakre, D. S., and Bhattacharya, D. (2024). Comparative analysis of statistical model and machine learning algorithms in forecasting black pepper price of Kerala. *Biological Forum – An International Journal*, 16(8), 63–68.

ARTIFICIAL INTELLIGENCE MODELS

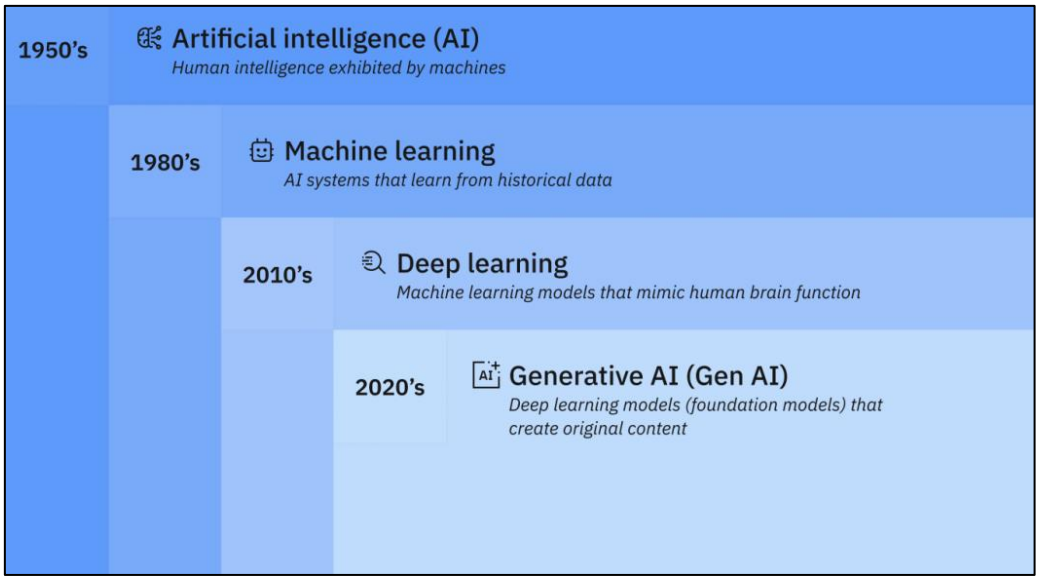
Dr. Muhammed Irshad M
Research Associate
Agro-Economic Research Centre
Visva Bharati
Email: 07737312101@visva-bharati.ac.in

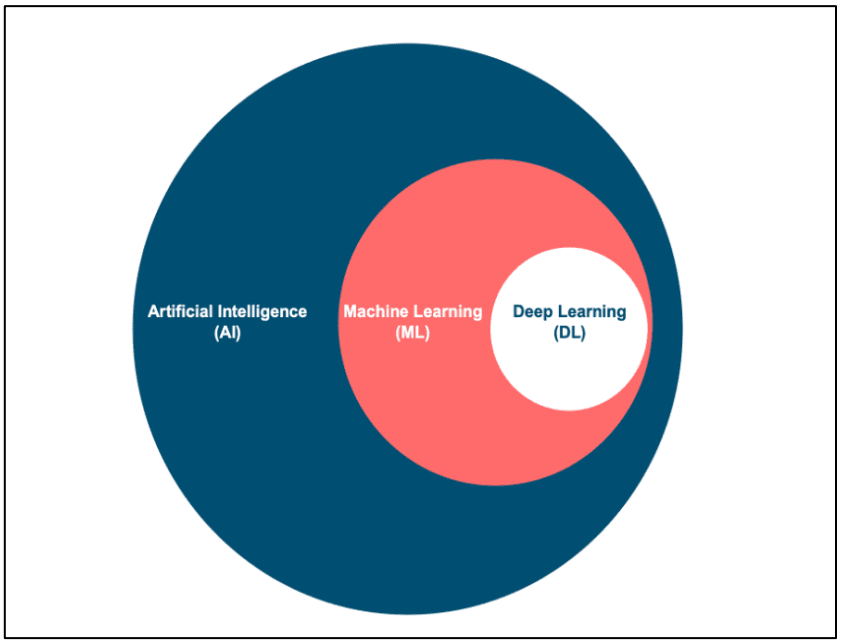
What is AI?

Artificial Intelligence (AI) seeks to imitate human brain to execute basic to sophisticated activities, including natural language interpretation, voice recognition, and decision-making etc. Applications and devices equipped with AI can see and identify objects. They can understand and respond to human language. They can learn from new information and experience. They can make detailed recommendations to users and experts. They can act independently, replacing the need for human intelligence or intervention (eg: self-driving car)

But in 2025, most AI researchers and practitioners and most AI-related headlines are focused on breakthroughs in generative AI (gen AI), a technology that can create original text, images, video and other content. To fully understand generative AI, it's important to first understand the technologies on which generative AI tools are built: Machine Learning (ML) and Deep Learning (ML).

A simple way to think about AI is as a series of nested or derivative concepts that have emerged over more than 70 years.





Venn diagram of Artificial Intelligence derivatives

Machine Learning

Directly underneath AI, we have machine learning, which involves creating models by training an algorithm to make predictions or decisions based on data. It encompasses a broad range of techniques that enable computers to learn from and make inferences based on data without being explicitly programmed for specific tasks.

Machine Learning Methods

Machine learning models fall into three primary categories.

Supervised Learning

It is defined by its use of labelled datasets to train algorithms for classifying data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids over fitting or under fitting. Supervised learning helps organizations solve a variety of real-world problems.

eg: Neural networks, Naïve Bayes, Linear regression, Logistic regression, Random forest, and Support vector machines

Unsupervised Learning

It uses machine learning algorithms to analyse and cluster unlabelled datasets (subsets called clusters). These algorithms discover hidden patterns or data groupings without the need for human intervention.

Unsupervised learning's ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction.

Eg: Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Neural Networks, K-Means Clustering, and Probabilistic Clustering etc.

Semi-Supervised Learning

It is balance between supervised and unsupervised learning. During training, it uses a smaller labelled data set to guide classification and feature extraction from a larger, unlabelled data set. Semi-supervised learning can solve the problem of not having enough labelled data for a supervised learning algorithm. It also helps if it's too costly to label enough data.

Reinforcement Learning

It is a machine learning model that is similar to supervised learning, but the algorithm isn't trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

Machine Learning Models

There are many types of machine learning techniques or algorithms, including Artificial Neural Networks, Decision Trees, Random Forest, Support Vector Machines (SVMs), K-Nearest Neighbour (KNN), clustering and more. Each of these approaches is suited to different kinds of problems and data.

Artificial Neural Network

An Artificial Neural Network (ANN) is a non-linear computational model inspired by the structure and functioning of the human brain, designed for pattern recognition and information processing tasks. Most important feature of ANN is that it doesn't require any assumptions to be satisfied in the pre-processing stage of data, rather it is data driven model and non-parametric model. Comprising interconnected nodes organized into layers. The architecture of an ANN

primarily consists of three layers: the input layer, a single hidden layer, and the output layer. Single hidden layer feed forward network is the most popular for time series modelling and forecasting.

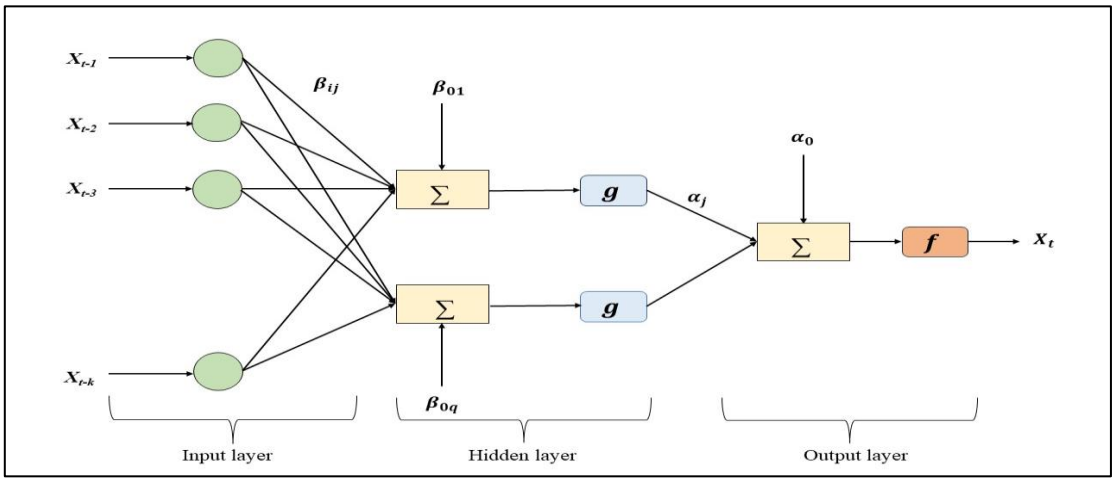
ANN model performs a nonlinear functional mapping between the input and output which characterized by a network of three layers of simple processing units connected by acyclic links. The relationship between the output (X_t) and the inputs (X_{t-1}, X_{t-2}..., X_{t-p}) can be mathematically represented as follows:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-p}, \boldsymbol{w}) + \varepsilon_t$$

where $\boldsymbol{w}$ is the vector of all parameters and f is a function of network structure and connection weights and ε_t is the error. The relationship between the output (X_t) and the inputs (X_{t-1}, X_{t-2}..., X_{t-p}) can be mathematically represented as follows:

$$X_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j g + (\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} X_{t-p}) + \varepsilon_t$$

where α_j (j=1,2, …, q) and β_{ij}(i=1, 2…, p, j=1, 2…, q) are the model parameters often called as connection weights, p is the number of input nodes and q is the number of hidden nodes.



Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity. Most commonly used activation function is logistic function which is often used as the hidden layer transfer function, i.e.,

$$g(X) = \frac{1}{(1+exp(-X_t))}$$

The selection of appropriate number of hidden nodes as well as optimum number of lagged observations p for input vector is important in ANN modelling for determination of the autocorrelation structure present in the time series. Though there are no established theory available for the selection of p and q, various training algorithms have been developed for the

determination of the optimal values of p and q. The objective of training is to minimize the error function that measures the misfit between the predicted value and the actual value.

Support Vector Machine

A Support Vector Machine (SVM) is a robust machine learning algorithm used for both classification and regression tasks. It works by finding the optimal hyper plane that best separates data points of different classes in a given dataset, maximizing the margin between the classes. The closest points to the hyper plane, known as support vectors, are critical in defining this boundary. In cases where data is not linearly separable, SVM utilizes the "kernel trick" to transform the data into a higher-dimensional space, making it easier to find a separating hyper plane. SVM is widely applied in various fields such as image recognition, text categorization, and bioinformatics, owing to its effectiveness in handling complex and high-dimensional data.

K-Nearest Neighbours

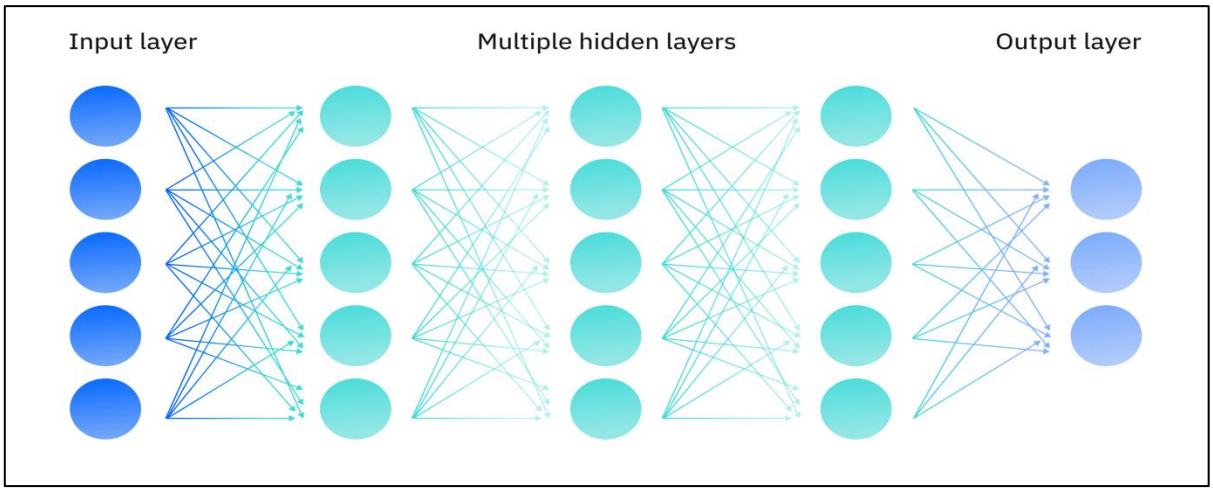
The K-Nearest Neighbours (KNN) algorithm is a simple, intuitive method used in both classification and regression tasks in machine learning. It works by comparing a new data point with existing data points to determine its category or value. In KNN, the "K" represents the number of nearest neighbours to consider when making the decision. For classification, the algorithm assigns the new data point to the majority class among its nearest neighbours, while for regression; it calculates the average value of the neighbours. KNN is a non-parametric, lazy learning algorithm that doesn't require explicit training, making it easy to implement and use for various applications such as image recognition, recommendation systems, and anomaly detection

Deep Learning

Deep learning is a subset of machine learning that uses multi-layered neural networks, called deep neural networks that more closely simulate the complex decision-making power of the human brain.

Deep neural networks include an input layer, at least three but usually hundreds of hidden layers, and an output layer, unlike neural networks used in classic machine learning models, which usually have only one or two hidden layers. These multiple layers enable unsupervised learning: they can automate the extraction of features from large, unlabelled and unstructured data sets, and make their own predictions about what the data represents.

Because deep learning doesn’t require human intervention, it enables machine learning at a tremendous scale. It is well suited to Natural Language Processing (NLP), computer vision, and other tasks that involve the fast, accurate identification complex patterns and relationships in large amounts of data. Some form of deep learning powers most of the Artificial Intelligence (AI) applications in our lives today.

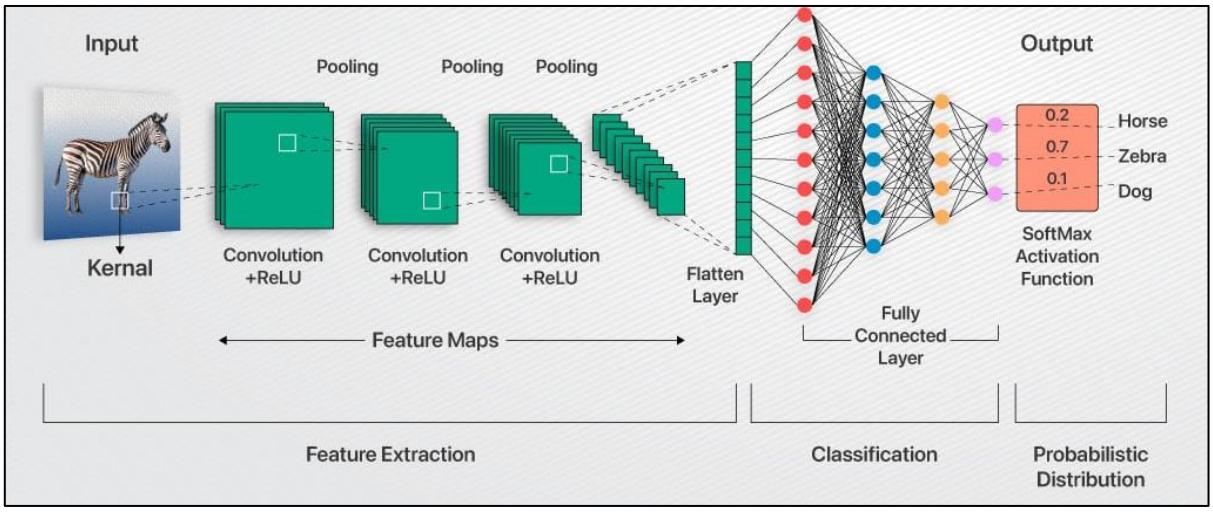


Structure of a deep learning model

Deep Learning Models

Convolutional Neural Networks

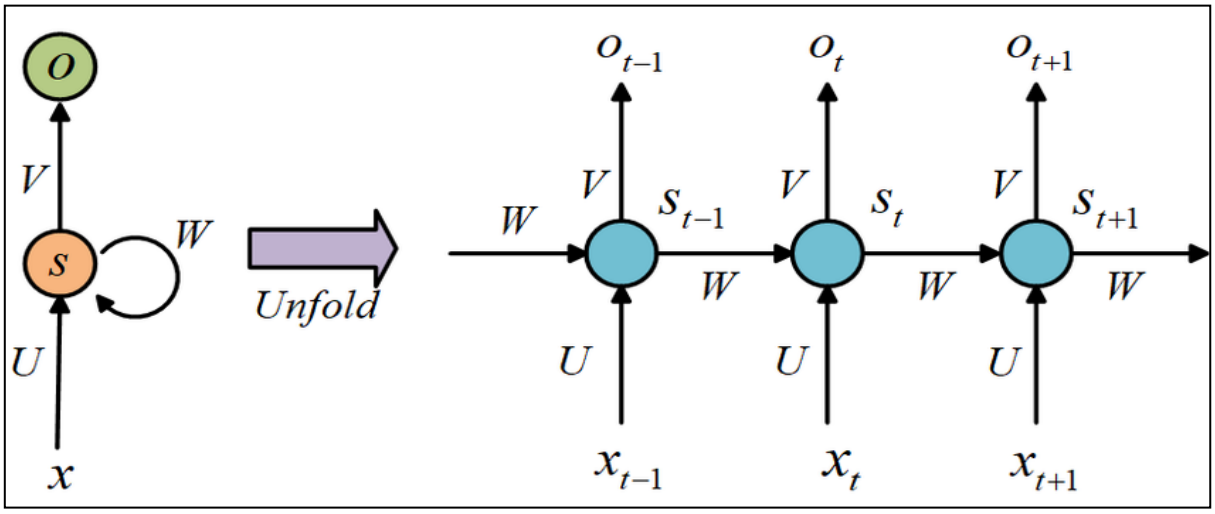
Convolutional Neural Networks (CNNs) are a specialized type of artificial neural network designed primarily for image processing tasks. They consist of multiple layers, including convolutional layers that automatically learn to detect features such as edges and textures from the input images. Pooling layers then reduce the spatial dimensions while retaining important information, and fully connected layers combine the features to make final predictions. Activation functions like ReLU introduce non-linearity, helping the network learn complex patterns. CNNs have been pivotal in advancing computer vision and are used in various applications like image recognition, object detection, and facial recognition.



Structure of a Convolutional Neural Network Model (Analytics LABS)

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks specialized for processing sequential data, such as time series, natural language, and speech. Unlike traditional neural networks, RNNs have connections that form directed cycles, allowing them to maintain a memory of previous inputs and capture temporal dependencies. This makes them particularly effective for tasks where context and order matter. However, standard RNNs can struggle with long-term dependencies due to vanishing gradient issues, which is why advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) were developed to address these challenges. RNNs are widely used in applications like language modelling, translation, and speech recognition.



Structure of a RNN model

Long Short-Term Memory Model

Hochreiter and Schmidhuber (1997) developed the Long Short-Term Memory (LSTM) neural network model to better handle long range dependence in the data. LSTM is an updated and modified version of RNN. The LSTM model worked really well with financial high-frequency time series data. It fixed issues that other models had, like vanishing gradient and gradient disappearance. In traditional RNNs, the gradients close to zero is multiplied many times which make the gradients much close to zero hence the problem of gradient vanishing or gradient disappearance problem arises. The primary function of the LSTM model is to retain essential information while discarding superfluous input. It does these using three essential memory components: the input gate, the output gate, and the forget gate. There are many LSTM models out there, but the one we used in this study was created by Hochreiter and Schmidhuber. At first, the forget gate gets rid of information which is not needed. After that, the model sorts out useful information using the input gate and a certain probability. Lastly, useful data is taken out through the output gate and sent to the next LSTM unit in the chain. Choosing the activation function is a

very important part of the LSTM process because it affects how the model handles data. This study used the standard sigmoid function and the hyperbolic tangent (tanh) function as activation functions. These functions shaped the behaviour of the LSTM units and made sure that the network processed information properly. The LSTM process can be summarized in five steps.

Step 1: The output value of the previous unit and input value of the current unit are integrated into the forget gate. The output value of the forget gate is calculated as,

$$f_t = \sigma\{W_f * (h_{t-1} * x_t)\} + b_f \quad ,$$

where W_f is the weight of forget gate, b_f is the bias, x_t is the input value and h_{t-1} is the output value of the prior unit.

Step 2: The output value of the prior unit and the input value of the current time are incorporated into the input gate. The output value and candidate cell state values are computed as,

$$i_t = \sigma\{W_i * (h_{t-1} * x_t)\} + b_i$$

$$C'_t = tanh\{W_c * (h_{t-1} * x_t)\} + b_c$$

where W_i and b_i are the weight and bias of the input gate respectively, W_c and b_c are the weight and bias of the candidate input respectively.

Step 3: Updating of the current cell is done using the formula,

$$C_t = f_t * C_{t-1} + i_t * C'_t$$

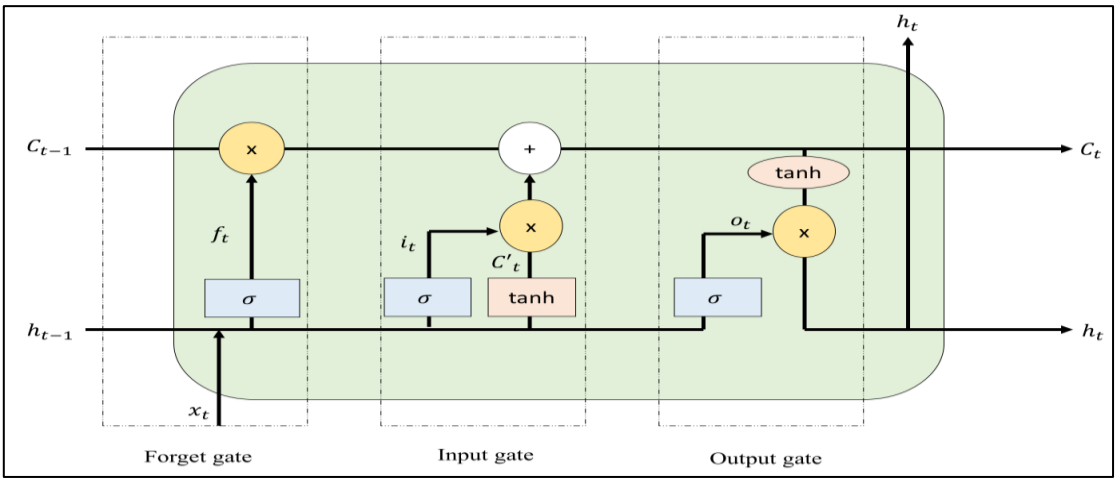
Step 4: the output gate takes h_{t-1} and x_t as input values, and its output is calculated using the formula

$$o_t = \sigma\{W_o * (h_{t-1} * x_t)\} + b_o$$

where W_o and b_o are the weight and bias of this gate respectively.

Step 5: The final output of the LSTM unit is generated by calculating the output of the output gate and the cell state, as follows,

$$h_t = o_t * tanh(C_t)$$



Structure of long short-term memory model

Gated Recurrent Unit Model

The Gated Recurrent Unit (GRU) was proposed by Cho et al. (2014) as an improvement of the RNN architecture. It was designed to mitigate some constraints of conventional RNNs, including the vanishing gradient problem, by integrating gating mechanisms like to those seen in LSTM networks. The GRU streamlines the LSTM design by combining forget and input gates into a single update gate and using a reset gate to regulate the information flow inside the network. This streamlined design resulted in expedited training durations and less computing complexity, while still enabling GRUs to proficiently capture long-term relationships in sequential data.

The activation h_t of the GRU at time t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\widetilde{h_t}$

$$h_t^j = (i - z_t^j)h_{t-1}^j + z_t^j h_t^j$$

where *update gate* z_t^j decides how much the unit updates its activation or content. The update gate is computed by,

$$z_t^j = \sigma(W_z \, x_t + U_z \, h_{t-1})^j$$

The candidate activation $\tilde{h}_t^j$ is computed as in Bahdanau et. al, (2014),

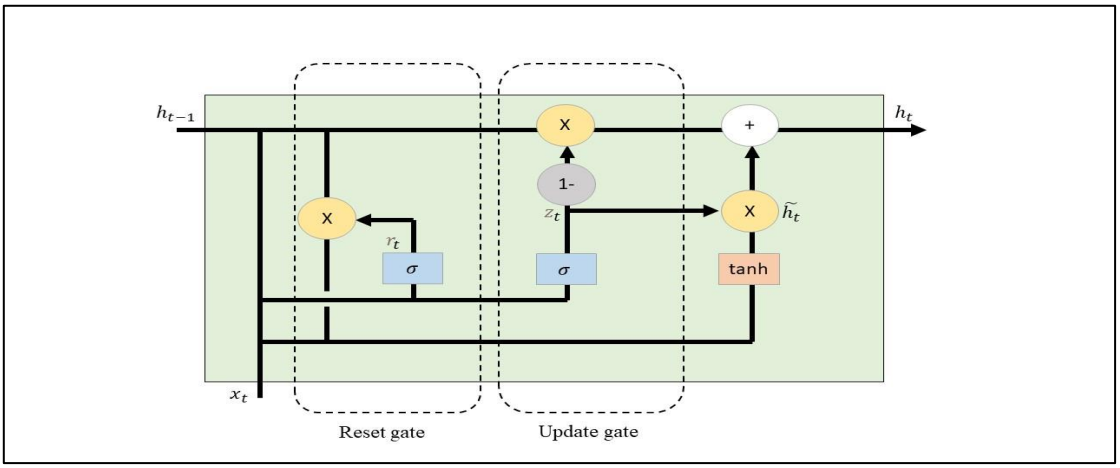
$$\tilde{h}_t^j = \tanh(Wx_t + U\,(r_t \odot h_{t-1}))^j$$

where r_t is the *reset gate* and $\odot$ is the element wise multiplication. When off (r_t close to 0) the reset gate effectively makes the unit act as it is reading the first symbol of an input sequence, allowing it to *forget* the previously computed state.

The *reset gate* r_i^j is computed similarly to the *update gate*:

$$r_i^j = \sigma(W_r \, x_t + U_r \, h_{t-1})^j$$

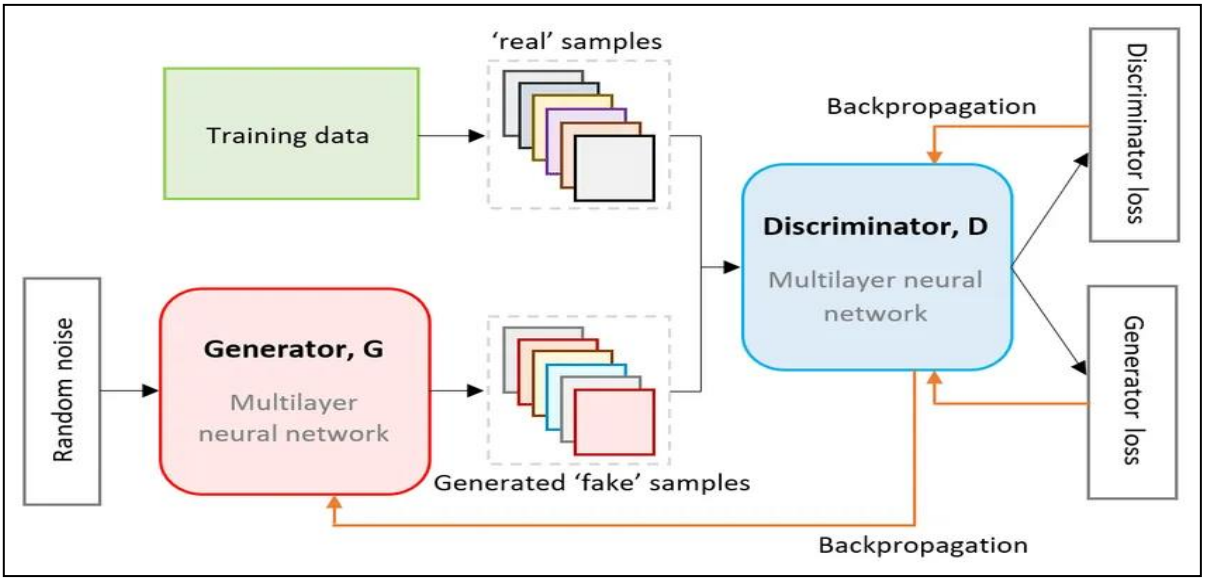
The Figure below represents the graphical illustration of GRU.



Structure of gated recurrent unit model

Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of deep learning models comprising two neural networks: the generator and the discriminator. These networks are set up in a competitive framework, where the generator aims to create realistic data, such as images, by learning from a dataset, while the discriminator attempts to distinguish between real and generated data. This adversarial process continues iteratively, with the generator improving its output to deceive the discriminator. GANs have been widely used for applications like image generation, style transfer, and even creating photorealistic artworks. Their ability to generate high-quality synthetic data has opened new avenues in various fields, from art to healthcare.



Structure of Generative Adversarial Networks

Transformer Models

Transformer models are a class of deep learning models that have revolutionized Natural Language Processing (NLP) by leveraging self-attention mechanisms to process input data in parallel rather than sequentially. This architecture allows transformers to capture long-range dependencies and contextual relationships within the data, making them highly efficient and effective for tasks like language translation, sentiment analysis, and text generation. Notable transformer-based models include BERT, which excels in understanding context, and GPT, which generates coherent and contextually relevant text. Their versatility and performance have made transformers the go-to models for a wide range of NLP applications.

Autoencoders

Autoencoders are used for tasks like dimensionality reduction and unsupervised learning. They learn to compress input data into a lower-dimensional representation and then reconstruct the output, capturing the essential features.

Deep Reinforcement Learning Models

These models combine deep learning with reinforcement learning principles. Agents learn to make decisions by interacting with their environment. AlphaGo, which famously defeated the world champion in the game of Go, is a notable example.

Graph Neural Networks

GNNs are designed to process data represented as graphs, making them ideal for tasks involving social networks, knowledge graphs, and recommendation systems.

Deep Learning Methods

- Semi-supervised learning, which combines supervised and unsupervised learning by using both labelled and unlabelled data to train AI models for classification and regression tasks.
- Self-supervised learning, which generates implicit labels from unstructured data, rather than relying on labelled data sets for supervisory signals.
- Reinforcement learning, which learns by trial-and-error and reward functions rather than by extracting information from hidden patterns.
- Transfer learning, in which knowledge gained through one task or data set is used to improve model performance on another related task or different data set.

Generative AI

Generative AI, sometimes called "gen AI", refers to deep learning models that can create complex original content such as long-form text, high-quality images, realistic video or audio and more in response to a user's prompt or request.

At a high level, generative models encode a simplified representation of their training data, and then draw from that representation to create new work that's similar, but not identical, to the original data.

Generative models have been used for years in statistics to analyse numerical data. But over the last decade, they evolved to analyse and generate more complex data types. This evolution coincided with the emergence of three sophisticated deep learning model types

- Variational autoencoders or VAEs, which were introduced in 2013, and enabled models that could generate multiple variations of content in response to a prompt or instruction.
- Diffusion models, first seen in 2014, which add "noise" to images until they are unrecognizable, and then remove the noise to generate original images in response to prompts.
- Transformers (also called transformer models), which are trained on sequenced data to generate extended sequences of content (such as words in sentences, shapes in an image, frames of a video or commands in software code). Transformers are at the core of most of today's headline-making generative AI tools, including ChatGPT and GPT-4, Copilot, BERT, Bard and Midjourney.

Weak AI vs. Strong AI

In order to contextualize the use of AI at various levels of complexity and sophistication, researchers have defined several types of AI that refer to its level of sophistication:

Weak AI: Also known as "narrow AI," defines AI systems designed to perform a specific task or a set of tasks. Examples might include "smart" voice assistant apps, such as Amazon's Alexa, Apple's Siri, a social media chatbot or the autonomous vehicles promised by Tesla.

Strong AI: Also known as "artificial general intelligence" (AGI) or "general AI," possess the ability to understand, learn and apply knowledge across a wide range of tasks at a level equal to or surpassing human intelligence. This level of AI is currently theoretical and no known AI systems approach this level of sophistication. Researchers argue that if AGI is even possible, it requires major increases in computing power. Despite recent advances in AI development, self-aware AI systems of science fiction remain firmly in that realm.

REFERENCES

- Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., and Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors, 21*(11), 3758. https://doi.org/10.3390/s21113758

- Kamilaris, A., and Prenafeta-Boldú, F. X. (2020). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2020). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. https://doi.org/10.3390/s18082674

- van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

- Wolfert, S., Ge, L., Verdouw, C., and Bogaardt, M. J. (2020). Big data in smart farming: A review. *Agricultural Systems, 153*, 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

- Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2021). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture, 151*, 61–69.

- Pathak, H. S., Brown, P., and Best, T. (2021). A systematic literature review of machine learning applications for sustainable agriculture. *Computers and Electronics in Agriculture, 153*, 69–81.

- Kaur, P., Sharma, M., and Mittal, M. (2022). Big data and machine learning in agriculture: A survey. *Computers and Electronics in Agriculture, 198*, 107017.

- Elavarasan, D., and Vincent, D. R. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access, 8*, 86886–86901. https://doi.org/10.1109/ACCESS.2020.2992480

- Shahhosseini, M., Hu, G., and Archontoulis, S. V. (2021). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science, 12*, 709819. https://doi.org/10.3389/fpls.2021.709819

BASICS OF ECONOMETRICS

Dr. Achiransu Acharyya
Hony. Deputy Director
Agro-Economic Research Centre
Visva Bharati
Email: dy.dir.aerc@visva-bharati.ac.in

What is Econometrics?

Econometrics refers to the application of economic theory and statistical techniques for the purpose of testing hypotheses and estimating and forecasting economic phenomena. Literally interpreted, econometrics means "economic measurement." Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations: Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results. Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference. Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena. Econometrics is concerned with the empirical determination of economic laws.

Basic Econometrics

The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him. Econometricians are a positive help in trying to dispel the poor public image of economics (quantitative or otherwise) as a subject in which empty boxes are opened by assuming the existence of can-openers to reveal contents which any ten economists will interpret in 11 ways. The method of econometric research aims, essentially, at a conjunction of economic theory and empirical measurement, using the theory and techniques of statistical inference as a bridge.

Objectives

I. Applications of economic theory need a responsible understanding of economic relationships and econometric methods.
II. Econometrics theory thus becomes a very powerful tool for understanding the applied economic relationships and for meaningful research in economics.
III. In this unit, we learn the basic theory of econometrics and the relevant applications of the methods.

Methodology of Econometrics

Broadly speaking, traditional econometric methodology proceeds along the following lines:

i. Statement of theory or hypothesis.
ii. Specification of the mathematical model of the theory
iii. Specification of the statistical, or econometric, model
iv. Obtaining the data
v. Estimation of the parameters of the econometric model
vi. Hypothesis testing
vii. Forecasting or prediction
viii. Using the model for control or policy purposes.

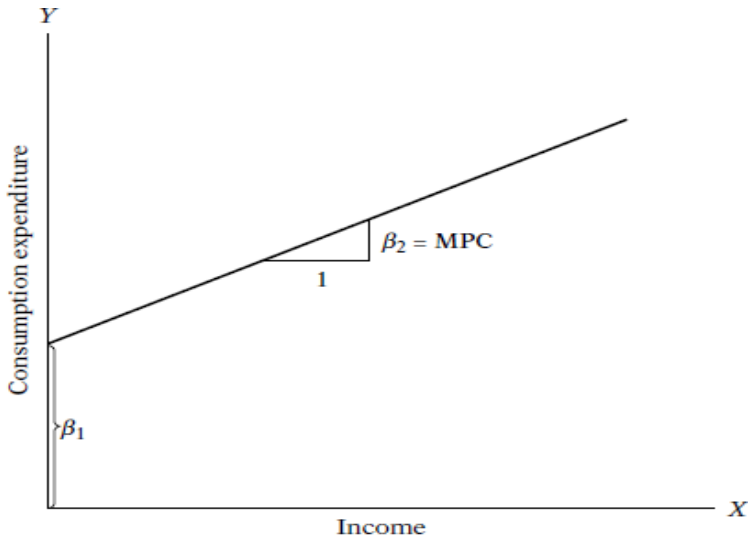
To illustrate the preceding steps, let us consider the well-known Keynesian theory of consumption:

Statement of theory or Hypothesis

Keynes postulated that the marginal propensity to consume (MPC), the rate of change of consumption for a unit change in income, is greater than zero but less than one. i.e., 0 < MPC< 1.

Specification of the Mathematical Model of Consumption

Keynes postulated a positive relationship between consumption and income.



Keynesian consumption function.

The slope of the coefficient β_2 measures the MPC.

Keynesian consumption function

$$Y=\beta_1 + \beta_2 x \qquad O < \beta_2 < 1$$

Y = Consumption expenditure

X = Income

β_1 and β_2 are known as model parameters and are, respectively, the intercept and slope of the coefficient.

It shows a precise and consistent relationship between consumption and income. The slope of the coefficient β_2 measures the MPC.

The equation states that consumption is linearly related to income (an Example of a mathematical model of the relationship between consumption and income, called the consumption function in economics).

A single equation is a single equation model, and a model with more than one equation is a multiple equation model.

Specification of the econometric model of consumption

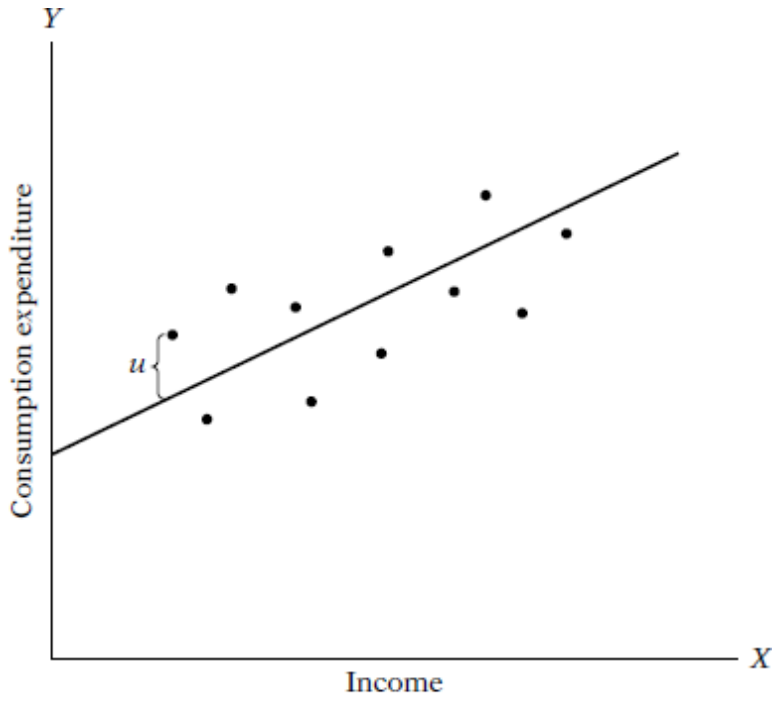
The inexact relationship between economic variables, the econometrician would modify the deterministic consumption function as follows:

$$Y = \beta_1 + \beta_2 x + U$$

This equation is an example of the econometric model. More technically, it is an example of a linear regression model.

This may represent all the factors that affect consumption but are not explicitly considered.

The econometric consumption function hypothesises that the dependent variable Y (consumption) is linearly related to the explanatory variable X (Income).



Econometric model of the Keynesian consumption function.

Q: Why do inexact (not exact) relationships exist?

A: Because, in addition to income, other variables affect consumption expenditure. For example, family, family members' ages, religion, and other factors are likely to influence consumption.

Original Data

To obtain the numerical values of β_1 & β_2, we need data.

{PCE: Personal consumption expenditure)

The Y variable in the graph is the average PCE, and the X variable is a measure of aggregate income.

Note: MPC: Average change in consumption over change in real income

Estimation of the Econometric Model

Regression analysis is the main statistical technique used to obtain estimates. The estimated consumption function

$\hat{Y}=\hat{\beta}_1+\hat{\beta}_2 x_i$

$\hat{Y}$ = Estimate of Y. The estimated consumption function (i.e., regression line).

Regression Analysis is used to obtain estimates.

Hypothesis Testing

Keynes expected the MPC to be positive but less than 1.

Confirmation or refutation of economic theories based on sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing)

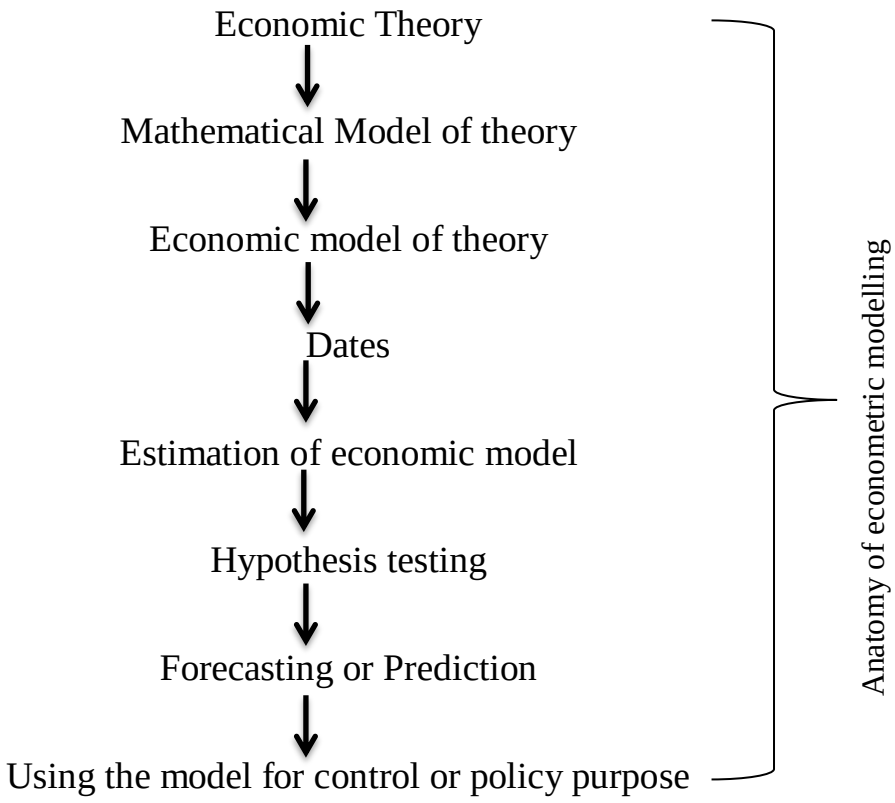
Forecasting or Prediction

If the chosen model does refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the dependent, or forecast, variable Y based on known or expected future value(s) of the explanatory, or predictor variable X. Macroeconomic theory shows that the change in income following a change in investment expenditure is given by the income multiplier M.

$$M = \frac{1}{1 - MPC}$$

The quantitative estimate of MPC provides valuable information for policy purposes. Knowing MPC, one can predict the future course of income, consumption expenditure, and employment following a change in the government's fiscal policies.

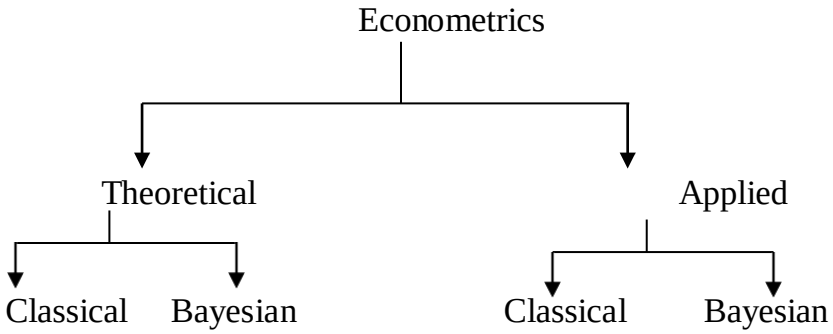
Use of the Model for control or Policy purposes



Note:

- Milton Friedman has developed a model of consumption theory, the permanent income hypothesis.
- Robert Hall has developed a model of consumption as the life cycle permanent income hypothesis.

Types of Econometrics



Theoretical economics is concerned with the development of appropriate methods of measuring economic relationships specified by economic models.

Applied economics uses the tools of theoretical economics to study some special fields of economics and business, such as the production function, etc.

Summary and Conclusions

Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves study for the following reasons.

Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative relationship between the price and the quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell how much the quantity will change in response to a given change in the price of the commodity. It is the econometrician's job to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory.

The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics, as noted previously, is mainly interested in the empirical verification of economic theory. As we shall see, econometricians often use the equations proposed by the mathematical economist but put them in a form that lends itself to empirical testing. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill.

REFERENCES

- Chow, G.C. (1983). Econometrics, McGraw-Hill, New York.
- Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
- Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
- Gujarati, D.N.(1995). Basic Econometrics. McGraw-Hill, New Delhi.
- Koutsoyiannis,A.(1977). Theory of Econometrics (2nd Edn.). The Macmillan Press Ltd., London.
- Maddala, G.S. (1997). Econometrics, McGraw-Hill, New York.

Agro-Economic Research Centre
(For the States of West Bengal, Sikkim, and Andaman & Nicobar Islands)
Ministry of Agriculture and Farmers' Welfare
Government of India
Visva-Bharati, Santiniketan
West Bengal-731235
E-mail: dir.aerc@visva-bharati.ac.in
Web: https://www.visvabharati.ac.in/visva_bharati/agro-economic-research-centre/

ISBN: 978-81-989525-0-9