TB-ICN: 306/2023

Training Manual

सामाजिक विज्ञान अनुसंधान को सशक्त बनाने हेतु विश्लेषणात्मक तकनीकें

Analytical Techniques for Empowering Social Science Research

August 07 – 18, 2023

World Bank – ICAR funded

National Agricultural Higher Education Project

Centre for Advanced Agricultural Science and Technology

On

Genomics Assisted Crop Improvement and Management

कृषि अर्थशास्त्र संभाग

भा0कृ0अ0प0-भारतीय कृषि अनुसंधान संस्थान

नई दिल्ली-110012

NAHEP Sponsored training programme

On

Analytical Techniques for Empowering Social Science Research

Course Director

Alka Singh
Professor and Head
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi – 110012
Email: asingheco@gmail.com

Coordinators

Praveen K V
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail: veenkv@gmail.com

Renjini V R
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail:renji608@gmail.com

Chiranjit Mazumder
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail: majumder.chira@gmail.com

TB-ICN: 299/2023

Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi- 110 012

About NAHEP-CAAST at IARI, New Delhi

Centre for Advanced Agricultural Science and Technology (CAAST) is a new initiative and student centric subcomponent of World Bank sponsored **National Agricultural Higher Education Project (NAHEP)** granted to the Indian Council of Agricultural Research, New Delhi to provide a platform for strengthening educational and research activities of post graduate and doctoral students. The ICAR-Indian Agricultural Research Institute, New Delhi was selected by the NAHEP-CAAST programme. NAHEP sanctioned Rs 19.99 crores for the project on "**Genomic assisted crop improvement and management**" under CAAST programme. The project at IARI specifically aims at inculcating genomics education and skills among the students and enhancing the expertise of the faculty of IARI in the area of genomics.

Objectives:

1. To develop online teaching facility and online courses for enhancing the teaching and learning efficiency, and scientific communication skills
2. To develop and/or strengthen state-of-the art next-generation genomics and phenomics facilities for producing quality PG and Ph.D.students
3. To develop collaborative research programmes with institutes of international repute and industries in the area of genomics and phenomics
4. To enhance the skills of faculty and PG students of IARI and NARES
5. To generate and analyze big data in genomics and phenomics of crops, microbes and pests for genomics augmentation of crop improvement and management

IARI's CAAST project is unique as it aimed at providing funding and training support to the M.Sc. and Ph.D. students from different disciplines who are working in the area of genomics. It will organize lectures and training programmes, send IARI students for training at expert laboratories and research institutions abroad, and cover students from several disciplines. It will provide opportunities to the students and faculty to gain international exposure. Further, the project envisages developing a modern lab named as **Discovery Centre** that will serve as a common facility for students' research at IARI.

Core-Team Members:

S.No.	Name of the Faculty	Discipline	Institute
1.	Dr. Ashok K. Singh	Genetics	ICAR-IARI
2.	Dr. Vinod	Genetics	ICAR-IARI
3.	Dr. Gopala Krishnan S	Genetics	ICAR-IARI
4.	Dr. A. Kumar	Plant Pathology	ICAR-IARI
5.	Dr. T.K. Behera	Vegetable Science	ICAR-IARI
6.	Dr. R.N. Sahoo	Agricultural Physics	ICAR-IARI
7.	Dr. Alka Singh	Agricultural Economics	ICAR-IARI
8.	Dr. A.R. Rao	Bioinformatics	ICAR-IASRI
9.	Dr. R.C. Bhattacharya	Molecular Biology & Biotechnology	ICAR-NIPB
10.	Dr. K. Annapurna	Microbiology **Nodal officer, Grievance Redressal, CAAST**	ICAR-IARI
11.	Dr. R. Roy Burman	Agricultural Extension **Nodal officer, Equity Action Plan, CAAST**	ICAR-IARI
12.	Dr. K.M. Manjaiah	Soil Science & Agri. Chemistry **Nodal officer, CAAST**	ICAR-IARI
13.	Dr.Viswanathan Chinnusamy	Plant Physiology **PI, CAAST**	ICAR-IARI

Associate Team

S.No.	Name of the Faculty	Discipline	Institute
14.	Dr. Kumar Durgesh	Genetics	ICAR-IARI
15.	Dr. Ranjith K. Ellur	Genetics	ICAR-IARI
16.	Dr. N. Saini	Genetics	ICAR-IARI
17.	Dr. D. Vijay	Seed Science & Technology	ICAR-IARI
18.	Dr. Kishor Gaikwad	Molecular Biology & Biotechnology	ICAR-NIPB
19.	Dr. Mahesh Rao	Genetics	ICAR-NIPB
20.	Dr. Veena Gupta	Economic Botany	ICAR-NBPGR
21.	Dr. Era V. Malhotra	Molecular Biology & Biotechnology	ICAR-NBPGR
22.	Dr. Sudhir Kumar	Plant Physiology	ICAR-IARI
23.	Dr. Dhandapani R	Plant Physiology	ICAR-IARI
24.	Dr. Lekshmy S	Plant Physiology	ICAR-IARI
25.	Dr. Madan Pal	Plant Physiology	ICAR-IARI
26.	Dr. Shelly Praveen	Biochemistry	ICAR-IARI
27.	Dr. Suresh Kumar	Biochemistry	ICAR-IARI
28.	Dr. Ranjeet R. Kumar	Biochemistry	ICAR-IARI
29.	Dr. S.K. Singh	Fruits & Horticultural Technology	ICAR-IARI
30.	Dr. Manish Srivastava	Fruits & Horticultural Technology	ICAR-IARI
31.	Dr. Amit Kumar Goswami	Fruits & Horticulture Technology	ICAR-IARI
32.	Dr. Srawan Singh	Vegetable Science	ICAR-IARI
33.	Dr. Gograj S Jat	Vegetable Science	ICAR-IARI
34.	D. Praveen Kumar Singh	Vegetable Science	ICAR-IARI
35.	Dr. V.K. Baranwal	Plant Pathology	ICAR-IARI
36.	Dr. Deeba Kamil	Plant Pathology	ICAR-IARI
37.	Dr. Vaibhav K. Singh	Plant Pathology	ICAR-IARI
38.	Dr. Uma Rao	Nematology	ICAR-IARI
39.	Dr. S. Subramanium	Entomology	ICAR-IARI
40.	Dr. M.K. Dhillon	Entomology	ICAR-IARI
41.	Dr. B. Ramakrishnan	Microbiology	ICAR-IARI
42.	Dr. V. Govindasamy	Microbiology	ICAR-IARI
43.	Dr. S.P. Datta	Soil Science & Agricultural Chemistry	ICAR-IARI
44.	Dr. R.N. Padaria	Agricultural Extension	ICAR-IARI
45.	Dr. Satyapriya	Agricultural Extension	ICAR-IARI
46.	Dr. Sudeep Marwaha	Computer Application	ICAR-IASRI
47.	Dr. Seema Jaggi	Agricultural Statistics	ICAR-IASRI
48.	Dr. Anindita Datta	Agricultural Statistics	ICAR-IASRI
49.	Dr. Soumen Pal	Computer Application	ICAR-IASRI
50.	Dr. Sanjeev Kumar	Bioinformatics	ICAR-IASRI
51.	Dr. S.K. Jha	Food Science & Post Harvest Technology	ICAR-IARI
52.	Dr. Shiv Dhar Mishra	Agronomy	ICAR-IARI
53.	Dr. D.K. Singh	Agricultural Engineering	ICAR-IARI
54.	Dr. S. Naresh Kumar	Environmental Sciences; **Nodal officer, Environmental Management Framework**	ICAR-IARI

Preface

Social science research, particularly in the applied disciplines of Agricultural Economics, Agricultural Extension and Agricultural Statistics, is characterised by a diversity of theoretical perspectives, substantive orientation, methodological strategies, data collection practices and analytical techniques. The students of these disciplines usually have to face challenges in research, since it involves conceptualizing the problems relevant to the stakeholders, collecting and handling large data sets (both primary as well as secondary), choosing appropriate methodology (qualitative and quantitative), executing the analysis using appropriate statistical packages, and interpreting and presenting the results in a meaning and useful format to all: farmers, academia and policymakers.

Recognizing the duty to impart essential research skills to the social science students, we have taken up the task of conducting the training and preparing this manual on “Analytical Techniques for Empowering Social Science Research”. The training and manual is sponsored by the Centre for Advanced Agricultural Science and Technology (CAAST), which is a new initiative and student-centric sub-component of World Bank sponsored National Agricultural Higher Education Project (NAHEP), granted to IARI to provide a platform for strengthening education and research activities of post-graduate students. Qualitative and quantitative methods are essential components of evidence-based research in Social Sciences. Since the last two decades have experienced rapid advancement in the methodology and analytical techniques, as well as their applications in the field of social sciences, it becomes imperative to disseminate the knowledge of these novel techniques to the students. This training manual is prepared considering the target of upgrading the research skills of the post-graduate students of social sciences.

The various chapters of this manual are contributed by the eminent social scientists of the country, with expertise in analytical methods. The primary goal of this training program on “Advanced Research Methods and Essential Skills for the Social Sciences” is to introduce students to and give a general overview of how social scientists formulate and address research problems. The course will introduce students to structuring their research design, gathering social data, methodologies of data analysis, and result interpretation through a systematic research paradigm in addition to providing an overview of recognising social problems and generating research questions. We take this opportunity to sincerely acknowledge the contribution of all the authors in the preparation of this manual. Considering the diversity and comprehensive nature of the topics covered, the manual can act as a quick and effective reference source for the students in their future research endeavours.

Alka Singh

Praveen K.V

Renjini V R

Chiranjit Mazumder

Date: 18.08.2023

Acknowledgments

1. Secretary DARE and Director General ICAR, New Delhi
2. Deputy Director General (Education), ICAR, New Delhi
3. Assistant Director General (HRD), ICAR, New Delhi
4. National Coordinator, NAHEP, ICAR, New Delhi
5. CAAST Team, ICAR-IARI, New Delhi
6. P.G. School, ICAR-IARI, New Delhi
7. Director, ICAR-IARI, New Delhi
8. Dean & Joint Director (Education), ICAR-IARI, New Delhi
9. Joint Director (Research), ICAR-IARI, New Delhi
10. Head, Division of Agricultural Economics, ICAR-IARI, New Delhi
11. Professor, Division of Agricultural Economics, ICAR-IARI, New Delhi
12. AKMU, ICAR-IARI, New Delhi
13. Staff & Students, Division of Agricultural Economics, ICAR-IARI, New Delhi

Contents

Basic statistical methods and sampling techniques for social sciences

Girish Kumar Jha
Division of Bioinformatics, ICAR-Indian Agricultural Statistical Research Institute

INTRODUCTION

The main aim of a program evaluation study is to learn about how a population of interest is affected by the program. In experimental studies, establishing such causation is much easier as the experimenter has complete control on the experiment and one can keep all other things constant across groups except for the treatment. In such case, the counterfactual outcome is directly observed and impact (change in outcome due to treatment or cause is simply the difference in outcomes between units which received the treatment and the counterfactual, which didn't receive the treatment. But, in observational studies, the experimenter is a passive collector of the data and the counterfactual outcome is difficult to observe. The treatment and control groups vary not only with respect to treatment but also with respect to many other variables and hence it is difficult to attribute the difference in outcomes only to the treatment. This is also called as 'self-selection bias' where units with certain attributes tend to self-select themselves into either treatment or control groups. The root cause of the problem is lack of random allocation-if the units were to be allocated to treatment and control groups randomly, on an average the two groups will be similar to each other on all observable and unobservable characters except for treatment. However, very rarely policies are implemented by selecting beneficiaries randomly or technologies are given at random. The main aim of this lecture note is to provide an overview of sampling procedures which is necessary to construct a suitable sampling plan for achieving the goals of an impact evaluation study.

In sample surveys, population or an aggregate to represent the whole exists in its own way and we need to device the sampling techniques in such a way that the sample should represent the population. Here the objective is to infer about the population on the basis of a 'part' of the population which is known as a sample. Sampling is frequently used in everyday life in all kinds of investigations. Almost instinctively, before deciding to buy a lot, we examine a few articles preferably from different parts of the lot. Another example relates to a handful of grain taken from a sack to determine the quality of the grain. These are examples where inferences are drawn on the basis of the results obtained from a sample. Sampling is not always necessary. When a population is small, one may choose to collect the data for each and every unit belonging to the population and this procedure of obtaining information is termed as complete enumeration. However, the effort, money and time required for carrying out complete enumeration for large population, generally is extremely large. For example, suppose we are interested in assessing the impact of a programme on the adoption of a new technology by farmers in a region. Should we collect data about the adoption of the new technology from each and every farmer of the region or data from a sample of farmers will serve the purpose? In most of the cases, sample can provide sufficient evidence in form of data useful for the programme evaluation. A sample can also save valuable time, money and the labour of Extension professionals. Time is saved because only fewer (part of the population) farmers, households, etc. must be interviewed or surveyed, thus the complete set of data can be collected quickly. Money and labour are saved because less data must be collected. In addition, errors from handling the data (e.g. entering data into a computer file) are likely to be reduced because there are fewer opportunities to make mistakes.

CONCEPTS AND DEFINITIONS

Sampling Unit: Elementary units or groups of units which are clearly defined, identifiable, observable and convenient for sampling purposes are called sampling units.

Sampling Frame: The list or map of sampling units is called the sampling frame and provides the basis for the selection of units in the sample. The common problem of sampling frames are noncoverage or incomplete frame, foreign units, duplicate listings and cluster of elements combined into one listings etc.

Population: A population consists of a group of units defined according to the objectives of the survey. The population may consist of all the households in a village/ locality, all the fields under a particular crop. We may also consider a population of persons, families, fields, animals in a region, or a population of trees, birds in a forest depending upon the nature of data required.

Probability and Non-probability Sampling: The two standard ways to draw a sample are probability and non-probability sampling. If the purpose of the evaluation is to generalize for the whole group on the basis of sample results or to provide a statistical basis for saying that the sample is representative, a probability sample is appropriate. If the aim of the evaluation is to learn about individuals or cases for some purpose other than generalizing to the population, or if random selection is not feasible, sometimes study is limited to those participants that agree to be included then non-probability sampling is appropriate.

Probability sampling is a method of selecting samples according to certain laws of probability in which each unit of the population has some known and positive probability of its being selected in the sample. Because the probability is known, the sample statistics can be generalized to the population at large (at least within a given level of precision).

In non-probability sampling procedures, choice of selection of sampling units depends entirely on the discretion or judgment of the sampler. This method is called purposive or judgment sampling. In this procedure, the sampler inspects the entire population and selects a sample of typical units which he considers close to the average of population. One thing common to all these non-probability sampling methods is that the selection of the sample is confined to a part of the population. None of the methods give a sample which can be considered to represent the entire population. A particular sample may prove to be very good or very bad but unless one has the knowledge of the complete population, it is not possible to know the performance of a particular sample. Moreover, since these methods lack a proper mathematical basis, these are not amenable to the development of the sampling theory. Nonprobabilty samples are quite convenient and economical.

Non-probability samples include haphazard, convenience, quota and purposive samples. Haphazard samples are those in which no conscious planning or consistent procedures are employed to select sample units. Convenience samples are those in which a unit is self-selected (e.g., volunteers) or easily accessible. Quota samples are those in which a predetermined number of units which have certain characteristics are selected. A sample of 50 small and 50 large farmers to be interviewed in a village is an example of this type. Researchers select units (e.g., individuals) for a purposive sample on the basis of characteristics or attributes that are important to the evaluation. The units used in a purposive sample are sometimes extreme or critical units. Suppose we are evaluating the adoption rates of a technology by farmers and we want to know if large farmers differ from small farmers. A sample of extreme units, e.g., farms of 10 or more hectares and farms of 2 or less hectares, would provide information to make this comparison.

Sampling and Non-sampling Errors: The sampling errors arise because estimates of parameter is based on a ‘part’ from the ‘whole’ population while non-sampling errors mainly arise because of some departure from the prescribed rule of the survey such as survey design, field work, tabulation & analysis of data etc. The sampling error usually decreases with increase in sample size (number of units selected in the sample) and is non-existent in a complete enumeration survey, since the whole population is surveyed. However, non-sampling errors are common to both to complete enumeration and sample surveys.

VARIOUS SAMPLING METHODS

A sampling method is a scientific and objective procedure of selecting units from the population and provides a sample that is expected to be representative of the population. One of the vital issues in sample surveys is the choice of a proper sampling strategy, which essentially comprise of a sampling method and the estimation procedure. In the choice of a sampling method there are some methods of selection while some others are control measures which help in grouping the population before the selection process. In the methods of selection, schemes such as simple random sampling, systematic sampling and varying probability sampling are generally used. Among the control measures are procedures such as stratified sampling, cluster sampling and multi-stage sampling etc. A combination of control measures along with the method of selection is called the sampling scheme.

We shall describe in brief the different sampling methods in the following sections.

Simple Random Sampling

Simple random sampling (SRS) is a method of selecting 'n' units (sample size) out of 'N' units (population size) such that every one of the non-distinct samples has an equal chance of being chosen. In practice, a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N are then drawn either by means of a table of random numbers or by means of a computer program that produces such a table. Sampling where each member of a population may be chosen more than once is called sampling with replacement. Similarly a method of sampling in which each member cannot be chosen more than once is called sampling without replacement.

Procedure of Selecting a Random Sampl*e*: Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

(i) Lottery Method,

(ii) Use of Random Number Tables.

Lottery Method: Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits/tickets are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits/tickets may be drawn one by one may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits/tickets and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method.

Use of Random Number Tables: A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern, where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1,2…,9 appear independent of each other. Some random Tables in common use are:

(a) Tippett's random number Tables

(b) Fisher and Yates Tables

(c) A million random digits Table.

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digits numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is by selecting a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The used numbers is, therefore, modified as remainder approach.

Remainder Approach: Let N be a r-digit number and let its r-digit highest multiple be N'. A random number k is chosen from 1 to N' and the unit with serial number equal to the remainder obtained on dividing k by N is selected. If the remainder is zero, the last unit is selected. As an illustration, let N = 123, the highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123, the remainder is 41. Hence, the unit with serial number 41 is selected in the sample.

Stratified Random Sampling

In SRS the precision of a sample estimate of the population mean depends not only upon the size of the sample but also on the population variability. Selection of a simple random sample from the entire population may be desirable when we do not have any knowledge about the nature of population, such as, population variability etc. However, if it is known that the population has got differential behaviour regarding variability, in different pockets, this information can be made use of in providing a control in the selection. The approach through which such a controlled selection can be exercised is called stratified sampling.

In stratified sampling, the whole population is divided into several homogenous groups (strata), thereby, controlling variability within each group and a random sample of pre-determined size is drawn independently from each one of the groups. To obtain full benefit from stratification, the strata sizes must be known. If a simple random sample is taken in each stratum then the procedure is termed as stratified random sampling. As the sampling variance of the estimate of mean or total depends on within strata variation, the stratification of population is done in such a way that strata are homogeneous within themselves with respect to the variable under study. However, in many practical situations it is usually difficult to stratify with respect to the variable under consideration especially because of physical and cost consideration. Generally the stratification is done according to administrative groupings, geographical regions and on the basis of auxiliary characters correlated with the character under study.

Cluster Sampling

A cluster may be defined as a group of units. When the sampling units are clusters, the method of sampling is known as cluster sampling. Cluster sampling is used when the frame of units is not available or it is expensive to construct such a frame. Thus, a list of all the farms in the districts may not be available but information on the list of villages is easily available. For carrying out any district level survey aimed at estimating the yield of a crop, it is practically feasible to select villages first and then enumerating the elements (in this case farms) in the selected village. The method is operationally convenient, less time consuming and more importantly such a method is cost-wise efficient. The main disadvantage of cluster sampling is that it is less efficient than a method of sampling in which the units are selected individually.

Multi-Stage Sampling

Generally, elements belong to the same cluster are more homogeneous as compared to those elements which belong to different clusters. Therefore, a comparatively representative sample can be obtained by enumerating each cluster partially and distributing the entire sample over more clusters. This will increase the cost of the survey but the proportionate increase in cost vis-à-vis cluster sampling will be less as compared to increase in the precision. This process of first electing cluster and then further sampling units within a cluster is called as two-stage sampling. The clusters in a two-stage sample are called as primary-stage units (psu) and elements within a cluster are called as second-stage units (ssu).

A two-stage sample has the advantage that after psu's are selected, the frame of the ssu's is required only for the sampled psu's. The procedure allows the flexibility of using different sampling design at the different stages of selection of sampling units. A two-stage sampling procedure can be easily generalized to multi-stage sampling designs. Such a sampling design is commonly used in large scale surveys. It is operationally convenient, provides reasonable degree of precision and is cost-wise efficient.

Systematic Sampling

In systematic sampling, only the first unit is selected at random and then proceeds with the selection of every k-th unit from then onwards. In this case, k = (population size/sample size). The method of systematic sampling is used on account of its low cost and simplicity in the selection of the sample. It makes control of field work easier. Since every k-th unit will be in the sample, the method is expected to provide an evenly balanced sample.

Systematic sampling can be used in situations such as selection of k-th strip in forest sampling, selection of corn fields every k-th mile apart for observation on incidence of borers, or the selection of every k-th time interval for observing the number of fishing craft landing at a centre.

For example, suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #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 every 10^{th} street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts.

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 then simple random sampling. Another drawback of systematic sampling is that it is not possible to get an unbiased estimation of the variance of the estimator.

Varying Probability Sampling

In simple random sampling without replacement (SRSWOR), the selection probabilities are equal for all the units in the population. However, if the sampling units vary in size considerably, SRSWOR may not be appropriate as it doesn't take into account the possible importance of the larger units in the population. To give possible importance to larger units, there are various sampling methods in which this can be achieved. A simple methods is of assigning unequal probabilities of selection to the different units in the population. Thus, when units vary in size and the variable under study is correlated with size, probabilities of selection may be assigned in proportion to the size of the unit e.g. villages having larger geographical areas are likely to have larger populations and larger areas under food crops. For estimating the crop production, it may be desirable to adopt a selection scheme in which villages are selected with probabilities proportional to their populations or to their geographical areas. A sampling procedure in which the units are selected with probabilities proportional to some measure of their size is known as sampling with probability proportional to size (pps). The units may be selected with or without replacement. In sampling with replacement, the probability of drawing a specified unit at a given draw is the same. In this case sample is selected either through cumulative total method or Lahiri's method.

Now let us move to the practical consideration of sample survey.

PLANNING AND EXECUTION OF SAMPLE SURVEYS

Sample Surveys are widely used as a cost effective instruments of data collection and for making valid inferences about population parameters. Most of the steps involved while planning a sample survey are common to those for a complete enumeration. These major stages of a survey are **planning, data collection and tabulation of data**. Some of the important aspects requiring attention of the planning stage are as follows:

- Formulation of data requirements – objectives of the survey
- Ad-hoc or repetitive survey
- Method of data collection
- Questionnaire versus schedules
- Survey, reference and reporting periods
- Problems of sampling frames
- Choice of sampling design
- Planning of pilot survey
- Field work

The different aspects listed above are inter-dependent.

(i) Formulation of Data requirements:

The user i.e., the person or organizations requiring the statistical information are expected to formulate the objectives of the survey. The user's formulation of data requirements is not likely to be adequately precise from the statistical point of view. It is for the survey statistician to give a clear formulation of the objectives of the survey and to check up whether his formulation faithfully reflects the requirements of the users.

(ii) Survey: Ad-hoc or repetitive:

An ad-hoc survey is one which is conducted without any intention of or provision for repeating it, whereas a repetitive survey is one, in which data are collected periodically for the same, partially replaced or freshly selected sample units. If the aim is to study only the current situation, the survey can be an ad-hoc one. But when changes or trends in some characteristics overtime are of interest, it is necessary to carry out the survey repetitively.

(iii) Method of collecting Primary Data:

There are variety of methods that can be used to collect information. The method to be followed has to be decided keeping in view the cost involved and the precision aimed at. The methods usually adopted for collecting primary data are:

- Recorded information
- Physical observation
- Face-to-face interviewing
- Postal enquiries
- Telephone interviews
- Web based survey etc.

(iv) Questionnaire Vs. Schedule:

In the questionnaire approach, the informants or respondents are asked pre-specified questions and their replies to these questions are recorded by themselves or by investigators. In this case, the investigator is not supposed to influence the respondents. This approach is widely used in mail enquiries. In the schedule approach, the exact form of the questions to be asked are not given and the task of questioning and soliciting information is left to the investigator, who backed by the training and instructions has to use his ingenuity in explaining the concepts and definitions to the informant for obtaining reliable information.

However, these two terms are often used synonymously. Designing questionnaire is one of the vital aspects of survey. Few suggestions for wording questions

- Use Simple words
- Questions should be concise
- Avoid multiple meaning questions
- Avoid ambiguous questions
- Minimum amount of writing on schedule
- Check on accuracy & consistency
- Handbook of instructions

(v) Survey, Reference and Reporting Periods:

Another aspect requiring special attention is the determination of survey period, reference period and reporting periods.

- Survey Period: The time period during which the required data is collected.
- Reference Period: The time period to which the collective data for all the units should refer.
- Reporting Period: The time period for which the required statistical information is collected for a unit at a time (reporting period is a part or whole of the reference period).

(vi) Choice of Sampling Design:

The principle generally adopted in the choice of a design is either reduction of overall cost for a pre-specified permissible error or reduction of margin of error of the estimates for given fixed cost. Generally a stratified uni-stage or multi-stage design is adopted for large scale surveys. For efficient planning, various auxiliary information which are normally available are utilized at various stages e.g. the area under particular crop available for previous years is normally used for size stratification of villages. If the information is available for each and every unit of the population and there is wide variability in the information then it may be used for selecting the sample through probability proportional to size methods.

(vii) Pilot surveys:

Where some prior information about the nature of population under study, and the operational and cost aspects of data collection and analysis is not available from part surveys. It is desirable to design and carry out a pilot survey. It will be useful for

- Testing out provisional schedules and related instructions
- Evolving suitable procedure for field and tabulation work, and
- Training field and tabulation staff
- Potential sources of measurement error
- Likely non-response rate
- Sensitive issues or sources of ambiguity
- Difficulties of access to chosen sample members

(viii) Field Work:

While planning the field work of the survey, a careful consideration is needed regarding choice of the field agency. For ad-hoc surveys, one may plan for ad-hoc staff but if survey is going to be a regular activity, the field agency should also be on a regular basis. Normally for regular surveys, the available field agencies are utilized. A regular plan of work by the Enumerators along with the rationalized supervision is an important consideration for getting a good quality of data. An initial quality check should be instituted while the interviewers are in the field to supply missing entries and correct apparent inconsistencies.

Determination of sample size

While planning a survey, a question often arises is that of fixing the size of the sample as unduly large sample size may mean wastage of resources while a smaller sample size limits the utility of results. The sample sizes are determined by fixing the precision of the estimate. It can be seen that sample size depends on population variance, which is generally not known. An estimated value of population variance can be obtained either from a pilot survey or by previous sampling of the same or similar population or by guess work about the structure of the population. Besides population variance, sample size depends on the minimum effect size we want to detect as well as power of the test. Illustration of power and sample size calculation through STATA will be discussed in the class. A do-file and log-file for power and sample size calculation will also be shared during the class.

REFERENCES

Cochran, W.G. (1977) Sampling *Techniques*, Third Edition, John Wiley & Sons, New York.

Murthy, M.N. (1967) *Sampling Theory and Methods*, First Edition, Statistical PublishingSociety, Calcutta.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C.(1984) *Sampling Theory of Surveys with Applications*. Third Edition, Iowa State University Press, USA and Indian Society of Agricultural Statistics, New Delhi.

Introduction to Regression Analysis

Chiranjit Mazumder & Utkarsh Tiwari

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

1. Why and when to use regression analysis

Regression analysis is a statistical method used to understand the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X). It's used when we want to predict the value of a variable based on the value of another variable.

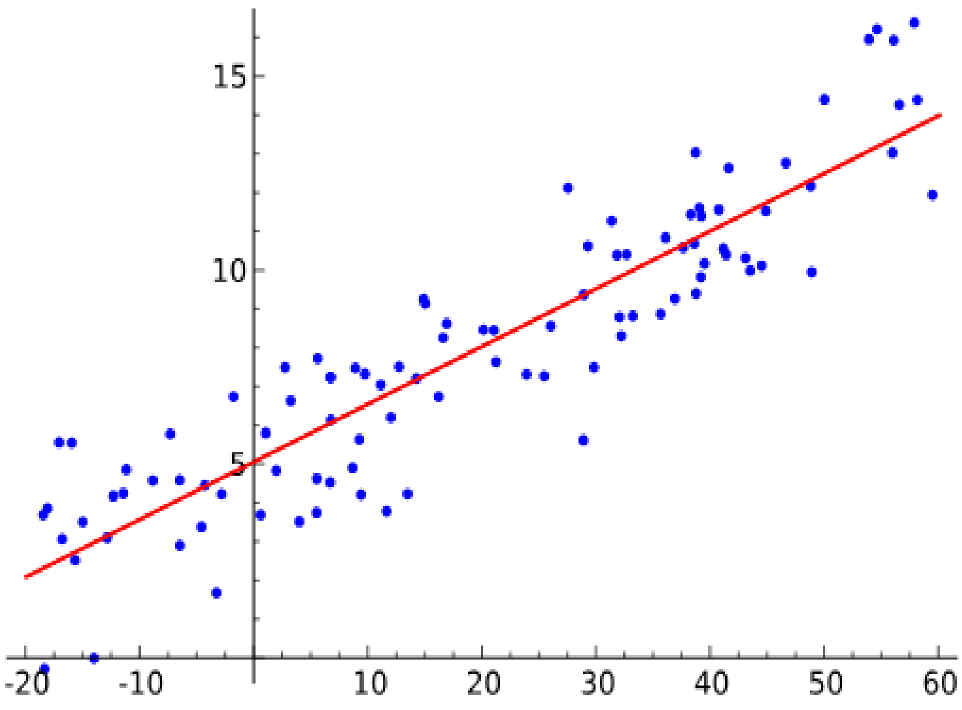
For example, predicting crop yield (Y) based on rainfall (X). The equation of a simple linear regression is:

$$Y = a + bX + e$$

where:

- Y is the dependent variable.
- X is the independent variable.
- a is the Y-intercept.
- b is the slope.
- e is the error term.

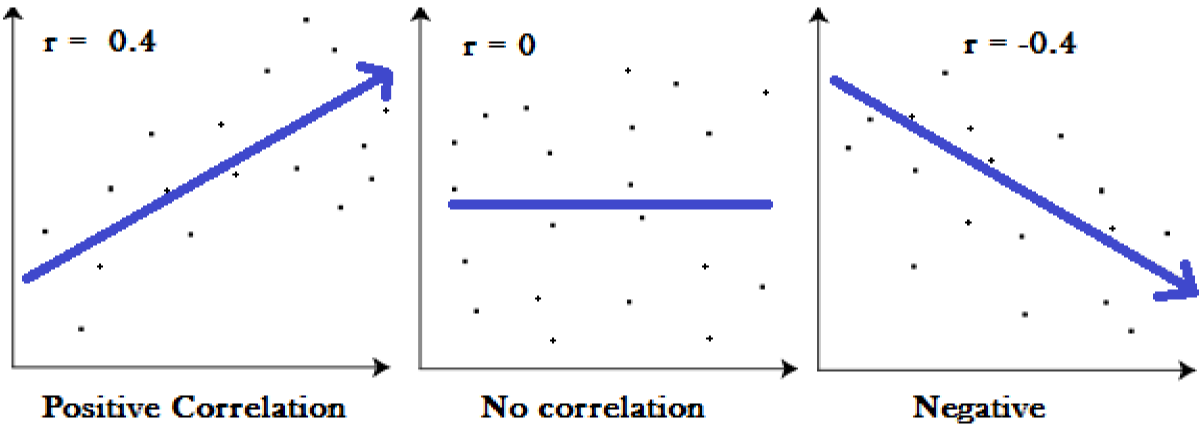
Here's a simple graph showing a linear regression line:



2. **Difference between Correlation and Regression**

Correlation measures the strength and direction of the linear relationship between two variables, while regression quantifies the nature of the relationship.

Correlation is a single statistic, while regression produces an entire equation. In correlation, both variables are treated equally, but in regression, there is a distinction between dependent and independent variables.



3. Types of Regression Analysis:

Simple Linear Regression:

This is the most basic type of regression where we have one predictor (independent variable) and one response (dependent variable). The goal is to find a linear relationship between these two variables. The relationship is represented by the equation:

$$y = a + bx + e$$

where:

- y is the dependent variable.
- x is the independent variable.
- a is the y-intercept.
- b is the slope of the line.
- e is the error term.

In the context of agriculture, an example could be predicting crop yield based on the amount of rainfall. The amount of rainfall is the independent variable and the crop yield is the dependent variable.

Multiple Linear Regression:

This is an extension of simple linear regression where we have more than one independent variable. The equation becomes:

$$y = a + b1 * x1 + b2 * x2 + \ldots + bn * xn + e$$

An example in agriculture could be predicting crop yield based on multiple factors like rainfall, temperature, and fertilizer used.

Polynomial Regression:

This is a type of regression where the relationship between the independent variable *x* and the dependent variable *y* is modelled as an n^{th} degree polynomial. The equation becomes:

$$y = a + b1 * x + b2 * x^2 + \ldots + bn * x^n + e$$

An example could be predicting the growth of a certain pest population based on temperature. The relationship might not be linear and could follow a polynomial pattern.

Robust Regression:

This type of regression is used when the data contains outliers or is contaminated with noise. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process.

Nonlinear Regression:

In this type of regression, the relationship between the independent variables *x* and the dependent variable *y* is modelled as a non-linear function. An example could be predicting the growth of a certain crop which might follow a logistic growth pattern.

General Linear Models:

These are a generalization of multiple linear regression models to allow for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

1. **Logistic Regression:**

This is a type of GLM used when the dependent variable is binary (i.e., it takes on only two values, like 0/1, Yes/No). The goal of logistic regression is to find the best fitting model to describe the relationship between the binary characteristic of interest (dependent variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

For example, you might use logistic regression to predict whether a plant will survive (1) or not (0) based on the amount of water it receives.

2. **Poisson Regression:**

Poisson regression is used for count data. If you have a dependent variable that represents the number of times an event occurs (like the number of pests on a plant), you might use a Poisson regression.

For example, you might use Poisson regression to predict the number of pests on a plant based on the amount of pesticide used.

3. **Multinomial Logistic Regression:**

This is an extension of logistic regression, and is used when the dependent variable is categorical and has more than two levels. That is, it can be used to model outcomes with multiple possibilities.

For example, you might use multinomial logistic regression to predict the type of crop (wheat, corn, soybeans) that a farmer will plant based on characteristics of the soil and climate.

4. **Ordinal Logistic Regression:**

This is another extension of logistic regression, used when the dependent variable is ordinal, i.e., it represents categories that have a specific order (like a rating of soil quality from poor to excellent).

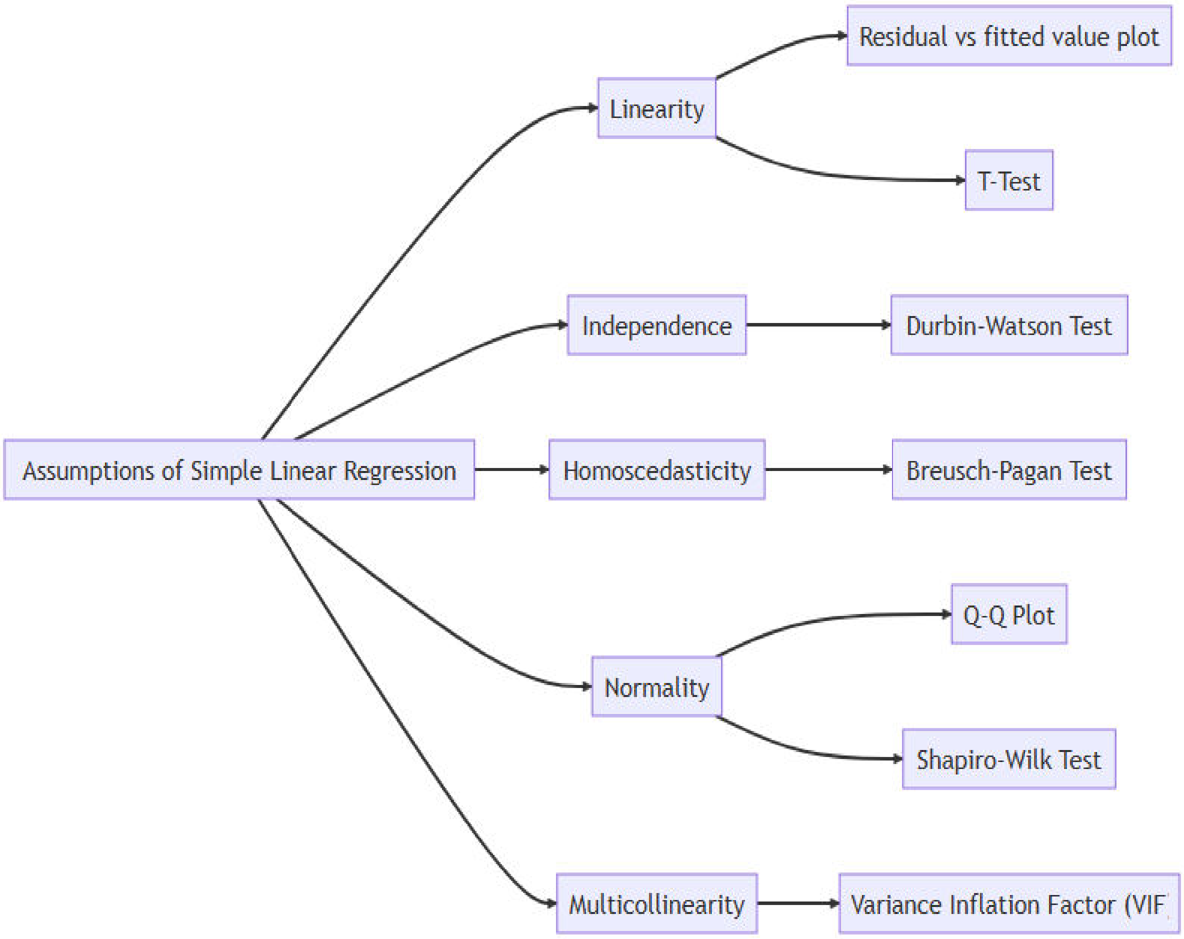
For example, you might use ordinal logistic regression to predict the quality of soil (poor, average, good, excellent) based on the amount of organic matter and minerals in the soil.

5. **Negative Binomial Regression:**

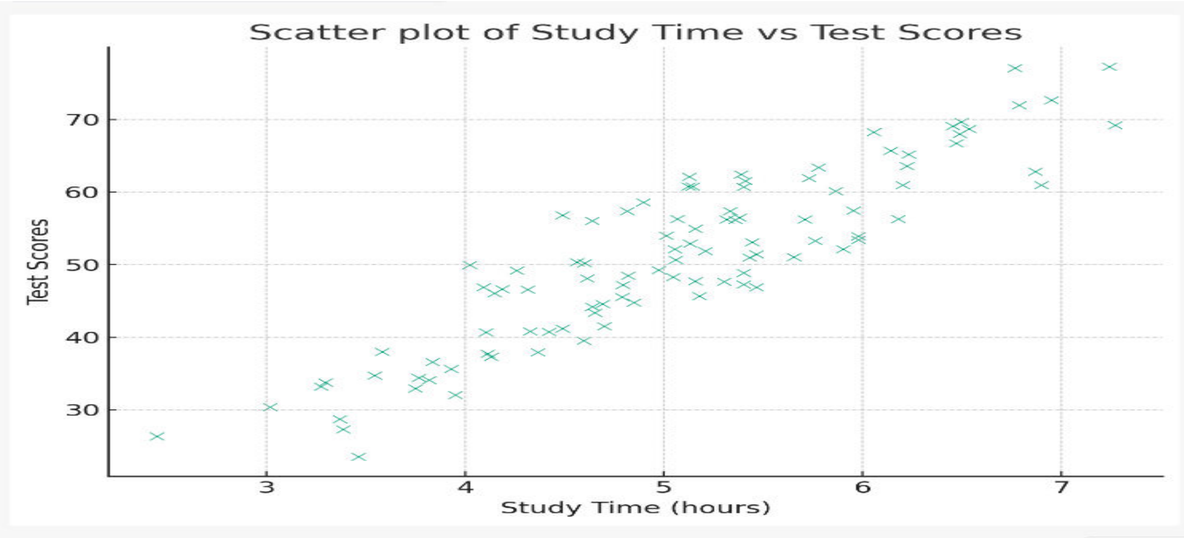
This is used when the dependent variable is a count of the number of times an event occurs, but where the variance is greater than the mean (overdispersion).

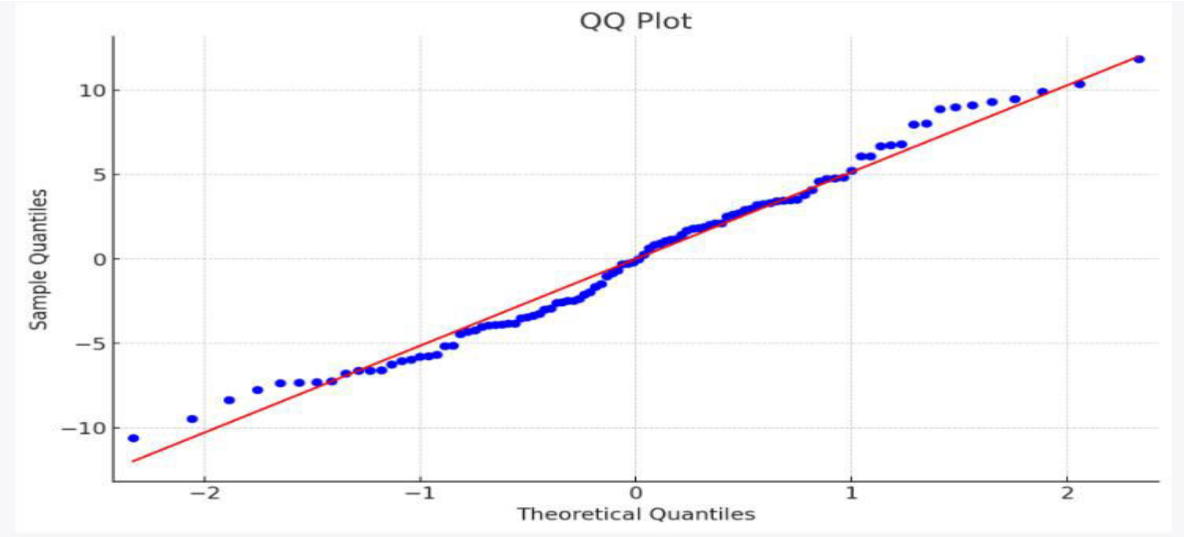
For example, you might use negative binomial regression to predict the number of times a particular insect pest damages a crop in a growing season.

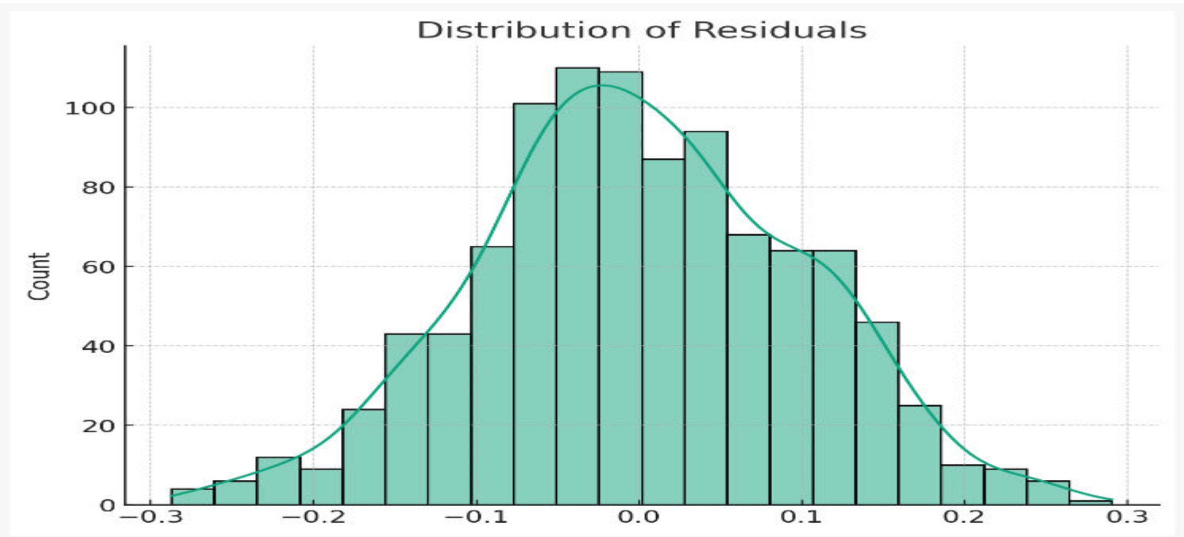
Each of these models has its own set of assumptions and is appropriate for different types of data and research questions. The choice of model depends on the nature of the dependent variable and the goals of the analysis.



- **Linearity**: This assumption can be checked by plotting residuals versus fitted values. A non-random pattern suggests that the data may not be linear. The t-Test or rainbow test can also be used to check for linearity.
- **Independence**: The Durbin-Watson Test is used to detect the presence of autocorrelation (a relationship between values separated from each other by a given time lag) in the residuals.
- **Homoscedasticity**: The Breusch-Pagan Test is used to test for homoscedasticity, which is the assumption that the variance of the errors is constant across all levels of the independent variables.
- **Normality**: The Q-Q Plot and the Shapiro-Wilk Test, Kolmogorov Smirnov test are used to check if the residuals are normally distributed.
- **Multicollinearity**: The Variance Inflation Factor (VIF) is used to check for multicollinearity, which is when there is high correlation between predictor variables, making it difficult to determine the individual effect of each predictor.


Scatter plot of Study Time vs Test Scores
Test Scores
70
60
50
40
30
3
4
5
6
7
Study Time (hours)


QQ Plot
Sample Quantiles
10
5
0
−5
−10
−2
−1
0
1
2
Theoretical Quantiles


Distribution of Residuals
Count
100
80
60
40
20
0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3

An Introduction to Stata Software for Statistical Analysis

Nithyashree M L
Division of Agricultural Economics, ICAR-Indian Agricultural Economics

This chapter describes an overview and basic commands used in the Stata software which will help the beginners to get familiar and hold a grip in using the software for further statistical analysis. Those who don't have access to Stata software can avail the short-term student license service by using the link https://www.stata.com/customer-service/short-term-license/. As shown below the main interphase (Figure 1) of the Stata has a menu system that enables the users to perform the task by interactive UI; drop-down menu. Alternatively, there is also a command window with which users can write commands directly for the statistical analysis. For example, to summarize data, by typing summarize variable/variables name, we can get the result and one can also do it by using the drop-down menu by going to current statistics (Figure 2). As a beginner one can start by exploring the possible option available by using the drop-down menu, after getting comfortable with the software, can directly type the command to perform any required analysis which is more efficient and professional, apart from this the executed commands can be copied from the history/review window and saved in do file, which can be easily shared with other researchers and this enables single-click execution. With the right side of the interphase, variables, and properties window variables and their related properties including labelling the variables and assigning variables values can be performed and visualized.

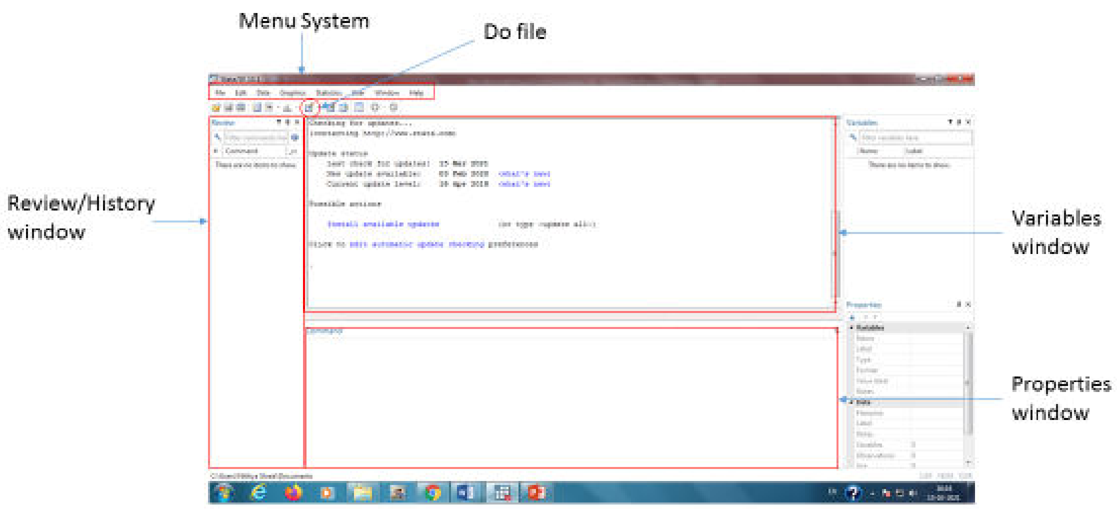


Figure 1. Main interphase of Stata

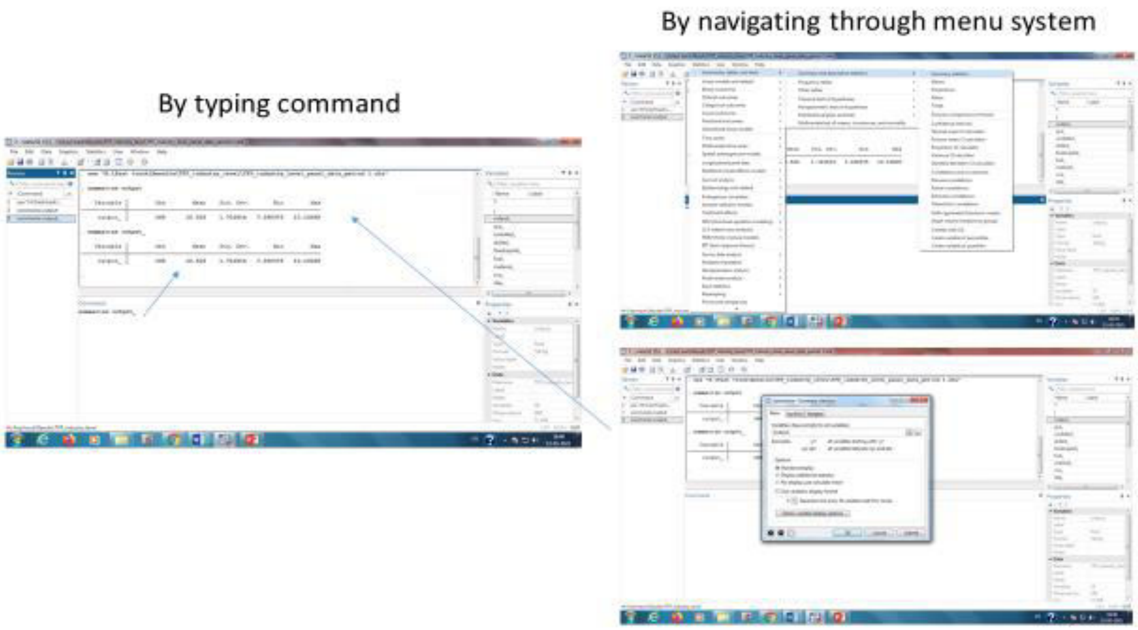
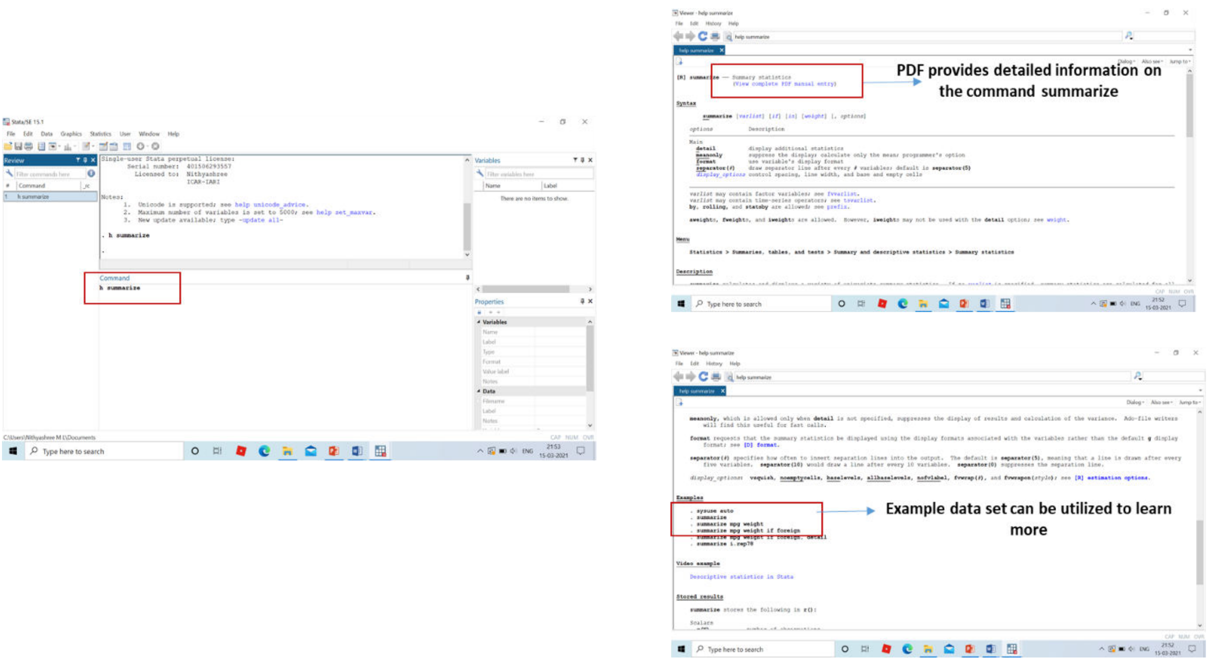
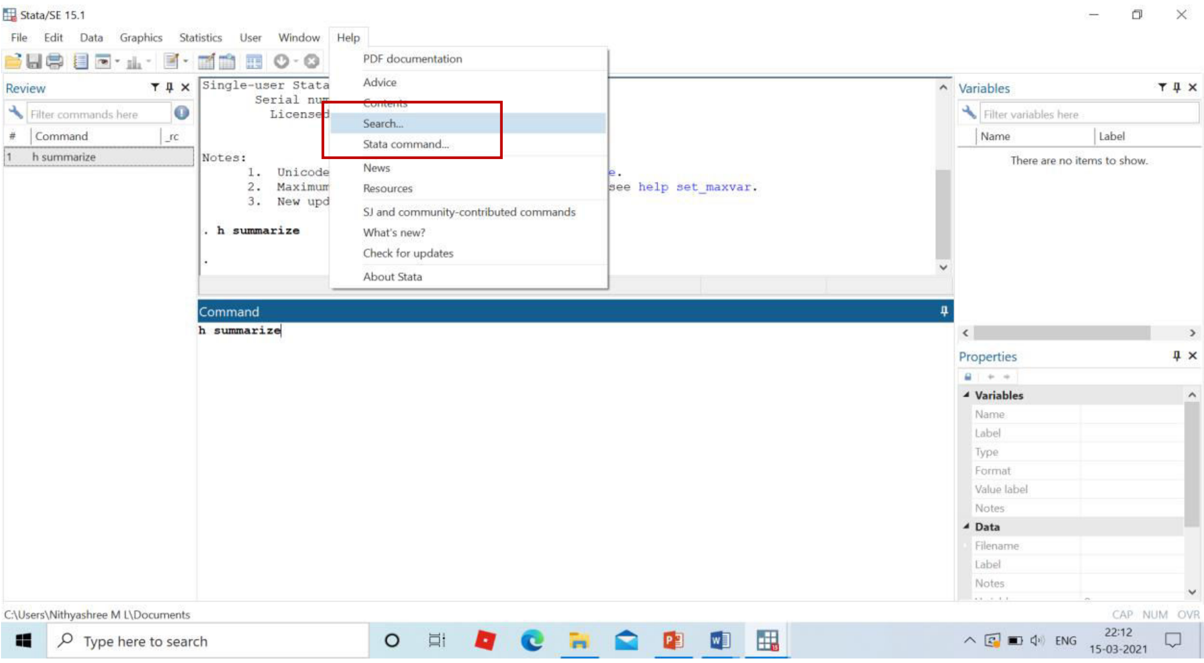


Figure 2. Two ways of performing task in Stata

Stata provides flexibility in exploring the various option to the users to learn the available functions and packages with the use of the help command. By using this command one can learn and use the applicability of different functions and packages. For example, by typing help summarize in the command window we get detailed information of the command summarize in the form of pdf also by scrolling down detailed information on using the summarize command directly as syntax as well as menu format along with some example data sets a shown below is a great source to learn Stata.



If the user is not sure about what to type in the command window, then the search option under the help in the menu bar can be explored. Besides there are many user-written commands are available in the Stata software in the form of packages, to use them first we need to install the packages. For this one can use findit command and after finding the suitable package install them to make use of those packages and also make sure your system connected to the internet while installing the packages.



It is always good practice to save the results of the analysis, that can be done in Stata by typing log using filename in the command window before beginning the analysis, which creates a log file by the given file name and after completing the analysis type log close and your analysis will be saved in the smcl format for example filename.smcl, which details all the activity which you carried out during the particular session or analysis. Which you may need at the later stage to communicate to the journals while submitting the research article.

Mathematical and logical operator in Stata are similar to those used in MS excel

a == b	if a equals b
a != b	if a not equal to b
a > b	if a greater than b
a >= b	if a greater than or equal to b
a < b	if a less than b
a <= b	if a less than or equal to b
a & b	**&** refers to **and**
a | b	**|** refers to **or**

Handling the data set

1. **Creating a new variable:** gen or egen command is used to create the new variable. These commands can be combined with arithmetic operators or logical operators

   Example: gen grade=1 if marks==2
   egen stdev_age= std(income)
   gen ln_wage = ln(wage)

2. **For labelling the variable:** to create a label to the variable, write the command lable variable and type the variable name need to be labelled then write the label with in the invited comma

   Example: label variable ln_wage "Log of hourly wage"

3. **The replace command:** Replace command generally helps in editing value of already existing or generated variable, for this command ***replace*** can be used.

   Example: replace gender=0 if missing(gender)
   replace gender=0 if gender==.

4. To sort the data in ascending order: use command ***sort variable name***

5. Command that can convert string to numeric variables: ***Tostring and destring***

Few Commands for Basic Statistical analysis

a. Regression: *reg or probit or logit*
b. Correlation: *corr or pwcorr*
c. Student T test: *ttest*
d. Factor analysis: *factor*
e. Marginal Effects after probit or logit: *mfx*
f. Chisquare test: *tab, all*
g. Principal Component Analysis: *PCA*

Few useful user written commands

h. *tatable2*: Calculate group wise mean value and test the significance

i. *orth_out*: Perform t-test for any number of variables at once

j. *dea*: Data envelopment analysis

k. *acfest*: Production function est. using Ackerberg-Caves-Frazer method

l. *levpet*: Production function est. using Levinsohn and Petrin approach

m. doubleb: Perform Double Bound Contingent Valuation.

n. clustersampsi: perform power calculations for RCT

Exercise- 1

1. Import the dataset "stata_intro", which has been shared with you already.
2. Summarize the data so that you see the means, standard deviation, min, and max of each variable
3. generate the logarithmic transformation of the variable wage as ln_wage
4. label variable ln_wage, wage, collgrad and union as Log of hourly wage, Hourly wage, College graduate and Union member respectively.
5. define label values for the variable collgrad
6. encode racecat as race
7. obtain mean wage sd wage for the variable union

9. draw histogram for the variable ln_wage and superimpose normal curve
10. generate scatter plot for the variables wage tenure
11. generate scatter plot for the variables wage tenure by union
12. obtain a liner regression equation of ln_wage and tenure
13. obtain a liner regression equation ln_wage and tenure along with by considering any one of the categorical variable and also an interaction component
14. save the commands in do file and results in log file.

Introduction to Impact Assessment

Aditya K.S. & Praveen K.V.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Impact assessment methods have their roots in theory of causal inference. Theory of causal inference is concerned with establishing causation and estimating the magnitude of effect of the cause. Within the framework of causal inference, impact is defined as the expected change in the outcome of interest in the absence of the treatment/ intervention. In other words, impact assessment methods aim at estimating the difference in value of the outcome variable of the units receiving the treatment/ intervention to what would have happened in absence of the treatment. Since the value of outcome variable of treatment group in the absence of treatment is never observed, the value of outcome from the control group is used as proxy. However, in observational studies, due to non-random allocation of the treatment and due to confounding variables, it is very difficult to find a suitable counterfactual. Assessing impact in absence of suitable counterfactual can lead to either over or under estimation of the impact of the treatment. Impact assessment methods are aimed at constructing the suitable counterfactual outcomes, either through experimental setting or through quasi experiments and regression based adjustments to minimize the element of bias in estimates of Impact.

In the sessions to follow, the term Impact indicates the change in the parameters/ outcome of interest of the beneficiaries due to the intervention/ program. This definition is derived from theory of causal inference. Let us examine basic notations used to understand impact.

$Y_{ij}|\ T = 1$ here the subscript $_j$ indicates the treatment status; If j=1, then the unit received the treatment, if it is 0, the unit did not receive the treatment, and T indicates which group it belongs to. T=1 means treatment group and T=0 means control group

Now, in a true sense,

$$Impact = E\{Y_{i1}|\ T = 1\} - E\{Y_{i0}|\ T = 1\}$$

But $Y_{i0}|\ T = 1$ is never observed (Same units cannot simultaneously receive and do not receive treatment at a particular point of time).

What we actually estimate is

$$\widehat{Impact} = E\{Y_{1i}|\ T = 1\} - E\{Y_{i0}|\ T = 0\}$$

As long as $E\{Y_{i0}|\, T = 0\} = E\{Y_{i0}|\, T = 1\}$, bias will be zero, for this to happen, the two groups must be similar to each other in all covariates, except for treatment. Add and subtract $\{Y_{i0}|\, T = 1\}$ – What would have happened if the control group were not receiving the treatment.

$$\widehat{Impact} = E\{Y_{i1}|\, T = 1\} - E\{Y_{i0}|\, T = 0\} + E\{Y_{i0}|\, T = 1\} - E\{Y_{i0}|\, T = 1\}$$

$$\widehat{Impact} = (\{Y_{i1}|\, T = 1\} - \{Y_{i0}|\, T = 1\}) + (E\{Y_{i0}|\, T = 1\} - E\{Y_{i0}|\, T = 0\})$$

$$\widehat{Impact} = Average\ Treatment\ Effect\ on\ Treated + (E\{Y_{i0}|\, T = 1\} - E\{Y_{i0}|\, T = 0\})$$

$$\widehat{Impact} = Average\ Treatment\ Effect + Bias$$

Our aim is to estimate the ATE, but we have no idea about Bias as $E\{Y_{i0}|\, T = 1\}$ is not observed. All the impact assessment methods are aimed at minimizing the bias. Further, there are few important categories of treatment effects, namely ATT, ATU and ATE.

ATT: Average Treatment Effect on Treated: Impact of the treatment on treated Units

ATU: Average Treatment Effect on Untreated: Impact of the treatment on Untreated units (What will happen if the untreated units get treatment)

ATE: Average Treatment Effect: Impact on the entire population.

When you decide to do an Impact Assessment study, try to understand which of the treatment effect indicator is most suited. In most of the applied economics papers, ATT is more relevant than ATE (In most scenarios 100% population in the treated group is unrealistic).

Improper use of term 'Impact' in Economics literature

In many economics literature, unfortunately, the term impact is loosely used. We highlight few such cases and elaborate on why the term in inappropriate. a) Simple difference of value of outcomes (say Yield) across treated and control, b) Percentage change in income across treated and control unit, and c) Working out cost and returns (budgeting). Going back to the impact assessment notations, we have established that (Difference in mean of Treated and Control) + (Bias) = Estimator of Impact. In all the cases mentioned above, only difference in mean value of outcome is considered and the direction or magnitude of bias is not known. Without adjusting for bias, the difference in outcome of interest are useless and cannot be considered as evidence for causal relationship.

Then, what is Impact Assessment?

Measure how much beneficiaries are affected by an intervention (new extension technique, technology, etc.). Compare units receiving a treatment to what would have happened in absence of that treatment, for the same point of time. In essence, the objective of credible impact assessment is to make it data exchangeable; the impact should be same if the control unit is given the treatment / treatment group is not given the treatment. We will elaborate more on it, while dealing with specific methods. So, when can we make this causal claim?

Three basic conditions to make causal claims (Adopted from Antonakis et al, 2010)

Let us denote 'E' be the effect (May be increase in income) and 'X" be the cause. Then to establish causation, there are 3 basic conditions

a) X must precede Y

b) There is a strong correlation between X and E

c) The relationship between the X and E should not be explained by other causes.

The point a) indicates that we should make sure that X happened before E. If there is two-way relationship, we require methodological adjustments to make causal claims. The point b)- the true and unarguable relationship between Effect and Cause requires us to have an impact pathway and strong theoretical support to say that why and how X Causes Y. However, even more challenging aspect is the last condition. More on it later, for now let us examine what are impact pathways.

Impact Pathways

Impact pathway are the conceptual or theoretical explanation that elaborates how the intervention affects the outcome of interest. The mechanism by which the intervention affects the outcome and intermediate variables should be highlighted. A good schematic impact pathway is provided here.

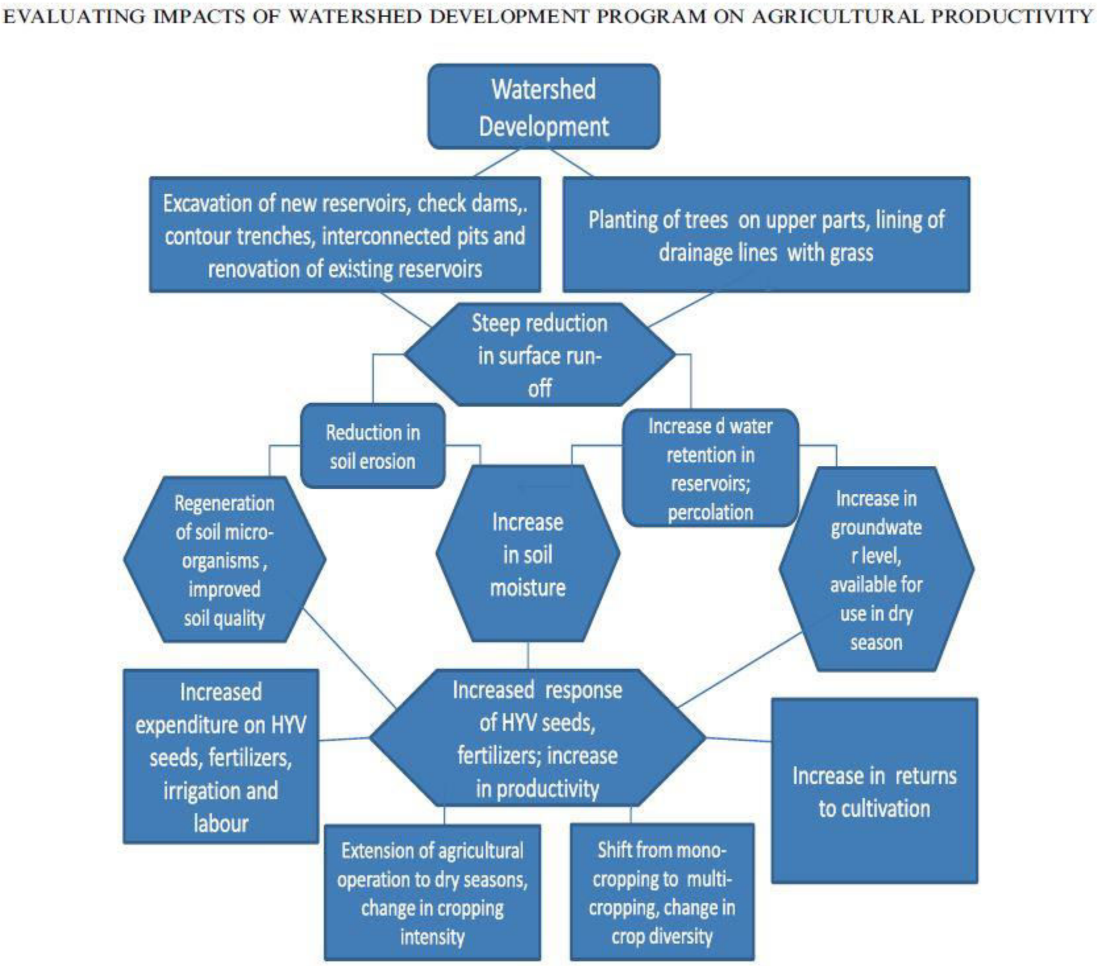


Figure 1. *Pathways through which watershed development impacts agriculture.*

Source: Datta 2015.

Another example of Impact pathway for lending through KCC, is provided in the form of verbal conceptual framework. (Kumar et al 2020)

"Farmers decide on the level of inputs needed to maximize their utility function based on various constraints, one of which is liquidity (Akudugu, 2012; Mani, 2018; Narayanan, 2015). Small and marginal farmers, with little savings, cannot invest in a profit-maximizing level of inputs unless they get credit (Guirkinger and Boucher, 2008; Karlan et al., 2014; Barrett et al., 2010). Not only does access to financing matter, but also its timeliness matters. The farmers who have access to a Kisan Credit Card are assured of credit and can withdraw the amount they see fit (Diwas et al., 2012; GoI, 2017; Kumar et al., 2011; Narayanan, 2016), which can promote an optimum use of resources. The program also has provisions for short-term and long-term investments in terms of assets, which can further increase productivity. The income increase resulting from higher input use and short-term investments creates what is called a 'liquidity effect' in literature (Binswanger and Khandkher, 1992; Narayanan, 2016). In addition, the KCC scheme has

provisions for consumption credit to facilitate consumption smoothening and prevent credit diversion for consumption purposes. With access to a KCC, farmers' dependence on informal sources of credit—like moneylenders who charge exorbitant interests' rates—can also be reduced. In this study, we empirically estimate the impact of the KCC scheme on input use, income, and farmers' dependence on moneylenders for credit. "

Confounders and confounding effects

When there are confounders, the relationship between Effect (E) is not only affected by Cause (X), but also affected by confounders, violating the condition c) mentioned before. Formally, Confounders are those variables which can simultaneously affect the X and E. For example, education of a farmer is a confounding variable. Educated farmers are more likely to adopt new technology (which is the cause 'X') and similarly, education also affects the yield (difference in yield is our 'E'). But one should note that mediating variables are not necessarily confounders. For example, because of the promise of new technology, farmers may use better quality inputs and hence gets more yield. Here fertilizer use is affecting only the outcome and not selection 'X'. The intermediate variables in the impact pathways are not always confounders. When there are confounders, the variable 'X' is not exogenous – it is affected by the confounders. Hence, the error term and 'X' are correlated, resulting in biased estimates. This is also similar to problem of selection bias- a case of specification bias/ omitted variable bias, where the estimate of impact is obtained ignoring the selection process/ equation.

It is easy to visualize the problem in the following way: When there are confounders, it means there are some variables that affect the Cause (X). Let us think that Cause (X) is a variable indicating adoption. If there are some variable affecting X, it is implicit that the selection into 'X' is not random and it depends on set of covariates and we observe X* not X.

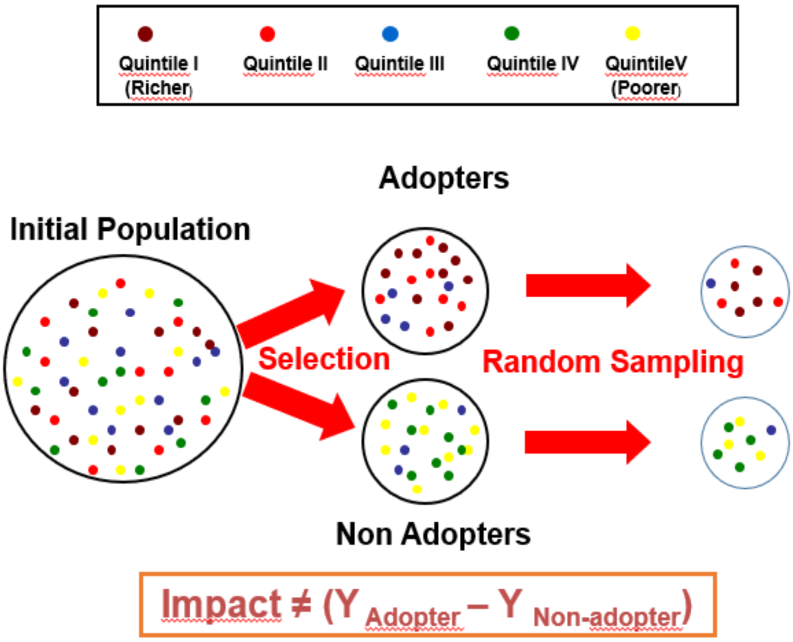
$X^* = f\,(\text{correlates of selection}) + u_i$

$E = f\,(X^* | \text{ Control variables}) + e_i$

The two error terms u_i and e_i are in fact correlated as there are variables which affects both X and E. This results in biased estimates.

Can random sampling help?

People often confuse random allocation with random sampling. Random sampling makes sure that the sample is representative of the population. In population itself, the adopters and non-adopters are different with respect to observables, then the sample drawn from random sampling will also have similar structure. So, random sampling is of no help in presence of selection bias. Selection bias is not a problem associated with sampling; it is a problem of non random allocation in the population itself.



Causal inference methods

Any book on Impact assessment invariable quotes "Correlation is not Causation" meaning that the association doesn't is not a sufficient condition for establishing causation. In presence of confounders, it is not straight forward to establish causation. One immediate question is; why can't we use a simple dummy variable regression to estimate impact. Such model will have outcome variable as dependent variable and a dummy variable to indicate treatment status, along with other control variables. However, the coefficient of the treatment variable cannot be considered as an estimate of impact. This is because of the reason that we made earlier; Because of the confounders, the dummy variable for treatment is not exogenous. Thus, the error term and the dummy variable are correlated and the estimate of impact could be underestimated/ overestimated or have a different sign altogether. This is what we call as selection bias. Figure 1 and 2 elaborate this point.

Estimating Impact in presence of a counterfactual outcome: An ideal scenario

When the treated units and the control units are similar with respect to observables (on an average), and the only difference between the two groups is the treatment itself, then the outcome of the control group becomes the counterfactual outcome. However, this can be achieved only when the treatment allocation is random.

In this case, there is no selection equation- X is random

Since data is exchangeable, $(E\{Y_{i1} | T = 1\} = E\{Y_{i1} | T = 0\})$, hence bias is also zero. So, if this condition is satisfied, the simple difference in mean itself is impact. Having said it, it is very rare for an economist to get his/her hand on such data. In most of the observational studies, counterfactual is rarely observed as most of the programs/ interventions, the beneficiaries are selected non- randomly. For example, a new variety will be mostly adopted by innovative, motivated farmers. There are a very few programs in the world where the program allocation is done randomly; like progress scheme of Mexico. In absence of random allocation, with and without comparison (or treated and control comparisons) are biased. Even before and after comparisons are biased because of element of time.

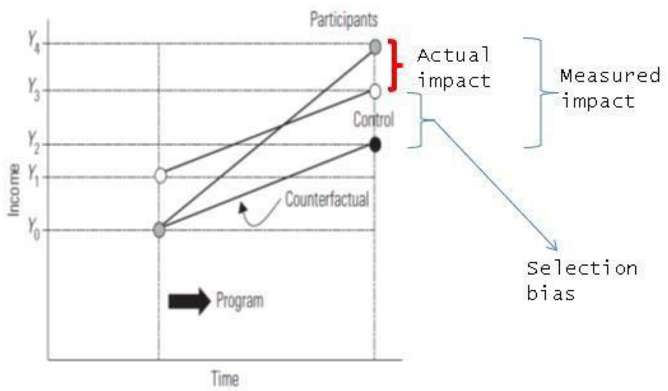


Figure 1: Bias in before and after trails

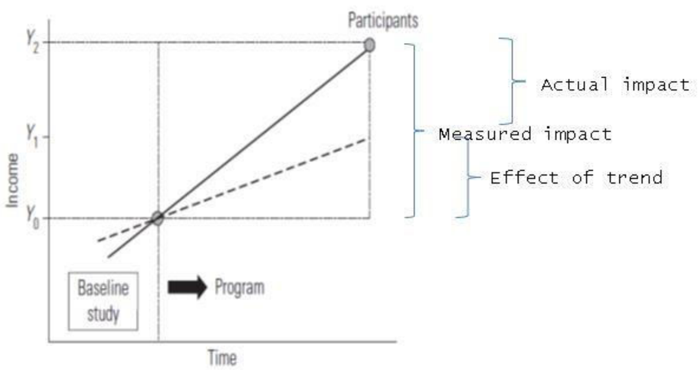


Figure 2: With and without comparisons

Further Readings

Cunningham, S. (2020). Causal Inference. The Mixtape, 1. Yale University Press.

Khandker, Shahidur R. Gayatri B. Koolwal, Hussain A. Samad. (2010). "Handbook on impact evaluation : quantitative methods and practices". The International Bank for Reconstruction and Development / The World Bank. Washington DC.

References

Datta, N. (2015). Evaluating impacts of watershed development program on agricultural productivity, income, and livelihood in bhalki watershed of Bardhaman District, West Bengal. *World Development*, *66*, 443-456.

Kumar, A., Sonkar, V. K., & Aditya, K. S. (2021). Assessing the Impact of Lending through Kisan Credit Cards in Rural India. IFPRI Discussion paper, International Food Policy Research Institute, Washington, DC.

An application of Difference-in-Difference model – A case study in the Nagarjuna Sagar Project of Krishna river basin

M Balasubramanian

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction

Climate change has tremendous effect on many parts of the world influencing people life in all aspects. It adversely affects agriculture, food security, water resources and biodiversity as a whole and its impact on agriculture has great significance in the world economy. Production of rice, wheat and maize in the past few decades has declined in many parts of the South Asia due to water stress arising partly from increasing temperature, increasing frequency of El Nino and reduction in the number of rainy days (Cruz *et al.,* 2007).Adaptation in the context of climate change comprises the measures taken to minimize the adverse impacts of climate change. Agricultural adaptation to climate change is a complex, multidimensional, and multi-scale process that takes on a number of forms (Bryant *et al.,* 2000). The Indian Council of Agricultural Research, state agricultural universities and international institutes are proposing various adaptation strategies to farmers (www.nicra-icar.in). In purview of this ClimaAdapt project (funded by the Ministry of Foreign Affairs, Royal Norwegian Embassy, New Delhi) was implemented from 2012-2015 to improve the resilience of the agriculture and water sectors in the state of Andhra Pradesh, Telangana and Tamil Nadu. The main objective of the project was to improve the adaptive capacity of stakeholders and farmers' groups. In course, situation analysis was carried out for the study areas and found that farming practices, water, insurance, capacity building, gender, information and services are the important focus areas to address the climate change impacts. The water saving technologies, alternate livelihood practices, capacity building programs and stakeholder integration were proposed in the project areas. Consequently, adaptation strategies were implemented in the project area for the last 4 years. Hence, the present study aims to know the socio-economic impacts of the selected water saving interventions under NagarjunaSagar Project in the ClimaAdapt program.

An impact evaluation is essentially a problem of missing data, because one cannot observe the outcomes of program participants had they not been beneficiaries. Impact evaluation is an effort to understand whether the changes in well-being are indeed due to project or program intervention. Specifically, impact evaluation tries to determine whether it is possible to identify the program effect and to what extent the measured effect can be attributed to the program and not to some other causes. The main challenge of an impact evaluation is to determine what would have happened to the beneficiaries if the program had not existed. A beneficiary's outcome in the absence of the intervention would be its counterfactual. So the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in light of this missing data. Without information on the counterfactual, the next best alternative is to compare outcomes of treated with and not been treated. Successful impact evaluations hinge on finding a good comparison group: - (a) create a comparator group through a statistical design (b) modify the targeting strategy of the program itself. A number of different methods can be used in impact evaluation theory to address the fundamental question of the missing counterfactual. Difference –in-Difference (DD) methods assume that unobserved selection is present and that it is time invariant and the treatment effect is determined by taking the difference in outcomes across treatment and control units before and after the program intervention. DD methods can be used in both experimental and non-experimental settings and uses both panel data and cross section data taken at repeated time interval.

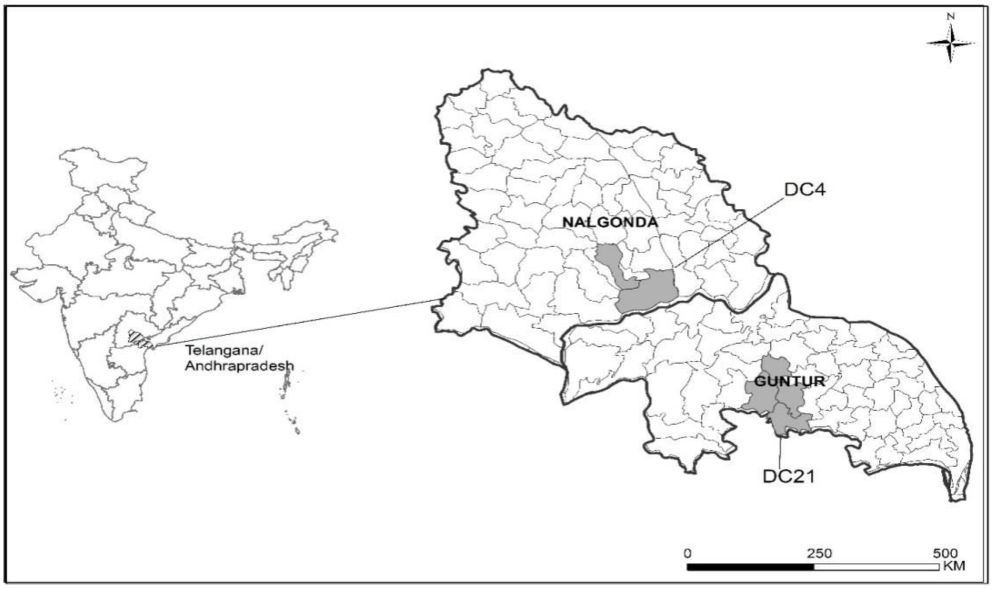
The Difference-in-Difference (DiD) approach is a research design for estimating causal effects. It is popular in empirical economics, for example to estimate the effects of certain policy interventions and policy changes that do not affect everybody at the same time and in the same way. It is used in other social sciences as well. (Cook and Campbell (1979), Rosenbaum (2001), and Shadish et al. (2002)).DiD could be an eye-catching option when using research designs based on controlling for confounding variables or using instrumental variables is deemed inappropriate, and at the same time, pre-treatment information is available. The DiD design is usually based on comparing *de facto* four different groups of objects. In many applications, "time" is an important variable to differentiate the groups. The idea behind the empirical strategy is that if the two treated and the two non-treated groups are subject to the same time trends, we use the mean changes of the outcome variables for the non-treated over time and add them to the mean level of the outcome variable for the treated prior to treatment to obtain the mean outcome the treated would have experienced if they had not been subjected to the treatment.

The History of DiD

The first scientific study using explicitly a DiD approach done by Snow (1854). He was interested in the question whether cholera was transmitted by (bad) air or (bad) water. He used a change in the water supply in one district of London, namely the switch from polluted water taken from the Themes in the centre of London to a supply of cleaner water taken upriver. Rose (1952) investigated the effects of a regime of 'mandatory mediation' on work stoppages by a DiD design. He tested the effectiveness of mandatory mediation in preventing work stoppages, it is necessary to make two simultaneous comparisons: (1) comparisons of states with the law to states without the law; (2) comparisons of the former states before and after the law is put into operation. The first comparison can be achieved by taking percentages of each of the three states to the total United States, for the measures used. The second comparison can be achieved by setting the date of the passage of the law at zero for each of the states. Obenauer and von der Nienburg (1915) analyzed the impact of a minimum wage by using an introduction of the minimum wage (in the retail industry) in the state of Oregon that, for a particular group of employees, led to higher wage rates in Portland, the largest city, compared to the rest of the state. Lester (1946) was concerned with the effects of wages on employment. He based his analysis on a survey of firms that had operations in both the northern and the southern US states. His idea was to compare employment levels, before and after various minimum wage rises, of groups of firms with low average wages to groups of firms with higher wage levels. The wage bills of the latter were naturally only mildly affected, if at all, by the rise in the minimum wage. An important aspect of DiD estimation highlighted by these early applications is that it does not require high powered computational effort to compute the basic DiD estimates, at least as long as further covariates are not needed and no complicated inference methods are used. Rose (1952), uses changes in state laws and regulations to define pretreatment periods (prior to the introduction of the policy) and unaffected comparison groups (states having a different policy than the one of interest). Later on, DiD designs have been used to address many other important policy issues, like the effects of minimum wages on employment (e.g., Card and Krueger, 1994), the effects of training and other active labor market programmes for unemployed on labor market outcomes (e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Heckman and Robb, 1986; Heckman and Hotz, 1989; Heckman et al., 1998; Blundell et al., 2004), the effect of immigration on the local labor market (e.g., Card, 1990), or the analysis of labor supply (e.g., Blundell et al.,1998). Methodological extensions of DiD methods often focus on this standard two periods, two groups setup; see, e.g., Heckman et al., 1997, Heckman et al., 1998, Abadie (2005), Athey and Imbens (2006), Qin and Zhang (2008), Bonhomme and Sauder (2011), de Chaisemartin and D'Haultfœuille (2017), Botosaru and Gutierrez (2018), Callaway et al. (2018), and Sant'Anna and Zhao (2020).

Methodology

The study was taken-up under NagarjunaSagar Project (NSP) of Krishna river basin covering both right and left canal of Andhra Pradesh and Telangana states (Figure 1). The NSP provides 7,465 mm3 water to irrigate 0.89 million hectares from five districts (Khammam, Krishna, Nalgonda, Guntur and Prakasam). The NSP have two main canals namely Jawahar (right canal) and LalBahadur (left canal) covering five districts. The right canal flowsthrough 203 km and 172 km on the left canal irrigating 0.47 million ha and 0.42 million ha area respectively. The present impact study was undertaken in the Nalgonda and Guntur districts where the climaAdpat project was implemented particularly focusing Distributary Committees 4 and 21in respective districts (Figure 1). After finalization of the questionnaire, data collection was carried out with the sample farmers through interview method, initially, mandals were selected purposively from the project implemented area where farmers, agricultural and irrigation officials and other line departments were involved in the program as stakeholders. The farm household samples were purposively selected for collecting the data from the selected mandals to cover both adopters and non-adopters. Farmers adopting water saving adaptation practices like Alternate Wetting and Drying (AWD), Modified System of Rice Intensification (MSRI), and Direct Sowing of Rice (DSR) were selected for analyzing the impact of adaptation and trainings. Despite of the various training programs attended for crop related practices, agricultural allied activities and risk management aspects, focus was given towards water saving interventions for sampling.The total sample covered for the study was 178, of which 138 are adopters and 40 non-adopters. Information on the adaptation of technologies in the others areaswasalso collected from the state/district agricultural department,irrigation departmentand scientists to know the upscaling impacts. Data was also collected from village knowledge centers (VKCs) on the trainings, dissemination information and adaption rate. Secondary data regarding agriculture information related to climate change aspects, farmer's response towards water saving interventions, adoption rate and their personal opinion about the climaadapt program both at the field level and macro level in a district was collected from respective Mandal Agricultural officer. Irrigation particulars like amount of canal water release in a period, discharge rate, loss in the distribution, relation with water user associations and any special scheduling of water for interventions were collected from executive engineer at Mandal level with the help of Assistant engineer at field level, where they have given their own experience about the delivery system, problems in scheduling and information about the meetings with agricultural and revenue officials to support the tail end farmers need.



Technology is nothing but the application of improved knowledge on production relationships and thus technology has the effect of raising the production function. More output per unit of input (water) is possible with the new technology such as water saving interventions (AWD, MSRI and DSR practices). This indicates that production can be increased with improved technology through the same amount of inputs that were used with traditional technology or the current production level can be reached with fewer inputs with improved technology. Consider the Figure 4, where curve AA refers the traditional irrigation technology production function, curve AB refers the improved technology production function and curve AC refers the improved technology with capacity building. With X units of water, traditional technology produces Y1 units of output, water saving interventions (DSR, AWD, MSRI) produces Y2 units of output whereas improved technology with capacity building produces Y3 units of output. The difference between Y2 and Y3 is the additional output due to capacity building/trainings (Palanisami K et al., 2014).

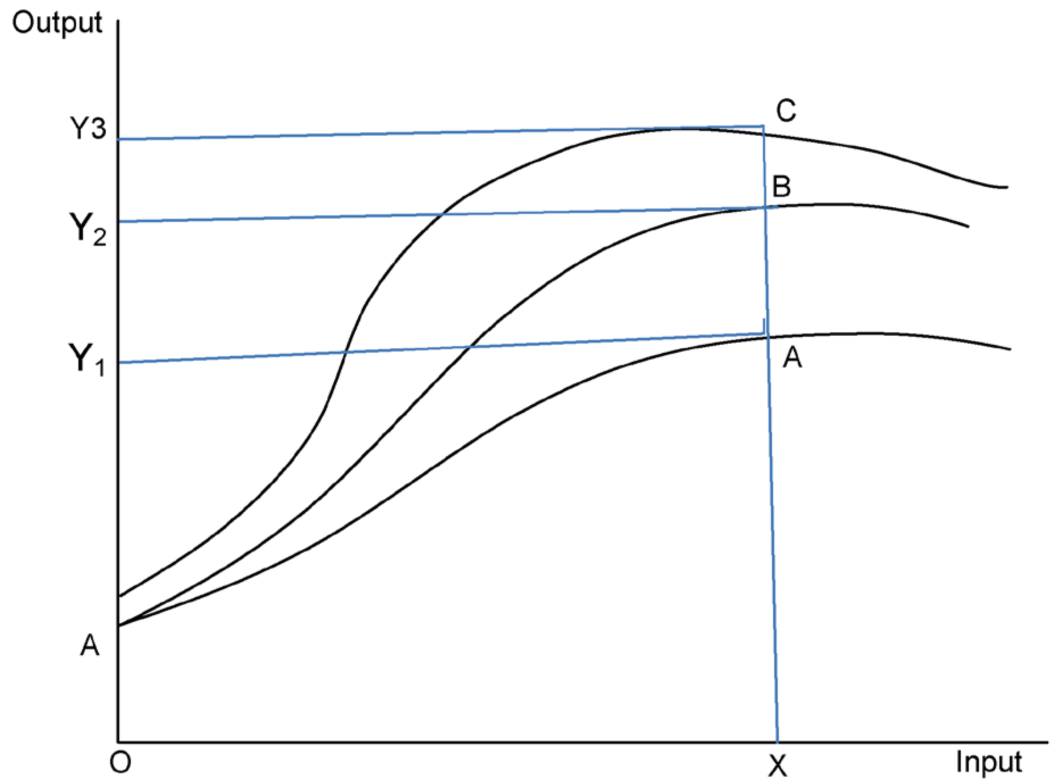


Figure 2.Technology adoption and crop yield. Note: AA refers the production function with traditional irrigation technology, AB refers the production function with water saving interventions; AC refers the production function with water saving interventions along with capacity building.

Several tools or approaches are used for impact evaluation. The most commonly used tools are the financial measures like the benefit-cost (B-C) ratio and internal rate of return (IRR). The major problems with this approach is that the benefits and costs are calculated using either before and after or with and without concept which ignores some of the benefits that are considered as residual which may occur even without the intervention. Hence an approach that considers both, with and without as well as before and after situations is important. The approach of any analysis of impact can be accomplished into two ways. Firstly, "with project" parameters compared to the "pre-project" situation gives the incremental benefits due to the project. But these increments in the parameters intrinsically include the changes due to state of art of technology. Thus sometimes, the benefits may be exaggerated. Secondly, the literature on project analysis unanimously suggests the use of comparison between the "project parameters" with the "non-projectcontrol region'. This method automatically incorporates the correction for the impact of technology in the absence of the project. For the present study, the information was collected for the pre and post-project period and compared with the control as well. Hence, the approach is a combination of both with and without and before and after approach i.e. difference- in-difference method (Table 1).

Table 1: Impact assessment of capacity building and implementation of WSI by Double Difference method

S.No	Particulars	Adopters	Non- Adopters	Differences across groups
1	After adapting WSI method	D1	C1	D1 – C1
2	Before adapting WSI method	D0	C0	D0 – C0
3	Difference across time	D1 – D0	C1 – C0	Double difference (D1 – C1) – (D0 – C0)

Farm level data were collected from both types of farmers i.e. who have participated in the capacity building program and adopted the water saving interventions method in the field and who haveparticipated in the training program but not adopted. This enables the use of the double difference method to study the impact of the capacity building program on water saving intervention methods. The resulting measures can be interpreted as the expected effect of implementing the capacity building program on water saving intervention method. The columns distinguish between groups with and without the program and the rows distinguish between before and after the program. Before the capacity building program, one would expect the average yield of paddy crop could be similar for the two groups, so that the quantity (D0 - C0) would be close to zero. Once the capacity building program has been implemented, however, one would expect yield differences between the groups as a result of the improvement in knowledge of the farmers about the water saving techniques due to the program. The impact of the program, however, would be better assessed considering any pre-existing observable or unobservable differences between the two randomly assigned groups i.e. the double-difference estimate, which is obtained by subtracting the pre-existing differences between the groups, (D0 - C0), from the difference after the program has been implemented, (D1 - C1). This is explained in Figure 3.

Double Difference

$$= E(Y_1^T - Y_0^T | T_1 = 1) - E(Y_1^C - Y_0^C | T_1 = 0) - - - - (1)$$

Where, Y_t^T and Y_t^C respectively denote the outcome responses for the trained and control groups at period t (= 0, 1) where the time period t = 0 corresponds to the period before program implementation and the period t = 1 corresponds to after program implementation. Further, $T_1 = 1$ means presence of the program at time t = 1 and $T_1 = 0$ means absence of the program. The first term in Equation (1) represents the average difference between before-after for the trained group and hence it is given by

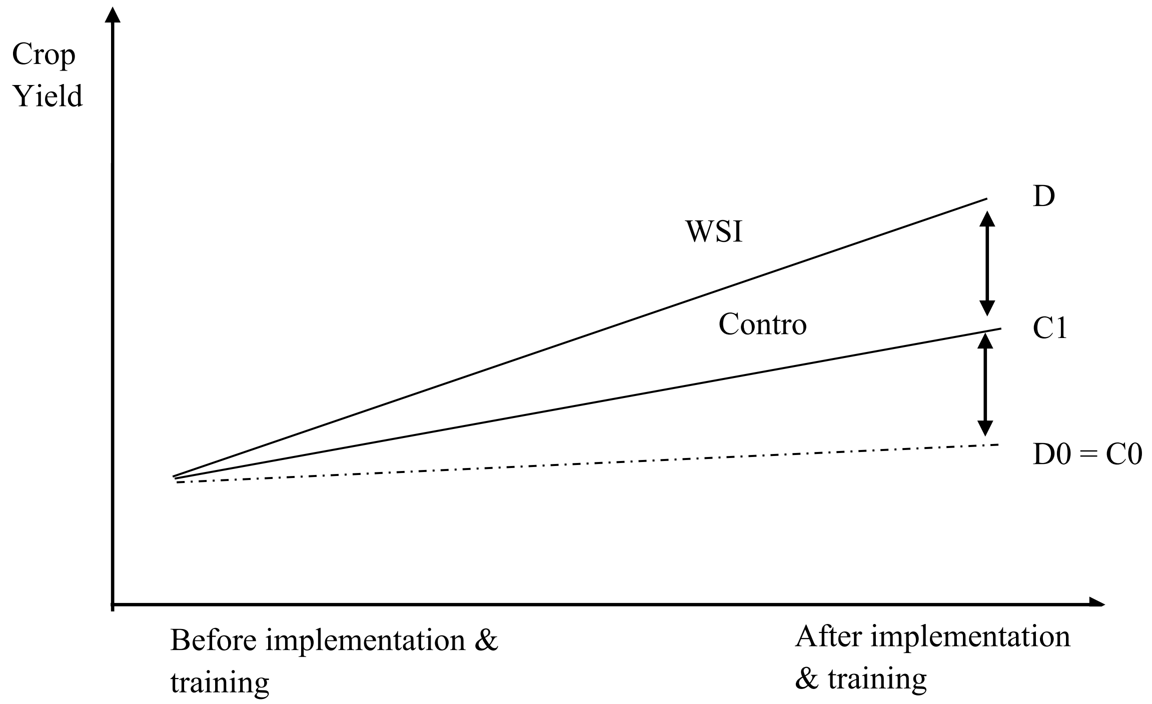


Figure 3: Illustration of impact of capacity building program by double difference method.

$$= E(Y_1^T - Y_0^T \mid T_1 = 1) = \frac{1}{N_T}\sum_{i\varepsilon T}(Y_{i1} - Y_{i0}) = \bar{y}_{r1} - \bar{y}_{r0} ----(2)$$

Similarly, for the control group the second term is given by

$$= E(Y_1^C - Y_0^C \mid T_1 = 1) = \frac{1}{N_C}\sum_{j\varepsilon C}(Y_{j1} - Y_{j0}) = \bar{y}_{c1} - \bar{y}_{c0} ----(3)$$

Substituting these values in (1), the impact of the program can be obtained as

$$Impact = (\bar{y}_{r1} - \bar{y}_{r0}) - (\bar{y}_{c1} - \bar{y}_{c0}) ----(4)$$

The same results can be obtained by following a regression approach as follows: For each observation i, let us define a variableδ_i as $\delta_i = 0$if the observation is from the control group and $\delta_i = 1$if it is from the trainedgroup. Similarly for each observation i define a variable $T_i$as$T_i = 0$ if the observation belongs to time t = 0, that is before the WSI implementation andcapacity building program and$T_i = 1$ if the observation belongs to time t = 1, that is, after the program. Now form the regression equation,

$$y_i = a + b\delta_i + cT_i + d\delta_i T_i ----(5)$$

Observation belongs to	***Δ***	***T***	***yi***
Control group before the program	0	0	$\bar{y}_{co} = a$
Control group after the program	0	1	$\bar{y}_{c1} = a + c$
Trained group before the program	1	0	$\bar{y}_{ro} = a + b$
Trained group after the program	1	1	$\bar{y}_{r1} = a + b + c + d$

So using equation (4)

Impact of the program = ((a+b+c+d) – (a+b)) – ((a+c)-a) = d $----$(6)

Results and discussion

The data collected from primary and secondary sources was analyzed for DC- 4 & DC – 21 areas of Telangana and Andhra Pradesh states. Of total 178 sample households 60 per cent of the farmers had middle and high school level of education which in turn boost their understanding capacity about the new techniques. Around 27 per cent of non- adopters have college education. At the same time, 18 per cent of adopters are illiterate, still they attended training and adopted new water saving interventions. So, education alone not influences the adaptation rate among the farmers rather contact with lead/ contact farmers, agriculture and irrigation officials and risk taking ability also influence the adoption behavior of the farmers. Around 44 percent of adopter has experience of farming above 30 years followed by middle age farmers sharing 42 per cent having farming experience of 15-30 years. Availability of canal water in kharif is 113 days as low as 90 days reported by the non-adopters against 137 days given by the adopters. Though, availability of canal water in any season is same for the command area but between head and tail reach varies hardly between 20 - 25 days (Irrigation official).It is observed that 86 per cent of adopters attended training on AWD followed by 70 per cent on DSR method and 43 per cent attended trainings on MSRI. In total, 83 per cent of sample farmers attended trainings on AWD method followed by 55 per cent on DSR and 47 per cent on MSRI including non-adopter category. Paddy is cultivated predominantly under irrigated condition where crop failure seldom occurs. Unlike for many other crops, the minimum support price (MSP) scheme is effectively implemented for paddy crop. The partial analysis of economic impacts showed that economic benefits in general by applying the technology were actually perceived by 81% of the farmers. The number of irrigations was reduced by 28% on average. Such a level of monetary profit, however, was observed only in cases where the payment system for irrigation is consumption-based compared to a fixed rate system. Weeds, however, were affected differently by the AWD regime, which often led to increased occurrence of weeds resulting in increased expenditure in hired labor for hand weeding. Farmers under DSR effectively managed the increased weed growth through the application of herbicides. Yields of rice increased by about 0.4 to 0.5t/ha, which is equivalent to about 10%.The net revenue for the DSR was also higher than the other two interventions due to the reduced cost.

The present study, aims to analyze the impact of capacity building and implementation program on water saving intervention (WSI) methods. Whether the capacity building and implementation program created any kind of improvements in yield or not can be addressed. The capacity building program on WSI created additional knowledge on water use efficiency and water conservation practices with the main aim of improving the crop yield. When the yield parameter was taken into consideration farmers were more interested to adapt new techniques which had direct impact on income leading to betterment of life. Hence the study sample was segregated into four methods of cultivating paddy for further analysis. The mean yield difference was presented in table 24.Farmers who have adapted new interventions got higher yield than non-adopter or control group. In the present study double difference method of analysis was used to assess the impact of capacity building program and implementation on WSI methods and crop yield. The three different kinds of WSI methods and one control group were compared and assessed the net impacts of the program.

Table 2: Rice yield under different WSI method by farmers in DC-4 & 21 area

Sl.No	Intervention	Sample	Mean	Minimum	Maximum	SD
1	**AWD**					
	Before	56	53.03	47.50	65.00	3.86
	After	56	64.33	55.80	72.50	5.3
2	**MSRI**					
	Before	38	58.57	43.00	80.00	8.8
	After	38	69.60	53.00	88.00	9.8
3	**DSR**					
	Before	44	55.54	38.00	75.00	7.2
	After	44	65.04	50.00	95.00	9.9
4	**Control**					
	Before	40	55.75	39.38	75.00	8.3
	After	40	57.42	41.25	75.00	7.36

From the above table 2, it is inferred that of all three methods average maximum yield was obtained from MSRI method followed by DSR and AWD methods with 69.6, 65.04 and 64.33 qtl/ha respectively. In case of control or non-adopters average maximum yield was 57 qtl/ha.

Table 3: Mean yield difference of WSI adapted

S.no	Observations from	AWD	MSRI	DSR
1	Trained farmers before (a)	53.03	58.57	55.54
2	Trained farmers after (a+c)	64.33	69.60	65.04
3	Control group before program (a+b)	55.75		
4	Control group after program (a+b+c+d)	57.42		
5	Net impact due to capacity building and interventions (d)	9.63	9.63	7.83

The average mean yield difference among non-adopters before and after the program was 1.67 qtl/ha (table 3) due to accumulated knowledge, experience on farming, use of better quality inputs and technology growth. Similarly, the yields for AWD adopters were 53.03 and 64.33 qtl/ha, following the yield difference of 11.3 qtl/ha. After capturing the effect of training or capacity building program on WSI yield difference was 9.63 (11.3-1.67) qtl/ha.

Likewise for other methods had been calculated and presented in the table 27. Net impact due to capacity building and implementation for DSR method was low than other two methods because of less use of inputs and other cultural practices. The results of the double-difference method using regression analysis on rice yield are presented in table 4.

Table 4: Regression analysis on impact of the training program on rice yield

Sl.No	Method	Constant	Δ	T	δT	R^2 Value
1	AWD	55.75 (56.83)	-2.714** (-2.113)	1.67 (1.20)	9.623*** (5.29)	0.34
2	MSRI	55.75 (40.74)	2.60 (1.32)	1.67 (0.86)	9.38*** (3.38)	0.28
3	DSR	55.75 (42.21)	-0.352 (-0.193)	1.67 (0.89)	7.78*** (3.01)	0.18

Note figure in the parenthesis indicates t values at 5 (**) per cent and 1(***) per cent significant level

It is inferred that capacity building with implementation and technology growth by time (T) interaction has significant impact for all three methods at 1 per cent level. It was represented by the effects of capacity building program on WSI (with and without) and technology growth (before and after) i.e. combined effect on yield was significant for all three methods indicating the importance of capacity building program and technology transfer overtime. There was significant difference between adopters and non-adopters on yield of WSI. The AWD adoption is able to overcome the yield losses by 2.71 times in the study area. Similarly, farmers adopting DSR were avoiding yield loss of 0.352 times. MSRI adoption has also significant impact at 1 per cent level in the interaction of capacity building program and technology transfer.

The R^2 is worked out to 0.34, 0.28 and 0.18 for AWD, MRSI and DSR respectively indicating the 34, 28 and 18 per cent of the variations were explained by the explanatory variables. The intercept term indicated the mean yield of the non-adopting farmers. It is evident that there is significant difference between yield of adopting and non-adopting farmers in the base period. Similarly, there is a significant increase in yield due to time period among the non-adopting farmers. It is evident that 1.67 qt/ha increase in yield was realized overtime period among the non-adopting farmers. The impact of capacity building program was significant on the expected positive line which showed that the program alone has increased the crop yield by 9.62, 9.38 and 7.78 through AWD, MSRI and DSR respectively.

Conclusion

Climate change is affecting the water resource availability and agricultural food production system in many parts of the country. To mitigate the climate change effects various adaptation strategies were developed and implemented under Nagarjunasagar project of Krishna river basin. The present study address the impacts of water saving interventions implemented through capacity building programs in the study area. The adaptation interventions are enhancing the performance of crop water use efficiency and riceyields through awareness and capacity building programs. Village knowledge centersestablished at cluster level are playing a key role in creating awareness on new interventions and providing online servicesto the farmers. Policy makers can understand the role of village knowledge center which can be initiatedin other parts of the state and make more sustainable where similar kind of situation exists. Significant rise in yield between adopters and non-adopters provide sufficient evidence to policy makers to promote more climate adaptation frameworks. This helps in improving water use efficiency and agriculture production to achieve food secure nation.

REFERENCES

Ashenfelter, O. and D. Card (1985), 'Using the longitudinal structureof earnings to estimate the effect of training programs'. The Reviewof Economics and Statistics 67, 648–660.

Blundell, R., A. Duncan, and C. Meghir (1998), 'Estimating labor supplyresponses using tax reforms'.Econometrica 66, 827–861.

Blundell, R., C. Meghir, M. Costa Dias, and J. van Reenen(2004), 'Evaluating the employment impact of a mandatory jobsearch program'.Journal of the European Economic Association 2,569–606.

Bryant, R. C., Smit, B., Brklacich, M., Johnston, R.T., Smithers, J., Chiotti, Q. and Singh, B. 2000.Adaptation in Canadian Agriculture to Climatic Variability and Change. Climatic Change, 45: 181-201.

Card, D. (1990), 'The impact of the mariel boatlift on the miami labormarket'. Industrial and Labor Relations Review 43(2), 245–257.

Card, D. and A. B. Krueger (1994), 'Minimum wages and employment:A case study of the fast-food industry in New Jersey and Pennsylvania'.The American Economic Review 84, 772–793.

Cook, P. J. and G. Tauchen (1982), 'The effect of liquor taxes on heavydrinking'. Bell Journal of Economics 13, 379–390.

Cruz, R.V., Harasawa, H., Lal, M., Wu, S., Anokhin, Y., Punsalmaa, B., Honda, Y., Jafari, M., Li, C. and HuuNinh, N. 2007. Asia. Climate Change (2007): Impacts, Adaptation and Vulnerability

Heckman, J. J. (1996), 'Comment on eissa: Labor supply and the economicrecovery act of 1981'. In: M. Feldstein and J. Poterba (eds.):Empirical Foundations of Household Taxation. pp. 5–38.

Heckman, J. J. and R. Robb (1986), 'Alternative methods for solvingthe problem of selection bias in evaluating the impact of treatmentson outcomes'. In: H. Wainer (ed.): Drawing Inferences from Self-Selected Samples. pp. 63–113.

Heckman, J. J. and V. J. Hotz (1989), 'choosing among alternativenonexperimental methods for estimating the impact of social programs:The case of manpower training'. Journal of the AmericanStatistical Association 84, 862–880.

Lester, R. A. (1946), 'Shortcomings of marginal analysis for the wageemploymentproblems'. American Economic Review 36, 63–82.

Obenauer, M. and B. von der Nienburg (1915), 'Effect of minimumwagedeterminations in oregon'. Bulletin of the U.S. Bureau of LaborStatistics, 176, Washington, D.C.: U.S. Government Printing Office.

Palanisami, K. et al. (2014).Enhancing the crop yield through capacity building programs: Application of double difference method for evaluation of drip capacity building program in Tamil Nadu State, India.Agricultural Sciences vol.5, no.1, 33-42.

Rose, A. M. (1952), 'Needed research on the mediation of labour disputes'.Personal Psychology 5, 187–200.

Rosenbaum, P. (2001), 'Stability in the absence of treatment'. Journalof the American Statistical Association 96, 210–219.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002), Experimentaland Quasi-Experimental Designs for Generalized Causal Inference.Boston: Houghton-Mifflin.

Simon, J. L. (1966), 'The price elasticity of liquor in the U.S. and asimple method of determination'. Econometrica 34, 193–205.

Snow, J. (1854), 'The cholera near golden square and at deptford'.Medical Times and Gazette 9, 321–322.

www.ccafs.cgiar.org

www.climaadapt.org

www.climarice.com

www.fao.org/docrep/013/am307e/am307e00.pdf

www.nicra-icar.in

www.worldbank.org

Synthetic control method for assessing impact of policy change

Prabhat Kishore

National Institute of Agricultural Economics and Policy Research

<u>Theoretical concept</u>

Basic evaluation designs that have often been used for impact assessment of any interventions are "with and without" and "before and after" approach. The "with and without" approach needs information of desired population with some unit with intervention and other without of it. For this, underlying assumption is, unit without intervention has to be good proxy for unit which has received intervention. Usually, this is an unrealistic assumption as every units of considered population may have some differences either observable or unobservable. In reality, it not possible to observe desired unit with and without intervention at same point of time. This is also known as problem of missing counterfactual. Other evaluation approach, "before and after" requires observations on the same units before and after the intervention. This approach is considered to have more credible control group as the desired unit are without intervention at earlier and same unit have received intervention latter. But, the difference on observed outcome before and after could not be assigned to only treatment as there could be other factor which could have influenced desired unit over time. To address the concerns of these two approaches, researchers have relied upon "Difference in Difference" (DID) for observational studies. DID combine a "with and without" with "before and after" approach wherein control group considered are subset of population which never received the intervention. An alternative approach has been the randomization of study unit. Random assignment of treatment creates a credible counterfactual that tells us what would have happened if the intervention does not take place. With this methodology, observed difference can be attributed to intervention alone. But in social science, randomization of treatment is subjected to time and money constraints. All these above stated methods are used to evaluate any intervention with help of affected individual unit.

However, in some cases intervention takes place at state or country and policy maker are interested to know impact of the intervention at that macro level. With traditional approach it seems to be difficult as the unit of intervention itself is single or may be few some time. So there is lack of sufficient number of treated and control unit for inferences. Major policy intervention occurred at macro level like Government of Bihar has repealed APMCs Act in 2006 with motif to remove barriers in agricultural marketing or Punjab Government enacted

Punjab Preservation of subsoil water act 2009 to regulate groundwater depletion in the state. To know the impact of these interventions at aggregate level, there is need to have a robust method which can provide suitable comparison unit for the inference. In this context, Synthetic Control Method provides new insight to tackle stated problems in impact assessment methodology.

Analytical method

In recent times, Synthetic Control Method (SCM) has been appearing in many research articles for impact assessment at aggregate level such as state or country. However, in agriculture SCM application is very few. The synthetic control method pioneered by Abadie and Gardeazabal (2003), bridged gap between qualitative and quantitative methodologies. SCM is a data-driven approach in choosing comparative units. It gives insight for systematic selection of comparison unit based on similarity of parameter considered for selected units. SCM construct counterfactual of treated unit by considering weighted average of non treated units based on parameter considered. In contrast to a difference-in-differences (DID) design, SCM does not give same weight to untreated unit in the comparison (Galiani and Quistorff, 2016). Further, it also allows the effects of observed and unobserved predictors of the outcome to change over time, while assuming that pre-intervention covariates have a linear relationship with outcomes post-treatment (Kreif et al., 2016). The advantage of constructing counterfactual unit with this method is that the pre-intervention characteristics of the treated unit can often be much more accurately approximated by a combination of untreated units than by any single untreated unit (Abadie et al. 2015). The central idea behind the synthetic control method is that the outcomes from the control units are weighted so as to construct the counterfactual outcome for the treated unit, in the absence of the treatment (Kreif et al., 2016). The weights estimated using pre-treatment data, can be applied to generate post-treatment outcomes for the synthetic unit. Those post-treatment outcomes can then be interpreted as if they were the counterfactual outcome values if treated and its synthetic track each other closely in pre intervention period. The divergence in outcome values between the synthetic and treated unit in the post-treatment period if the intervention has a significant impact.

In the recent time, desirable property of SCM and DID have been combined for the assessment of policy change that occurs at aggregate level (Dmitry et al., 2021)

Econometric model

Suppose there is S+1 state in India where one state got intervention and remaining non intervention states considered as potential control or donor pool. Let Y_{it}^{N} be the outcome that would be observed for state i^{th} at time t in absence of intervention where i= 1, 2...,S+1 and time t=1, 2,…,T. let T_0 be intervention year where $1 \leq T_0 < T$. Further, Y_{it}^{I} be the outcome that would be observed for unit i^{th} at time t if i^{th} unit for intervention in period T_0+1 to T. Here assumption is outcome of untreated unit does not affected by intervention in treated unit. Impact of intervention is quantified by δ_{it} where

$$\delta_{it} = Y_{it}^{I} - Y_{it}^{N}$$

Let D_{it} be the indicator which takes value 1 if unit i^{th} received intervention at time t otherwise zero i.e.

$$D_{it} = \begin{cases} 1 & if\ i = 0\ and\ t > T_0 \\ 0 & Otherwise \end{cases}$$

The synthetic control technique, subjects the comparison units' predictor variables' attribute data in the pre treatment period to a dual optimization process that minimizes:

$$\sum_{m=1}^{k} Vm\ (X_1m - X_0mW)^2$$

by selecting the optimal values of W and Vm where X_1m is the value of the m^{th} attribute of the treated unit; X_0m is a 1 x j vector containing the values of the m^{th} predictor attribute of each of the S potential comparison or control units; W is a vector of weights on control units; and Vm is a vector of weights on attributes of the control units such that they maximize the ability to predict the outcome variable of interest (Abadie et al. 2010). This optimization process minimizes prediction error between the actual and the synthetic in the pre-treatment period.

Y_1 is the observed outcome data for the treated, unit. Y_0W is the weighted average of outcome variables for the included control units. If there are no important omitted predictor variables then a reliable synthetic match will be created such that Y_1 -Y_0W, the distance between the actual unit's outcome variable and the synthetic unit's outcome variable will be small in the pre-intervention period (Abadie et al. 2010). This is particularly likely when the pre-intervention period is sufficiently long. If the outcome variable of the synthetic control diverges significantly from the actual outcome in the post-treatment period, the gap between actual and synthetic may be attributed to the effect of the treatment.

For post estimation the fake treatments are applied to donor units that were not subjected to the intervention to analyse the divergence between synthetic and treated unit. Basic idea is that replicating the same analysis should not generate a significant divergence between synthetic and actual outcomes in the absence of treatment. These tests bolster confidence in methodology. Creating a synthetic for each donor unit in the population enables researchers to ascertain whether the estimated treatment effect for the treated unit is of unique magnitude and direction.

ILLUSTRATION

In year 2006, government of Bihar has repealed its APMC Act of 1960s in order to open up space for private investment in new market to improve the market efficiency. Other state continued with their APMCs Act except Jammu & Kashmir, Kerala and Manipur. This intervention of government in form of APMCs repeal has been considered for further elaboration of chapter with motif to find out whether this has led any change in agricultural GDP of Bihar or not.

Here Bihar has been considered as treated unit and rest of the Indian states except Kerala, Jammu and Kashmir and Manipur were taken as control or donor pool. With help of Synthpackage of stata, SCM is being employed see impact of APMCs repeal. Agricultural GDP has been taken as outcome variables and predictor variables are agricultural area, gross cropped area, net irrigated area, cropping intensity and per cent electrified villages. Panel data constructed for considered outcome and predictor variable of desired Indian states for the period of 1984-2012. The length of the pre-intervention period over which prediction error is to be minimized, is about 20 years. Table 1 obtained as SCM result compare considered variables characteristics of Bihar with its synthetic. Average of 26 control states in pre intervention period depicted in last column does not provide suitable control group for Bihar. But synthetic produced with weighted average of considered control groups is similar to real Bihar.

Table 1: Comparison of variable characteristics in Pre-treatment for Bihar with its synthetic

Variables	Bihar		Average of 26 control state
	Real	Synthetic	
Agricultural land (thousand hectare)	9933.76	9914.75	6806.34
Net cropped area (thousand hectare)	6994.77	6981.73	4475.73
Net irrigated area (thousand hectare)	3342.52	3336.83	1854.18
Electrified villages (Per cent)	65.58	65.52	85.66
Cropping intensity (Per cent)	135.80	135.70	136.6

Figure 1 shows agricultural GDP of Bihar (blue bold line) and its synthetic (black dotted line) during 1985-2012. Synthetic Bihar's agricultural GDP very closely tracks the trajectory of real Bihar's agricultural GDP for entire pre treatment period. Close trajectory of real and synthetic Bihar in pre-treatment period indicate toward better approximation of the agri. GDP in post treatment period. Synthetic Bihar (black dotted line) in post treatment period represent trajectory of Bihar agricultural GDP in absence of intervention considered. The estimate did not turned out to be noticeable divergence between actual and synthetic Bihar in post treatment revealing toward the insignificant impact of repeal on agricultural GDP.

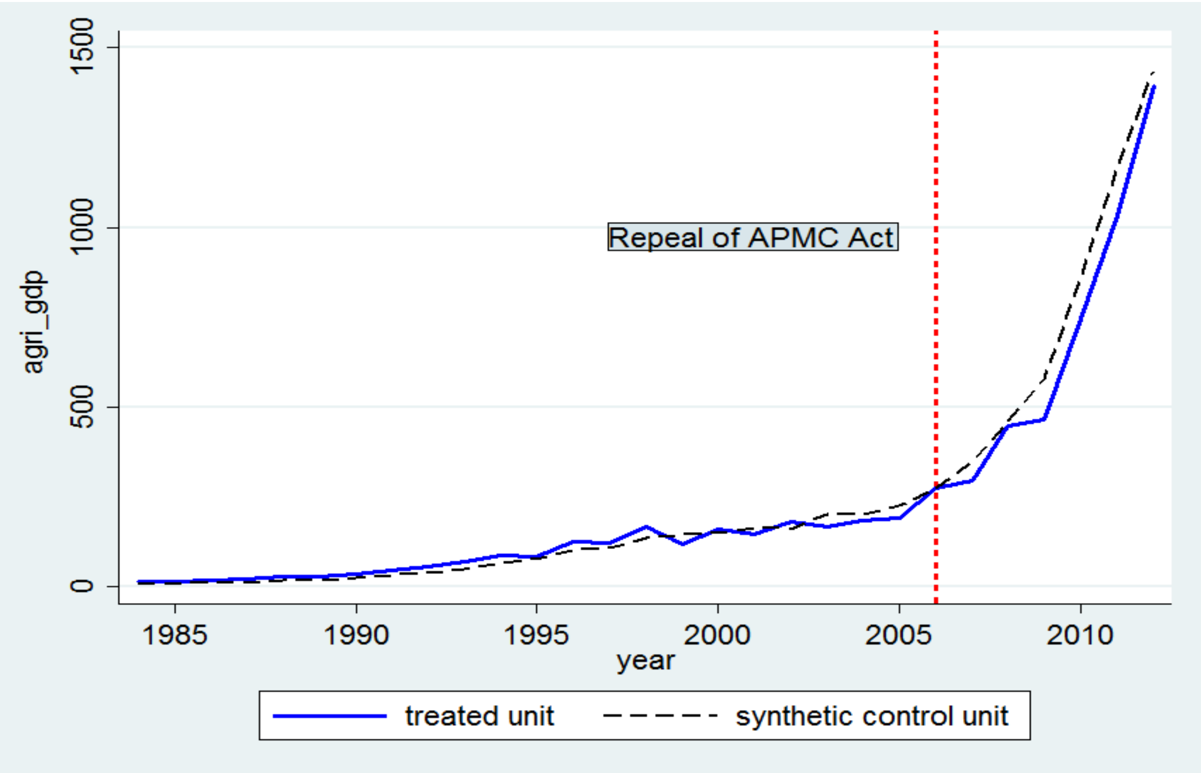


Fig 1: Trends in Agri. GDP: Bihar vs. synthetic Bihar

Table 2 present weights assigned to each state in order to construct counterfactual of Bihar agricultural GDP. This weight is being estimated based on similarity of variable characterise considered for the study i.e. more weight to particular state in constructing counterfactual if that state variable are more like treatment state.

Table 2: State weight in Synthetic Bihar

State	Assigned weight	State	Assigned weight
Andhra Pradesh	0.009	Mizoram	0.007
Arunachal Pradesh	0.210	Nagaland	0.006
Assam	0.010	Orissa	0.017
Delhi	0.008	Pondicherry	0.010
Goa	0.006	Punjab	0.028
Gujarat	0.006	Rajasthan	0.159
Haryana	0.011	Sikkim	0.009
Himachal Pradesh	0.024	Tamil Nadu	0.007
Karnataka	0.006	Tripura	0.011
Madhya Pradesh	0.027	Uttar Pradesh	0.184
Maharashtra	0.005	West Bengal	0.016
Meghalaya	0.223		

Post estimation

Placebo test done to check the validity of the result obtained following SCM to test whether result is driven by chance or it is factual. So, a series of placebo test iteratively applied to every other state in the donor pool. In each iteration, every state is assigned same treatment in same year and rest of the state shifted to donor pool including Bihar. This iterative procedure provides counterfactual of each state agricultural GDP and also distribution of estimated gaps for each state with its counterfactuals. Figure 2 displays the results for the placebo test conducted for each state considered in this study. Blue line presents treatment state; Bihar and other line represent state under donor pool. As the figure make apparent, trajectory for Bihar does not vary significantly relative to other state in donor pool after treatment applied. Other state considered in donor pool have same type trajectory even without any treatment. This reinforced the result earlier obtained that there is no impact on agricultural GDP with repeal of APMC Act. Ratio of post and pre root mean squared prediction error (RMSPE), ranked Bihar on 21th number out of 27 states considered. For significant impact there would have been wider gap between actual and synthetic trajectory of Bihar agricultural GDP. And this would have led to large value of post and pre ratio of root mean squared prediction error placing Bihar at first place. This result has further bolster result that there is no significant change in agricultural GDP after the repeal of APMC Act.

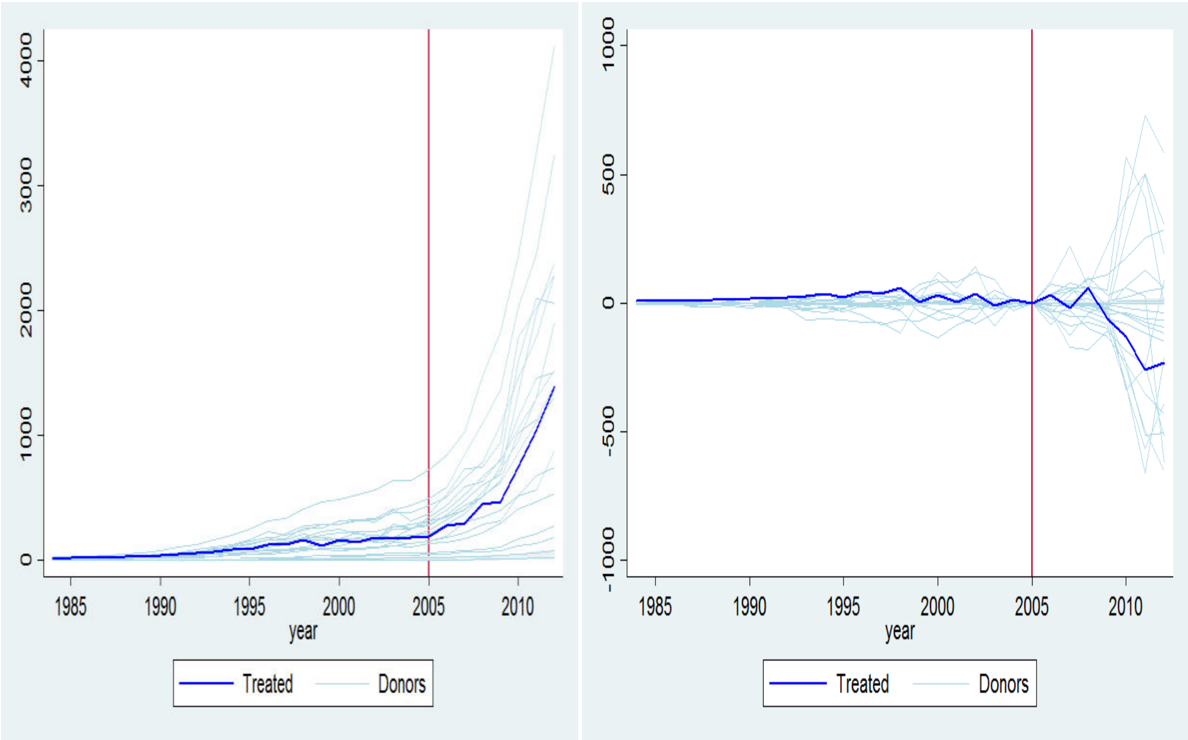


Fig 2: Placebo test: Counterfactual for agricultural GDP and gap between actual and synthetic of Bihar

Acknowledgment: The content was published as book chapter in "Qunatitative methods for social science" edited by Vinayak Nikam, Abimanyu Jhajharia and Suresh Pal.

References

Abadie, A, Diamond, A. and Hainmueller, J. (2015). Comparative politics and the synthetic control method, *American Journal of Political Science*, **59**(2): 495-510.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country, *American Economic Review*, **93**(1):113-132.

Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program, *Journal of the American Statistical Association*, **105**(490):493-505.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. Synthetic Difference-in-Differences. *American Economic Review*, **111** (12): 4088-4118.

Galiani, S. and Quistorff, B. (2016). The synth_runner Package: Utilities to Automate Synthetic Control Estimation Using synth, University of Maryland.

Kreifa, N., Grievea, R., Hangartnerb, D., A. J., Nikolovad, S. and Suttonc, M. (2016). Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units, *Health Economics Health Econ*. 25: 1514–1528.

Draft paper:

Kishore, A., Kishore, P., Roy, D. and Saroj, S. (2021) Impact of sweeping agricultural marketing reforms in a poor state of India Evidence from repeal of the APMC Act. Paper presented in *31st International Conference of Agricultural Economist* (ICAE, 2021) and *16th Conference of Indian Statistical Institute*, New Delhi.

Kishore, P., D.R. Singh, S.K. Srivastava, P. Kumar, and G. Jha (2021). Impact of Subsoil Water Preservation Act, 2009 on burgeoning trend of Groundwater depletion in Punjab, India. Paper presented in *31st International Conference of Agricultural Economist (ICAE,* 2021*)*.

Univariate Time Series Analysis: ARIMA and G(ARCH) Models

Achal Lama & Girish K Jha
ICAR-Indian Agricultural Statistics Research Institute

A time series (TS) is a collection of observations on a quantitative characteristic of a phenomenon observed sequentially in time. Here we are interested on those observations which are collected at equally spaced as well as at discrete time intervals, may be collected hourly, daily, weekly, monthly, or yearly, and so on. For example, daily maximum temperatures, weekly agricultural price data, monthly sales, yearly gross national product, annual crop production, etc. A basic assumption in any TS analysis is that some aspects of the past pattern will continue into the future. Hence, dependency through time is used for extrapolation in the future. The notation such as $\{X_t\}$ or $\{Y_t\}$ (t =1,…,T) is used to denote a time series of length T. The goals of time series models include smoothing irregular series, forecasting series into the medium or long term future and causal modelling of variables moving in parallel through time.

Time series methods

Time series methods use different statistical methods to treat the time series data approximately to draw inferences. These models may be univariate, i.e., modelling of single series of data or multivariate that includes multiple series of data containing different variables. Besides, non-linear time series techniques are also popular nowadays. Here it is tacitly assumed that information about the past is available in the form of numerical data. Ideally, at least 50 observations are necessary for performing TS analysis/ modelling, as propounded by Box and Jenkins who were pioneers in TS modelling.

Decomposition models are among the oldest approaches to TS analysis *albeit* a number of theoretical weaknesses from a statistical point of view. These were followed by the crudest form of forecasting methods called the moving averages method. As an improvement over this method which had equal weights, exponential smoothing methods came into being which gave more weights to recent data. Exponential smoothing methods have been proposed initially as just recursive methods without any distributional assumptions about the error structure in them, and later, they were found to be particular cases of the statistically sound Auto-Regressive Integrated Moving Average (ARIMA) models.

This write-up is intended to provide an overview on both linear and non-linear time series models within the ARMA framework and some frequently used parametric nonlinear models such as Autoregressive conditional heteroscedastic (ARCH) and its generalised form GARCH models. At the end we have given R code for the use of linear and non-linear models on a real data set for better understanding and acceptability.

1. Linear Time Series Models

In a data series containing observations spaced at equal intervals of time often may be correlated. Such correlation between consecutive observations is called *autocorrelation*. When the data is autocorrelated, most of the standard modeling methods based on the assumption of independent observations may become misleading. We therefore need to consider alternative methods that consider the serial dependence in the data which can be achieved by employing time series models such as autoregressive integrated moving average (ARIMA) models.

The most popular class of linear time series models consists of autoregressive moving average (ARMA) models, including purely autoregressive (AR) and purely moving-average (MA) models as special cases. ARMA models are frequently used to model linear dynamic structures, to depict linear relationships among lagged variables, and to serve as vehicles for linear forecasting. A particularly useful class of models contains the so-called autoregressive integrated moving average (ARIMA) models, which includes stationary ARMA - processes as a subclass.

1.1 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, 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 $t, t-1, t-2, \ldots$ by $y_t, y_{t-1}, y_{t-2}, \ldots$ then y_t can be described as

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t$$

If we define an autoregressive operator of order p by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

where B is the backshift operator such that $By_t = y_{t-1}$, autoregressive model can be written as

$\varphi(B)\, y_t = \varepsilon_t$.

1.2 Moving Average (MA) Model

Another kind of model of great practical importance in the representation of observed time-series is finite moving average process. MA (q) model is defined as

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

If we define a moving average operator of order q by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

where B is the backshift operator such that $By_t = y_{t-1}$, moving average model can be written as $y_t = \theta(B)\varepsilon_t$.

1.3 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 mixed autoregressive-moving average model

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots \theta_q \varepsilon_{t-q}$$

or

$$\varphi(B)\, y_t = \theta(B)\varepsilon_t$$

and is written as ARMA(p, q). In practice, it is quite often 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.

1.4 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

$$\varphi(B)(1-B)^d\, y_t = \theta(B)\varepsilon_t$$

ε_t are assumed to be independently and identically distributed with a mean zero and a constant variance of σ^2.

2. Non-linear models: ARCH and GARCH models

After the dominance of the ARIMA model for over two decades, the need of such model was felt which could predict with varying variance of the error term. The solution was provided by Engle (1982) when he developed ARCH model to estimate the mean and variance of the United Kingdom inflation. This model has few interesting characteristics; it models the conditional variance as the square of the function of the previous error term and assumes the unconditional variance to be constant. Along with the ARCH models can model heavy tail data which are common in financial market. Besides these, Bera and Higgins (1993) pointed out that ARCH models are easy and simple to handle, can take care of clustered errors, non-linearity and importantly takes care of changes in the econometrician's ability to forecast.

The ARCH (q) model for the series $\{\varepsilon_t\}$ is defined by specifying the conditional distribution of given the information available up to time $t-1$. Let denote this information. ARCH (q) model for the series is given by

$$\varepsilon_t | \psi_{t-1} \sim N(0, h_t)$$

$$h_t = a_0 + \sum_{i=1}^{q} a_i \, \varepsilon_{t-i}^2$$

where, $a_0 > 0$, $a_i \geq 0$, for all i and $\sum_{i=1}^{q} a_i < 1$ are required to be satisfied to ensure non-negativity and finite unconditional variance of stationary $\{\varepsilon_t\}$ series. Bollerslev (1986) and Taylor (1986) proposed the Generalized ARCH (GARCH) model independently of each other, in which conditional variance is also a linear function of its own lags and has the following form

$$\varepsilon_t = \xi_t h_t^{1/2} \tag{1}$$

where $\xi_t \sim$ N (0,1). A sufficient condition for the conditional variance to be positive is

$$a_0 > 0, \;\; a_i \geq 0, \;\; i = 1,2,...,q. \;\; b_j \geq 0, \;\; j = 1,2,..., p$$

The GARCH (p, q) process is weakly stationary if and only if

$$. \sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1$$

The conditional variance defined by (1) has the property that the unconditional autocorrelation function of ε_t^2 ; if it exists, can decay slowly. For the ARCH family, the decay rate is too rapid compared to what is typically observed in financial time-series, unless the maximum lag q is long. As (1) is a more parsimonious model of the conditional variance than a high-order ARCH model, most users prefer it to the simpler ARCH alternative. The most popular GARCH model in applications is the GARCH (*1,1*) model.

Model Building

Step 1: Determine whether the time series is stationary.

The series being analysed must be stationary. 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 which means that stationary time series has the property that its statistical properties such as the mean and variance are constant over time. The presence of stationarity in the data can be obtained by simply plotting the raw data or by plotting the autocorrelation and partial autocorrelation function. Statistical tests like Dickey- Fuller test, augmented Dickey-Fuller test, KPSS (Kwiatkowski, Phillips, Schmidt, and Shin) test, Philips-Perron test are also available to test the stationarity.

Step 2: Identify the model.

After the time-series is stationary we go for identifying the mean model for the series. This is done by fitting the simple ARIMA (Autoregressive integrated moving average) model. The ARIMA (p,d,q) is determined by the ACF (Autocorrelation function) and PACF (Partial autocorrelation function) values of the stationary series. The parameter p is determined by the ACF value and q by the PACF value and d refers to order of differencing done to the original series to make it stationary.

Step 3: Estimate the model parameters and diagnostic checking.

Once few tentative models are specified, estimation of the model parameters is straightforward. The parameters are estimated through maximum likelihood function such that an overall measure of errors is minimized or the likelihood function is maximized. This step is basically to check if the model assumptions about the errors are satisfied. This is achieved by performing portmanteau test. The test is utilized to see whether the model residuals are white noise. The null hypothesis tested is that the current set of residual is white

noise.

The Ljung-Box statistic is given by:

$$Q = n(n+2)\sum_{k=1}^{h}(n-k)^{-1}r_k^{\,2}$$

where, *h* is the maximum lag, *n* is the number of observations, *k* is the number of parameters in the model. If the data are white noise, the Ljung-Box Q statistics has a chi-square distribution with *(h-k)* degrees of freedom.

Step 4: Select the most suitable ARIMA model

The most suitable ARIMA model is selected using the smallest Akaike Information Criterion (AIC) or Schwarz-Bayesian Criterion (SBC). AIC is given by

$$AIC = (-2\log L + 2m)$$

where, $m = p+q$ and L is the likelihood function. SBC is also used as an alternative to AIC which is given by

$$SBC = \log\sigma^2 + (m\log n)/n$$

If the model is not adequate, a new tentative model should be identified, which is again followed by the parameter estimation and model verification. Diagnostic information may help suggest alternative model(s). The steps of model building process are typically repeated several times until a satisfactory mean model is finally selected. The final model can then be used for prediction purposes.

Step 5: Determination of residuals and heteroscedasticity test.

After finding the mean model now the residuals are to be determined. And we create a new variable called 'rsquare' by squaring the residuals. Then the ACF and PACF values of the 'rsquare' are determined and the lags in which these values are found to be significant are identified. The test for heteroscedasticity is done at identified significant lags. The test employed is the ARCH-LM test.

Step 6: Residuals and diagnostic checking.

The residuals obtained from the mean model used for fitting the different GARCH models were squared and stored in a new variable called 'esquare'. As already mentioned previously, the diagnostic tests are employed to check whether the residuals are white noise or not.

Step 7: Estimation of parameters.

The parameters of the obtained model are estimated using method of maximum likelihood (MLE). And then forecasting is done using the selecting model.

5. Illustration

In this illustration Cotlook A index data is used and was collected from the commodity price bulletin, published by the United Nations Convention of Trade and Development (UNCTAD). The series contains 360 data pints, 346 data points are used for modelling and remaining 14 points for forecasting. At first the ARIMA model was applied to the data set and on unsatisfactory performance of the model, the GARCH model was used.

5.1 Fitting of the Cotlook A index

Various combinations of the ARIMA models were tried, among all, the AR (1) model had minimum AIC and BIC values. The AIC value for fitted GARCH model has been found to be minimum when the mean equation depends on two recent pasts only. Investigating the autocorrelation function (Acf) of squared residuals of AR (2) model, it is found that the Acf and Pacf are maximum at lag 3, which is 0.226 and 0.221 respectively. But if we go for AR (2)-ARCH (3) model, a large number of parameters are needed to be estimated. So, to get a parsimonious model, the AR (2)-GARCH (1, 1) model is selected.

The mean and conditional variance for fitted AR (2)-GARCH (1, 1) model is computed as follows:

$$y_t = 141.9264 \; -1.3905 \; y_{t-1} + 0.4538 \, y_{t-2} + \varepsilon_t$$

$$(3.94) \quad (0.05) \qquad (0.05)$$

where

$$\varepsilon_t = h_t^{1/2} \xi_t \,,$$

and h_t satisfies the variance equation

$$h_t = 8.470 + 0.208 \; \varepsilon_{t-1}^2 + 0.215 \; h_{t-1}$$

$$(1.97) \qquad (0.09) \qquad (0.079)$$

The values within brackets denote corresponding standard errors of the estimates. The AIC value, for fitted GARCH model is 2288.88.

Table 1. Forecast of the Cotlook A index series

MONTH	ACTUAL VALUE	FORECAST ARIMA(1,1,0)	FORECAST AR(2)-GARCH(1,1)
Feb-11	469.98	408.34(8.30)	389.59(26.46)
Mar-11	506.34	416.47(15.56)	371.55(25.74)
Apr-11	477.56	421.40(22.35)	348.54(25.05)
May-11	364.91	424.53(28.55)	324.69(24.39)
Jun-11	317.75	426.66(34.17)	301.98(23.75)
Jul-11	268.96	428.23(39.29)	281.25(23.13)
Aug-11	251.55	429.49(43.97)	262.76(22.54)
Sep-11	257.63	430.57(48.29)	246.50(21.97)
Oct-11	243.85	431.55(52.30)	232.32(21.42)
Nov-11	230.78	432.48(56.05)	220.01(20.90)
Dec-11	210.43	433.37(59.58)	209.35(20.39)
Jan-12	222.91	434.25(54.45)	200.15(19.91)
Feb-12	222.12	435.12(57.13)	192.21(19.44)
Mar-12	219.36	435.99(59.68)	185.37(19.01)

Table 2. Forecast evaluation of the Cotlook A index series

MODEL	RMSE	RMAPE (%)
ARIMA(1,1,0)	44.03	60.72
AR(2)-GARCH(1,1)	15.38	9.36

6. R code for analysing a time series data

```
library("tseries")
library("forecast")
library("fgarch")
setwd("C:/Users/ACHAL/Desktop") # Setting of the work directory
data<-read.table("data.txt") # Importing data
```

```
datats<-ts(data,frequency=12,start=c(1982,4)) # Converting data set into time series
plot.ts(datats) # Plot of the data set
adf.test(datats) # Test for stationarity
diffdatats<-diff(datats,differences=1) # Differencing the series
datatsacf<-acf(datats,lag.max=12) # Obtaining the ACF plot
datapacf<-pacf(datats,lag.max=12) # Obtaining the PACF plot
auto.arima(diffdatats) # Finding the order of ARIMA model
datatsarima<-arima(diffdatats,order=c(1,0,1),include.mean=TRUE) # Fitting of ARIMA model
forearimadatats<-forecast.Arima(datatsarima,h=12) # Forecasting using ARIMA model
plot.forecast(forearimadatats) # Plot of the forecast
residualarima<-resid(datatsarima) # Obtaining residuals
archTest(residualarima,lag=12) # Test for heteroscedascity
# Fitting of AR-GARCH model
garchdatats<-garchFit(formula = ~ arma(2)+garch(1, 1), data = datats, cond.dist = c("norm"), include.mean = TRUE, include.delta = NULL, include.skew = NULL, include.shape = NULL, leverage = NULL, trace = TRUE,algorithm = c("nlminb"))
# Forecasting using AR-GARCH model
forecastgarch<-predict(garchdatats, n.ahead = 12, trace = FALSE, mse = c("uncond"), plot=FALSE, nx=NULL, crit_val=NULL, conf=NULL)
plot.ts(forecastgarch) # Plot of the forecast
```

References:

Bera, A. K., and Higgins, M. L. (1993), ARCH Models: Properties, Estimation and Testing, *Journal of Economic Surv*ey, **7**, 307-366.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, **31**, 307-327.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2007). Time-Series Analysis:Forecasting and Control. 3rd edition. *Pearson education*, India.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987-1008.

Fan, J. and Yao, Q. (2003). *Nonlinear time series:nonparametric and parametric methods*. Springer, U.S.A.

Taylor, S. J. (1986). Modeling financial time series. Wiley, New York.

Structural Break Analysis and its application

P. Anbukkani & Haritha K.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction

In econometrics and statistics, a structural break is an unexpected change over time in the parameters of regression models, which can lead to huge forecasting errors and unreliability of the model in general. This issue was popularised by David Hendry, who argued that lack of stability of coefficients frequently caused forecast failure, and therefore we must routinely test for structural stability. Structural stability − i.e., the time-invariance of regression coefficients − is a central issue in all applications of linear regression models.

Structural change is a statement about parameters, which only have meaning in the context of a model. To focus our discussion, we will discuss structural change in the simplest dynamic model, the first-order autoregression:

$$Yt = \alpha + \rho Y_{t-1} + e_t$$

where e_t is a time series of serially uncorrelated shocks. The parameters are (α, ρ, σ^2).

The assumption of stationarity implies that these parameters are constant over time. We say that a structural break has occurred if at least one of these parameters has changed at some date—the break date—in the sample period. While it may seem unlikely that a structural break could be immediate and might seem more reasonable to allow a structural change to take a period of time to take effect, we most often focus on the simple case of an immediate structural break for simplicity and parsimony. A structural break may affect any or all of the model parameters, and these cases have different implications. Changes in the autoregressive parameter ρ reflect changes in the serial correlation in Yt . The intercept α controls the mean of Yt through the relationship E($Yt = \mu = \alpha/(1-\rho)$). Since Yt is the growth rate in labor productivity, changes in m are identical to changes in the trend and are probably the issue of primary interest. Finally, changes in σ^2 imply changes in the volatility of labor productivity. The econometrics of structural change looks for systematic methods to identify structural breaks. In the past 15 years, the most important contributions to this literature include the following three innovations:

1) Tests for a structural break of unknown timing
2) Estimation of the timing of a structural break
3) Tests to distinguish between a random walk and broken time trends.

These three innovations have dramatically altered the face of applied time series econometric (Hansen, 2001)

Testing for Structural Change of Unknown Timing

Testing for structural change has always been an important issue in econometrics because a myriad of political and economic factors can cause the relationships among economic variables to change over time. The early works of Quandt (1958) and Chow (1960) consider tests for structural change for a known single break date. The researches headed for the modelling where this break date is treated as an unknown variable. Quandt (1960) extends the Chow test and proposes taking the largest Chow statistic over all possible break dates. In the same context, the most important contributions are those of Andrews (1993) and Andrews and Ploberger (1994) who consider a comprehensive analysis of the problem of testing for structural change.

The classical test for structural change is typically attributed to Chow (1960). His famous testing procedure splits the sample into two subperiods, estimates the parameters for each subperiod, and then tests the equality of the two sets of parameters using a classic F statistic. This test was popular for many years and was extended to cover most econometric models of interest. For a recent treatment, see Andrews and Fair (1988). However, an important limitation of the Chow test is that the break date must be known a priori. A researcher has only two choices: to pick an arbitrary candidate break date or to pick a break date based on some known feature of the data. In the first case, the Chow test may be uninformative, as the true break date can be missed. In the second case, the Chow test can be misleading, as the candidate break date is endogenous—it is correlated with the data—and the test is likely to indicate a break falsely when none in fact exists. Furthermore, since the results can be highly sensitive to these arbitrary choices, different researchers can easily reach quite distinct conclusions—hardly an example of sound scientific practice.

Chow tests

Chow tests assess the stability of coefficients β in a multiple linear regression model of the form $y = X\beta + \varepsilon$. chowtest splits the data at specified break points. Coefficients are estimated in initial subsamples, then tested for compatibility with data in complementary subsamples.

We can use the following steps to perform a Chow test.

Step 1: Define the null and alternative hypotheses.

Suppose we fit the following regression model to our entire dataset:

- $y_t = a + bx_{1t} + cx_{t2} + \varepsilon$

Then suppose we split our data into two groups based on some structural break point and fit the following regression models to each group:

- $y_t = a_1 + b_1x_{1t} + c_1x_{t2} + \varepsilon$
- $y_t = a_2 + b_2x_{1t} + c_2x_{t2} + \varepsilon$

We would use the following null and alternative hypotheses for the Chow test:

- Null (H_0): $a_1 = a_2$, $b_1 = b_2$, and $c_1 = c_2$

Alternative (H_A): At least one of the comparisons in the Null is not equal.

If we reject the null hypothesis, we have sufficient evidence to say that there is a structural break point in the data and two regression lines can fit the data better than one.

If we fail to reject the null hypothesis, we do not have sufficient evidence to say that there is a structural break point in the data. In this case, we say that the regression lines can be "pooled" into a single regression line that represents the pattern in the data sufficiently well.

Step 2: Calculate the test statistic.

If we define the following terms:

- S_T: The sum of squared residuals from the total data
- S_1, S_2:The sum of squared residuals from each group
- N_1, N_2: The number of observations in each group
- k: The number of parameters

Then we can say that the Chow test statistic is:

Chow test statistic = $[(S_T - (S_1+S_2))/k]$ / $[(S_1+S_2)/ (N_1+N_2-2k)]$

This test statistic follows the F-distribution with *k* and and N_1+N_2-2k degrees of freedom.

Step 3: Reject or fail to reject the null hypothesis.

If the p-value associated with this test statistic is less than a certain significance level, we can reject the null hypothesis and conclude that there is a structural break point in the data.

Fortunately, most statistical software is capable of performing a Chow test so you will likely never have to perform the test by hand.

Bai-perron Test

Bai-Perron (BP) procedure are briefly discussed below. The BP) methodology considers the following multiple structural break model with m breaks (m+1 regimes).

$$
\begin{aligned}
y_t &= x_t'\beta + z_t'\delta_1 + u_t, t = 1,\ldots\ldots,T_1 \\
y_t &= x_t'\beta + z_t'\delta_2 + u_t, t = T_1+1,\ldots\ldots,T_2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_t &= x_t'\beta + z_t'\delta_{m+1} + u_t, t = T_m+1,\ldots\ldots,T
\end{aligned}
\quad (1)
$$

where yt is the observed dependent variable at time t; x_t is px1 and z_t is qx1, and β^1 and δ_j (j=1,…..,m+1) are the corresponding vectors of coefficients; and ut is the disturbance term at time t. the break points (T_1, ….,T_m) are treated as unknown, and are estimated together with the unknown coefficients when T observations are available. The purpose is to estimate the unknown regression coefficients and the break dates (β, δ_1,…. δ_{m+1}, T_1,…T_m)when T observations on (y_t, x_t, z_t) are available. The above multiple linear regression models can be expressed in matrix form as

$$Y = X\beta + \bar{Z}\delta + U \quad (2)$$

where Y=(y_1,…..y_T)' , X=(x_1,….x_T)', Z is the matrix which diagonally partitions Z at the m-partition (T_1,…T_m), i.e. $\bar{Z} = diag(Z_1,\ldots Z_{m+1})$ with Zi=(z_{T-i}+1,….z_{Ti})' , $\delta = (\delta_1', \delta_2',\ldots.\delta_{m+1}')'$ and U=(u_1,……..u_T)'. Bai and Perron (1998) impose some restrictions on the possible values of the break dates. They define the following set for some arbitrary small positive number $\varepsilon : \Lambda_\varepsilon = \{(\lambda_1,\ldots\lambda_m); |\lambda_{i+1} - \lambda_i| \geq \varepsilon, \lambda_1 \geq \varepsilon, \lambda_m \leq 1-\varepsilon\}$ to restrict each break date to be asymptotically distinct and bounded from the boundaries of the sample where the λ_i (i=1,2…..m) gives the break fraction ($\lambda_i = T_i / T_m$).

R code for Bai-Perron

```
library(strucchange)
maize_area_kr <- breakpoints (maize_area_kr~year, h=8, data=tsdata)
maize_area_kr
summary(maize_area_kr)
plot(maize_area_kr)
maize_area_mp <- breakpoints (maize_area_mp~year, h=8, data=tsdata)
maize_area_mp
summary(maize_area_mp)
plot(maize_area_mp)
maize_area_tn <- breakpoints (maize_area_tn~year, h=8, data=tsdata)
maize_area_tn
summary(maize_area_tn)
plot(maize_area_tn)
```

References

Andrews, Donald W.K. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." Econometrica. July, 61:4, pp. 821–56.

Andrews, Donald W.K. and Ray C. Fair. 1988. "Inference in Nonlinear Econometric Models with Structural Change." Review of Economic Studies. October, 55:4, pp. 615–3

Bai, Jushan and Pierre Perron .2004. "Multiple Structural Change Models: A Simulation Analysis". In: Corbea, D., Durlauf, S., Hansen, B.E. (Eds.), Econometric Essays. Cambridge University Press. In press

Bai, Jushan and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica.* January, 66:1, pp. 47–78.

Bai, Jushan, Robin L. Lumsdaine and James H. Stock. 1998. "Testing for and Dating Common

Bai, Jushan. 1994. "Least Squares Estimation of a Shift in Linear Processes." *Journal of Time Series Analysis.* 15:5, pp. 453–72.

Bai, Jushan. 1997a. "Estimation of a Change Point in Multiple Regression Models." *Review of Economics and Statistics.* November, 79:4, pp. 551–63.

Bai, Jushan. 1997b. “Estimating Multiple Breaks One at a Time.” *Econometric Theory.* June, 13:3, pp. 315–52.

Breaks in Multivariate Time Series.” *Review of Economic Studies.* July, 64:3, pp. 395–432.

Chow, Gregory C. 1960. “Tests of Equality Between Sets of Coefficients in Two Linear Regressions.” Econometrica. 28:3, pp. 591–605

Gujarati, D.N. 2004. *Basic Econometrics*. 4th Edition, McGraw-Hill Companies

Hansen, Bruce E. 2000. “Testing for Structural Change in Conditional Models.” *Journal of Econometrics.*July, 97:1, pp. 93–115.

Hansen.,B.E. 2001. “The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity.” Journal of Economic Perspectives,15:4, 117–128

Perron, Pierre (Ed.) .2018. “Unit roots and structural breaks”, ISBN

Perron, Pierre. 1989. “The Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis.” *Econometrica.* November, 57:6, pp. 1361–401.

Perron, Pierre. 1997. “Further Evidence on Breaking Trend Functions in Macroeconomic Variables.” *Journal of Econometrics.* October, 80:2, pp. 355–85.

Quandt, Richard. 1960. “Tests of the Hypothesis that a Linear Regression Obeys Two Separate Regimes.” Journal of the American Statistical Association. 55, pp. 324–30.

Application of Structural Equation Modeling (SEM) in Social Science Research

Sitaram Bishnoi[1], Satyapriya[1] and Praveen K.V[2]

[1]Division of Agricultural Extension, ICAR-Indian Agricultural Research Institute

[2]Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Abstract

Structural Equation Modeling (SEM) is a powerful statistical technique widely used in social science research to analyze complex relationships among multiple variables. This chapter provides an in-depth exploration of the theoretical foundations, key concepts, and practical application of SEM in the context of social science research. We begin by introducing the basic principles of SEM and then delve into various aspects, including model specification, estimation methods; model fit assessment, and interpreting results. Throughout the chapter, we illustrate the application of SEM using real-world examples to help researchers better understand and utilize this valuable tool in their social science inquiries.

1. Introduction

First-generation multivariate data analysis techniques, such as multiple regression, logistic regression, and analysis of variance, belong to the core set of statistical methods employed by researchers to empirically test hypothesized relationships between variables of interest. Numerous researchers in various scientific disciplines have applied these methods to generate findings that have significantly shaped the way we see the world today. These techniques have three important limitations in common, namely (1) the postulation of a simple model structure, (2) requiring that all variables can be considered observable, and (3) the assumption that all variables are measured without error.

With regard to the first limitation, multiple regression analysis and its extensions postulate a simple model structure involving one layer of dependent and independent variables. It can only be estimated piecewise with these methods rather than simultaneously, which can have severe consequences for the quality of the results. With regard to the second limitation, regression-type methods are restricted to processing observable variables, such as age or sales (in units or dollars). Theoretical concepts, which are "abstract, unobservable properties or attributes of a social unit of entity", can only be considered after prior stand-alone validation by means of, for example, a confirmatory factor analysis (CFA). With regard to the third limitation and related to the previous point, one has to bear in mind that each observation of the real world is accompanied by a certain degree of measurement error, which can be systematic or random. First-generation techniques are, strictly speaking, only applicable when measured variables contain neither systematic nor random error. This situation is, however, rarely encountered in reality, particularly when the aim is to estimate relationships among measures of theoretical concepts. Since the social sciences, and many other Fields of scientific inquiry, routinely deal with theoretical concepts, such as perceptions, attitudes, and intentions, these limitations of first-generation techniques are fundamental. To overcome these limitations, researchers have increasingly been turning to second-generation techniques. These methods, referred to as structural equation modeling (SEM), enable researchers to simultaneously model and estimate complex relationships among multiple dependent and independent variables. The concepts under consideration are typically unobservable and measured indirectly by multiple indicators. In estimating the relationships, SEM accounts for measurement error in observed variables. As a result, the method obtains a more precise measurement of the theoretical concepts of interest.

Social scientists often face complex research questions that involve numerous variables with intricate interrelationships. Structural Equation Modeling (SEM) provides an effective solution to address these challenges. This chapter aims to equip researchers with the necessary knowledge and skills to employ SEM confidently in their social science studies. We will cover the following aspects of SEM:

2. Theoretical Foundations

SEM is a multivariate technique based on variants in both the measurement and structural models. In the measurement models, each set of indicators for a construct act collectively (as a variate) to define the construct. In the structural model, constructs are related to one another is correlational and dependence relationships. SEM is a family of statistical models that seek to explain the relationships among multiple variables. It examines the structure of interrelationships expressed in a series of equations, similar to a series of multiple regression equations. The equation depicts all of the relationship among constructs (the dependent and independent variables) involved in the analysis. Constructs are unobservable of latent factors represented by multiple variables (much like variables representing a factor in factor analysis). Thus each multivariate technique has been classified either as an interdependence or dependence technique. SEM can be thought of as a unique combination of both types of techniques because SEM foundation lies in two familiar multivariate techniques: factor analysis and multiple regression analysis.

SEM should never be attempted without a strong theoretical basis for specification of both the measurement and structural models. The theoretical support in SEM plays following fundamental roles: (1) specifying relationships that define the model (2) establishing causation, particularly when using cross-sectional data (3) the development of modeling strategy. Two popular methods dominate SEM in practice: covariance-based SEM (CBSEM) and partial least squares SEM (PLS-SEM, also called PLS path modeling). CB-SEM is primarily used to confirm (or reject) theories and their underlying hypotheses. This approach confirms/rejects hypothesis by determining how closely a proposed theoretical model can reproduce the covariance matrix for an observed sample dataset. In contrast, PLS has been introduced as a “causal–predictive” approach to, which focuses on explaining the variance in the model’s dependent variables.

2.1. Understanding Causal Relationships: The basis of SEM lies in the explicit modeling of causal relationships among observed and latent variables. We explore the fundamental principles of causality and how SEM facilitates the identification of direct and indirect effects.

2.2. The Path Diagram: The graphical representation of a SEM model is essential for understanding its structure. We discuss how to construct and interpret path diagrams to visualize complex models.

A model is a representation of theory. Theory can be thought of as a systematic set of relationship providing a consistent and comprehensive explanation of phenomena. A model should not be developed without some underlying theory. Theory is often a primary objective of academic research, but practitioners may develop or propose a set of relationship that are as complex and interrelated as any academically based theory. Path models are diagrams used to visually display the hypotheses and variable relationships that are examined when SEM is applied. Path models are developed based on theory and are often used to test theoretical relationships. Theory is a set of systematically related hypotheses developed following the scientific method that can be used to explain and predict outcomes. Thus, hypotheses are individual conjectures, whereas theories are multiple hypotheses that are logically linked together and can be tested empirically. Two types of theory are required to develop path models: measurement theory and structural theory. Measurement theory specifies which indicators and how these are used to measure a certain theoretical concept. In contrast, structural theory specifies how the constructs are related to one another in the structural model.

3. Model Specification

3.1. Latent Variables: In social science research, not all constructs of interest can be directly observed. We explain the concept of latent variables and the importance of specifying them in SEM. Many critical factors in agricultural extension, such as farmers' perceptions, attitudes, and motivations, are latent and cannot be directly measured. We discuss the identification and specification of latent variables based on theoretical concepts.

3.2. Measurement Model: The measurement model establishes the relationships between latent variables and their observable indicators. We guide researchers through the process of assessing and validating measurement models.

Latent Variable	Indicator	Factor Loading	Standard Error	t-value	p-value
Attitudes	Attitude1	0.85	0.03	29.67	<0.001
	Attitude2	0.78	0.04	20.53	<0.001
Subjective Norms	Norm1	0.72	0.05	14.85	<0.001
	Norm2	0.66	0.06	11.48	<0.001
Perceived Behavioral Control	Control1	0.79	0.04	18.59	<0.001
	Control2	0.71	0.05	14.70	<0.001

3.3. Structural Model: The structural model defines the relationships between latent variables. We demonstrate how to formulate hypotheses and specify the structural model.

Latent Variable	Path Coefficient	Standard Error	t-value	p-value
Attitudes -> Intention	0.57	0.05	11.38	<0.001
Subjective Norms -> Intention	0.42	0.04	10.10	<0.001
Perceived Behavioral Control -> Intention	0.34	0.03	9.92	<0.001

3.4. **Exogenous versus Endogenous latent variables:**

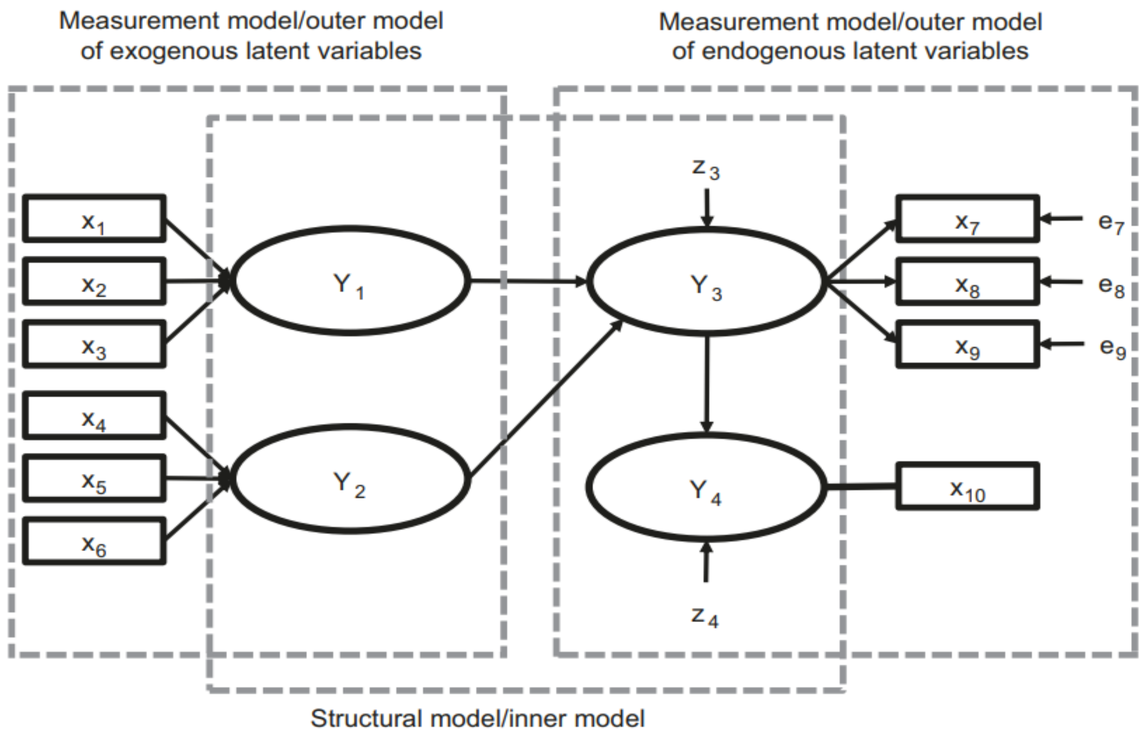
Exogenous latent variables are synonymous with independent variables; they “cause” fluctuations in the values of other latent variables in the model. Changes in the values of exogenous variables are not explained by the model. Rather, they are considered to be influenced by other factors external to the model. Background variables such as gender, age, and socioeconomic status are examples of such external factors. Endogenous latent variables are synonymous with dependent variables and, as such, are influenced by the exogenous variables in the model, either directly or indirectly. Fluctuation in the values of endogenous variables is said to be explained by the model because all latent variables that influence them are included in the model specification.

SEM Symbol Notation:

Structural equation models are schematically portrayed using particular configurations of four geometric symbols—a circle (or ellipse), a square (or rectangle), a single-headed arrow, and a double-headed arrow. By convention, circles (or ellipses;) represent unobserved latent factors, squares (or rectangles;) represent observed variables, single-headed arrows (→) represent the impact of one variable on another, and double-headed arrows (↔) represent covariances or correlations between pairs of variables. In building a model of a particular structure under study, researchers use these symbols within the framework of four basic configurations, each of which represents an important component in the analytic process. These configurations, each accompanied by a brief description, are as follows:

Basic composition

The general SEM model can be decomposed into two sub models: a measurement model, and a structural model. The measurement model defines relations between the observed and unobserved variables. In other words, it provides the link between scores on a measuring instrument (i.e., the observed indicator variables) and the underlying constructs they are designed to measure (i.e., the unobserved latent variables). The measurement model, then, represents the CFA model described earlier in that it specifies the pattern by which each measure loads on a particular factor. In contrast, the structural model defines relations among the unobserved variables.



Measurement Theory:

Measurement theory specifies how the latent variables (constructs) are measured. Generally, there are two different ways to measure unobservable variables. One approach is referred to as reflective measurement, and the other is formative measurement. In formative directional arrows are pointing from the indicator variables to the construct, indicating a predictive (causal) relationship in that direction. In reflective indicators, the direction of the arrows is from the construct to the indicator variables, indicating the assumption that the construct "causes" the measurement (more precisely, the covariation) of the indicator variables.

Structural Theory:

Structural theory shows how the latent variables are related to one another (i.e., it shows the constructs and their path relationships in the structural model). The location and sequence of the constructs are either based on theory or the researcher's experience and accumulated knowledge or both. When path models are developed, the sequence is from left to right. The variables on the left side of the path model are independent variables, and any variable on the right side is a dependent variable. Moreover, variables on the left are shown as sequentially preceding and predicting the variables on the right. However, when variables are in the middle of the path model (between the variables that serve only as independent or dependent variables), they serve as both independent and dependent variables in the structural model.

PLS-SEM and CB-SEM

There are two main approaches to estimating the relationships in a structural equation model. One is CB-SEM, and the other is PLS-SEM. Each is appropriate for a different research context, and researchers need to understand the differences in order to apply the correct method. A crucial conceptual difference between PLS-SEM and CB-SEM relates to the way each method treats the latent variables included in the model. CB-SEM represents a common factor-based SEM method that considers the constructs as common factors that explain the covariance between its associated indicators. This approach is consistent with the measurement philosophy underlying reflective measurement, in which the indicators and their covariance are regarded as manifestations of the underlying construct. PLS-SEM, on the other hand, assumes the concepts of interest can be measured as composites), which is why PLS is considered a composite-based SEM method. Model estimation in PLSSEM involves linearly combining the indicators of a measurement model to form composite variables. The composite-based approach is consistent with the measurement philosophy underlying formative measurement, but this does not imply that PLS-SEM is only capable of estimating formatively specified constructs. The reason is that the estimation perspective (i.e., forming composites to represent conceptual variables) should not be confused with the measurement theory perspective (i.e., specifying measurement models as reflective or formative). Researchers can include reflectively and formatively specified measurement models that PLS-SEM can straightforwardly estimate.

SEM has become a popular approach in a relatively short period of time. Researchers are attracted to SEM because it provides a conceptually appealing way to test theory. There are six stages in SEM:

Stage 1: Defining individual construct

Stage 2: Developing overall measurement model

Stage 3: Designing a study to produce empirical results

Stage 4: Assessing the measurement model validity

Stage 5: Specifying the structural model

Stage 6: Assessing structural model validity

4. Estimation Methods

4.1. Maximum Likelihood Estimation: We introduce the widely used Maximum Likelihood Estimation (MLE) method for estimating model parameters and discuss its assumptions.

4.2. Bayesian Estimation: An overview of Bayesian estimation as an alternative approach to MLE, highlighting its advantages and when it is suitable for social science research.

5. Model Fit Assessment

5.1. Goodness-of-Fit Indices: We discuss various fit indices used to assess the adequacy of a SEM model, such as chi-square test, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA).

Fit Index	Value	Recommended Threshold
Chi-square	145.23	--
df (Degrees of Freedom)	10	--
CFI (Comparative Fit Index)	0.95	>= 0.90
TLI (Tucker-Lewis Index)	0.94	>= 0.90
RMSEA (Root Mean Square Error of Approximation)	0.06	<= 0.08
SRMR (Standardized Root Mean Square Residual)	0.04	<= 0.08

5.2. Interpretation of Fit Indices: Researchers will learn how to interpret the fit indices and what values indicate a good model fit.

6. Model Modification

6.1. Model Re-specification: When a model does not fit the data adequately, modification is required. We explain common model modifications, such as adding or removing paths, allowing correlated errors, and exploring alternative models.

7. Reporting and Interpreting Results

7.1. Presenting SEM Results: Guidance on how to present SEM results in a clear and concise manner, including path coefficients, standardized coefficients, and significance levels. Researchers will gain valuable guidance on effectively presenting SEM results to make them accessible to diverse audiences, including policymakers, extension agents, and farmers.

7.2. Interpreting Findings: Understanding the implications of SEM results for social science research and drawing meaningful conclusions. The chapter emphasizes the significance of interpreting SEM results to gain insights into farmer behavior, adoption patterns, and the effectiveness of agricultural programs.

8. Advanced Topics

8.1. Multiple Group Analysis: The application of SEM in comparing groups and investigating group differences. The application of SEM to compare different groups of farmers, such as smallholders and large-scale farmers, and investigating group differences in adoption behavior.

8.2. Mediation and Moderation: How to explore mediation and moderation effects using SEM to gain a deeper understanding of complex relationships.

9. Conclusion

In conclusion, Structural Equation Modeling is a valuable tool for social science researchers to unravel complex relationships among variables. By providing a comprehensive overview of SEM, this chapter aims to empower researchers to apply this methodology effectively in their investigations, leading to more robust and insightful findings in the field of social sciences. Researchers are encouraged to keep refining their understanding of SEM and explore its vast potential for advancing knowledge and solving real-world challenges. By providing an in-depth understanding of SEM principles and its application in the agricultural domain, this chapter equips researchers with the knowledge and skills to harness SEM effectively, leading to more informed decision-making and the development of targeted and impactful agricultural extension interventions.

References

Hair Jr, J. F., William C.B., Babin B.J., Anderson R.E., (2014). Multivariate data analysis. Seventh Edition. Pearson New International Edition. Pp 546-99.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). **A primer on partial least squares structural equation modeling (PLS-SEM**). Sage publications.

Austin, J.T. & Calderón, R.F. (1996). Theoretical and technical contributions to structural equation modeling: An updated bibliography. Structural Equation Modeling, 3, 105–175.

Byrne, B.M. (2001). Structural equation modeling with AMOS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum.

Byrne, B.M. (2003). Confirmatory factor analysis. In R. Fernández-Ballesteros (Ed.), Encyclopedia of psychological assessment (Vol. 1, pp. 399–402). Thousand Oaks, CA: Sage.

Kline, R. B. (1998). Principles and Practice of Structural Equation Modeling. New York: The Guilford Press. Kline, R.B. (2011). Principles and practice of structural equation modeling (3rd ed.). New York, NY: Guildford.

Approaches for Estimating Ecosystem Services

Kiran Kumara T. M[1], & Suvangi Rath[2]

[1]ICAR-National Institute for Agricultural Economics and Policy Research

[2] Department of Agricultural Economics, College of Agriculture, Odisha University of Agriculture and Technology

Introduction

Human beings derive a plethora of benefits from nature, both directly and indirectly. The multitude of benefits that the environment provides to society are collectively known as 'Ecosystem Services(ES)' (Costanza et. al., 2017; ESB, FAO). ES is broadly defined as the distinct benefits derived by humans from their environment, that support and fulfill on earth (MEA, 2005). Ecosystem services play a fundamental role in sustaining human life, encompassing vital functions such as provisioning food, securing water, regulating climatic conditions, fostering pollination, shaping the soil, and even offering recreational, spiritual, and cultural enrichment. Despite an estimated value of $125 trillion, these invaluable assets remain inadequately integrated into political and economic policies, resulting in insufficient investments towards their preservation and effective management (ESB, FAO). These services can be categorized into four primary types (refer to Fig. 1).

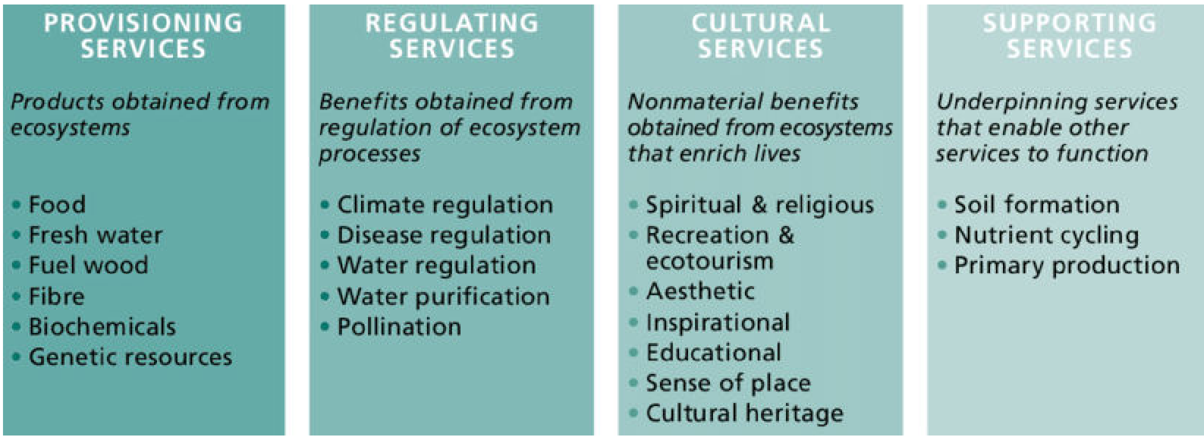


Fig. 1 Different Types of Ecosystem Services (*Source: MEA, 2005*)

Ecosystem Services: Why are they essential?

The environment around us offers an extensive array of indispensable services to humanity. These include the provision of clean water, essential raw materials like wood and lac for economic enterprises, the vital role of climate regulation, soil genesis, pollination of crops, and even intangible cultural advantages such as avenues for recreation, aesthetic enrichment, and spiritual connections. These numerous benefits that people derive from ecosystems are known as 'ecosystem services' (Costanza et. al., 2017). Ecosystem services significantly contribute to our health, economy, and, above all, our overall quality of life, yet paradoxically, many of these services have historically been overlooked. The concept of ecosystem services is an integral component of the broader "Ecosystem Approach," which strives to encompass a comprehensive geographical and socio-economic perspective. As outlined by the Conservation of Biodiversity, this approach represents a strategic framework for the holistic management of land, water, and living resources.

Importance of assessing agricultural ecosystem Services

Agricultural ecosystems inherently offer and depend upon pivotal ecosystem services (ES); however, these services often remain undervalued and uncompensated due to missing markets and market failure. In addition to their essential role in meeting the growing demand for food and other agricultural products, agricultural ecosystems also play an important role in sequestering carbon, managing watersheds, and providing landscapes. However, the intensification of farming practices and the continuation of unsustainable agricultural activities significantly contribute to the deterioration of several critical ES. Conversely, the preservation and safeguarding of biodiversity and ecosystem services ensures long-term food security. Healthy ecosystems are the best way to ensure productive agriculture and nutritious food. Thus, knowing their intrinsic and economic value can help us promote greater investment for the management of agro-ecosystems.

Compared to other ecosystems, the agricultural ecosystem is one that is mostly driven by humans. Agroecosystems are complex systems of climate zones that include factors like temperature, rainfall, and other parameters that influence the growth of crops through direct and indirect interactions with the flora and fauna in the soil including the growth–promoting microbiota (Yadav et. al., 2021). This complex interplay underscores their dual role in offering and being reliant upon ES (see Figure 2) (Garbach et al., 2014).

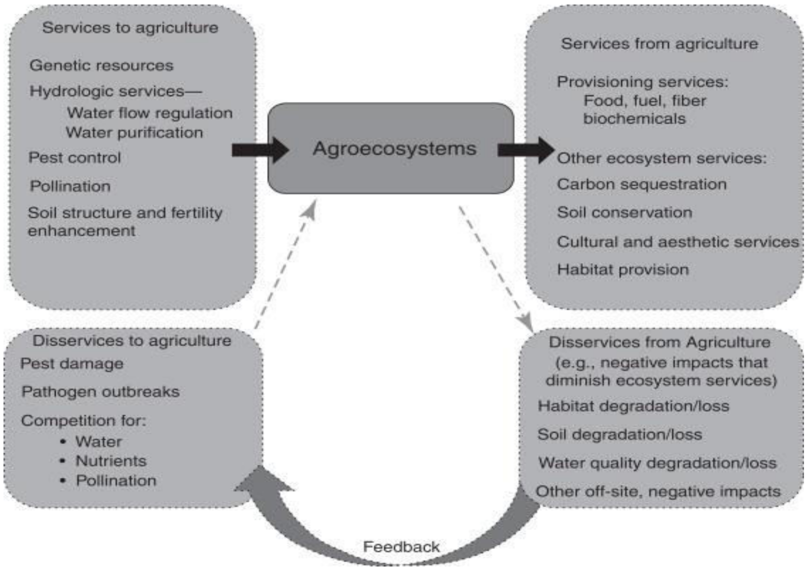


Fig. 2: Ecosystem services and disservices to and from agriculture

(Source: *Garbach et. al., 2014*)

Need for valuation of ecosystem services

Historically, the conventional economic analysis did not take into consideration the services provided by nature. Although the tangible natural resources used in the production process were valued at market prices, a significant array of intangible yet pivotal services have been disregarded. For example, while we factor in natural resources like minerals when estimating a firm's production costs, we tend to overlook the sink services provided by the atmosphere. We release increasing amounts of pollutants into the ecosystem. This has resulted in massive destruction of nature and ecosystems. In consequence, we are increasingly spending more and more resources to avail the services that we derived freely from the natural ecosystems. For example, in the absence of natural pollinators, we have to use hormones for increasing the agricultural output, in the absence of natural nitrogen fixation, we have to use factory-produced nitrogen. Consequently, due to the degradation of ecosystems and the consistent decline in ecosystem services, our realization of ecosystem services is increasing.

We must understand that humanity cannot flourish without the benefits and services provided by their natural environment. As the environment around us is a complex system interconnected with other systems of day-to-day life, a healthy, properly functioning natural environment is the foundation of sustained economic growth, prospering communities, and personal well-being. Hence, it becomes imperative to accurately assess the economic and social merits of a sound natural ecosystem while concurrently upholding the intrinsic value of nature. An ecosystems-based approach to valuation offers a framework that holistically evaluates entire ecosystems within decision-making processes, enabling the appraisal of the ecosystem services they furnish. This, in turn, ensure the preservation of a vibrant and resilient natural environment, both in the present and for future generations. (DEFRA, 2007). Therefore, realizing these multifaceted goals necessitates a comprehensive strategy that transcends isolated environmental considerations, encompassing actions across diverse sectors.

Why evaluate ecosystem services in monetary terms?

With the ongoing exploitation of the environment by the global population, assessment of ecosystem services becomes essential to add value to our decision-making processes. This recognition is further amplified by the stark reality that conventional economic yardsticks, including GDP and other economic metrics, fail to account for the invaluable contributions of nature. In this context, the concept of green accounting emerges as a promising paradigm to inform economic choices. Consequently, the valuation of ecological services assumes a foundational role, underscoring the very essence of our efforts. Being aware of the potential implications covering the range of ecosystem services contributes to averting unintended negative consequences and potentially optimizing the net advantages obtained thereof. However, the extent to which such assessments are warranted, along with the intricacies that follow, will inevitably fluctuate based on the specific context, encompassing factors like geographic area and study scale.

The economic valuation, functioning as a robust risk assessment instrument, extends its utility across every facet of society, fostering the broader integration of the Ecosystem Approach. It can thus be applied to communicate and determine the consequences of the decisions, policies, and schemes of the state and most importantly to consider options for the future use or management of habitats, to broaden the scope of impact assessments, to address the robustness of business plans, and to communicate with and better engage local communities. In regions where expansive biodiversity remains inadequately managed, marginalized, and subject to exploitation, particularly within developing nations, such valuation emerges as an essential imperative. This serves as a rallying point for responsible stewardship and underscores the dire need to ensure the preservation and responsible utilization of these invaluable natural resources.

Approaches to Valuation of Ecosystem Services

Valuation of ecosystem services involves assigning economic or non-economic values to the benefits that ecosystems provide to humans. There are several approaches to valuing ecosystem services, each with its own strengths, limitations, and suitability for different contexts. The major valuation methods are presented in Fig. 3 and explained in detail.

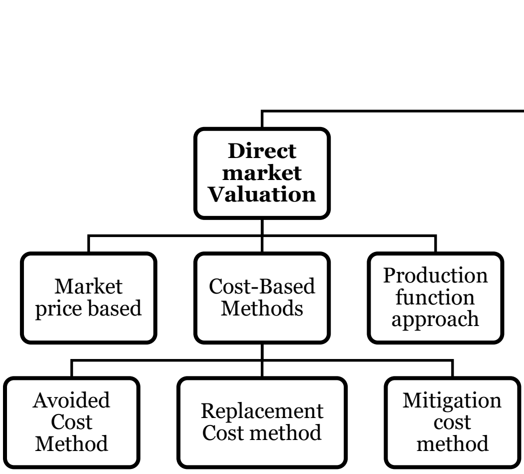


Fig.3. Popular methods used for the valuation of Ecosystem services

I. Direct market valuation

The approach utilizes data from established markets as the foundational framework for the valuation of ecosystem services. This entails employing actual market prices of goods and services, thereby anchoring the method in real-world market dynamics. The methods for direct market valuation can be broadly categorized into three primary approaches: (a) Market price based, (b) Cost-based methods, and (c) Production function based. This section is primarily based on the meta-analysis study conducted by Legesse et al. 2022 on valuation methods in ecosystem services.

(a) Market price based:

Market-price methods involve the utilization of observed prices from real markets that are pertinent to the provisioning of environmental goods or services (Arias et al. 2018). This approach is employed to approximate the economic value of ecosystem products or services that are actively exchanged within commercial markets. Consequently, this method can be effectively employed to assess alterations in either the quantity or quality of a good or service. However, not all ecosystem services have well-defined markets, and this approach may miss many non-market services.

(b) Cost-based methods:

The cost-based approaches are based on the assumption that the valuation of ecosystem services can be established by equating them to the expenses saved due to their natural existence. This methodology relies on estimations of the expenditures that would be required if the benefits derived from ecosystem services were to be replicated artificially (Moran, 2002). Various techniques are employed within this framework:

- *Avoided cost method*: The avoided cost method involves defining the value of ecosystem services as the expenses linked to the potential damage that was prevented through the presence of these services. This could also encompass costs that might have arisen in the absence of these services.
- *Replacement cost method*: This method estimates the value of an ecosystem service by calculating the cost of replacing it with a human-made alternative. For instance, if a wetland provides flood control, the cost of building and maintaining infrastructure like levees or floodwalls can be used as an estimate of the wetland's value.
- *Mitigation or Restoration cost method*: The mitigation or restoration cost method pertains to the expenses incurred in countering the effects resulting from the loss of ecosystem services or the costs associated with reinstating these services. This approach assumes that the valuation of ecosystem services is equivalent to the expenditure required for mitigating the negative consequences arising from ecosystem degradation (Kornatowska & Sienkiewicz, 2018; Koetse et al.2015).

c) Production function approach

The production function approach involves assessing the extent to which a specific ecosystem service enhances the provision of another service or commodity, which is actively traded within an established market. This method aims to estimate the economic valuation of ecosystem-related products or services that contribute to the generation of commercially viable goods. It is useful in scenarios where ecosystem products or services are utilized, in combination with other inputs, to produce a marketable product (Chee, 2004; Pascual et al.2010).

II. Revealed Preference Methods

The revealed-preferences method entails ascertaining the value attributed by consumers to an environmental good through the observation of their purchasing behavior in the market, particularly when their choices are directly or indirectly linked to the quality of the environment. Under this framework two major methods are:

- *Travel cost method:* This method is employed to assess the economic use value associated with ecosystems or sites utilized for recreational purposes. This approach involves analyzing the costs people incur to visit an area to enjoy its ecosystem services, such as recreational activities like hiking or birdwatching. The costs of travel, accommodation, and other expenses can help estimate the value people place on these experiences.
- *Hedonic pricing method*: The hedonic pricing method is employed to determine economic values for ecosystem or environmental services that have a direct influence on market

prices. This method involves studying property values to infer the value of nearby ecosystem services. For example, homes located near green spaces or bodies of water tend to have higher property values due to the aesthetic and recreational benefits provided by those ecosystems.

III. Stated Preference Methods

The methods are the only ones that can value environmental goods in cases where there are no markets to provide information on the value of environmental goods i.e. this approach can be used when no market prices are available and it is not possible to apply methods particular to the revealed preference approach. Under this approach the major techniques used for the valuation of ecosystem services are:

- *Contingent valuation method*: This method of valuation is primarily used in cost-benefit analysis and environmental accounting. In this approach, people are directly asked about their willingness to pay (WTP) for a specific ecosystem service or their willingness to accept (WTA) compensation for losing that service. It often involves surveys or hypothetical scenarios and can provide insights into the non-market value of services.
- *Choice experiments*: The choice experiments also use social surveys to extract individuals' preferences regarding various options, each characterized by distinct levels of ecosystem service attributes and their corresponding associated payments. This approach systematically analyzes individual responses to attribute levels (representing ecosystem services and payments) in order to determine the valuation of these ecosystem services (Reference 54 and 65).

Conclusion

The valuation of ecosystem services plays a pivotal role in recognizing the multifaceted contributions of natural systems to human well-being and sustainable development. Through a diverse array of valuation methods, we gain the tools to decipher the intricate web of benefits that ecosystems provide, translating their intrinsic worth into terms comprehensible to decision-makers and society. Each valuation method has its own strengths and limitations, and the choice of method depends on the specific ecosystem service being assessed, the available data, and the objectives of the valuation study. It is crucial to consider the context and potential biases when interpreting and applying the results obtained from these methods.

References

Arias-Arévalo, P., Gómez-Baggethun, E., Martín-López, B., & Pérez-Rincón, M. (2018). Widening the evaluative space for ecosystem services: A taxonomy of plural values and valuation methods. Environmental Values, 27(1), 29–53.

Chee, Y. E. (2004). An ecological perspective on the valuation of ecosystem services. Biological Conservation, 120(4), 549–565.

Costanza, R., De Groot, R., Braat, L., Kubiszewski, I., Fioramonti, L., Sutton, P., ... & Grasso, M. 2017. Twenty years of ecosystem services: how far have we come and how far do we still need to go? *Ecosystem services*, *28*, 1-16.

DEFRA.2007. An Introductory Guide to Valuing Ecosystem Services; Department for Environment, Food and Rural Affairs (DEFRA): London, UK.

ESB. Ecosystem Services & Biodiversity (ESB). (n.d.). Retrieved from https://www.fao.org/ecosystem-services biodiversity/en/#:~:text=Ecosystem%20services%20make%20human%20life,recreational%2C%20cultural%20and%20spiritual%20benefits.

Garbach, K., Milder, J. C., Montenegro, M., Karp, D. S., & De Clerck, F. A. J. 2014. Biodiversity and ecosystem services in agroecosystems. *Encyclopaedia of agriculture and food systems*, 2, 21-40.

Koetse, M. J., Brouwer, R., & Van Beukering, P. J. H. (2015). Economic valuation methods for ecosystem services. Ecosystem Services: From Concept to Practice, 108–131. https://doi.org/10.1017/CBO9781107477612.009

Kornatowska, B., & Sienkiewicz, J. (2018). Forest ecosystem services-assessment methods. Folia Forestalia Polonica, Series A, 60(4), 248–260. https://doi.org/10.2478/ffp018-0026

Legesse, F., Degefa, S., & Soromessa, T. (2022). Valuation Methods in Ecosystem Services: A Meta-analysis. World Journal of Forest Research,1(1):1-12.

MEA. 2005. Millennium Ecosystem Assessment Washington, DC: Island press.

Moran, D. (2002). Guy Garrod and Ken Willis 1999, Economic Valuation of the Environment: Methods and Case Studies . Environmental and Resource Economics 21, 101–102. https://doi.org/10.1023/A:1014547926638

Pascual, U., Muradian, R., Brander, L., Christie, M., Cornelissen, H., Eppink, F., Farley, J., Loomis, J., Pearson, L., Perrings, C., & Polasky, S. (2010). Chapter 5 The economics of valuing ecosystem services and biodiversity. The Economics of Ecosystems and Biodiversity. Ecological and Economic Foundations, January, 183–255.

Yadav, P., Jaiswal, D. K., & Sinha, R. K. 2021. Climate change: Impact on agricultural production and sustainable mitigation. *In Global Climate Change* (pp. 151-174). Elsevier.

Analyzing trade patterns of agricultural commodities-Application of the SMART Model

Renjini V R

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Concept and definitions

In the past decades, the world trading community has observed new regional integration initiatives called regional and bilateral trade agreements, collectively known as Regional Trade Agreements (RTAs). According to WTO, RTAs refer to reciprocal trade agreements between two or more countries to liberalize tariffs and services. Depending upon their level of integration, RTAs can be broadly divided into five categories: Preferential Trade Agreements, Free Trade Agreements, Customs Unions, Common Markets, and Economic Unions.

Preferential Trade Agreements (PTA)

A *Preferential Trade Agreement* is a union in which member countries impose lower trade barriers on goods produced within the bloc, with some flexibility for each member country on the extent of the reduction.

Free Trade Area (FTA)

Free Trade Area/Agreement is a special case of PTA where member countries completely eliminate trade barriers for goods originating within the member countries. In most cases, governments do not abolish trade barriers entirely, even within Free Trade Areas; most agreements tend to exclude sensitive sectors.

Customs Union (CU)

Customs Union provides deeper integration than FTA, where member countries are free to keep their individual tariff barriers for goods imported from non-member countries and apply a common external tariff on a good imported from outside countries.

Common Markets (CM)

The next level of integration is Common Markets, where member countries attempt to harmonize some institutional arrangements, commercial and financial laws, and regulations among themselves, which also entails free movements of factors of production.

Economic Union (EU)

The final integration level is the economic union, where member countries implement common economic policies and regulations and adopt a single currency (Das, 2001). Even though CU is perceived as optimal compared to FTA, the latter is more prevalent globally than in the former. Individual FTA members have the flexibility to form their own future FTAs policies, whereas CU members must jointly engage in future preferential agreements (Missios et al., 2012; Lake, 2014).

In recent years, global trade has been increasing through these kinds of bilateral and regional agreements. From about 50 till 1990 to over 250 in the 2000s, growth in the number of RTAs has been phenomenal and unprecedented. According to the WTO, as of 31st December 2022, the WTO received about 582 notifications of Regional Trade Agreements (RTAs), of which 355 were in force.

SMART (Software for Market Analysis and Restrictions on Trade) model is a partial equilibrium model to analyze the impact of tariff reduction policies on importing countries. This model and the simulation tools are part of the World Integrated Trade Solution (WITS) trade database and software suite provided jointly by the World Bank and the United Nations Conference on Trade and Development. The SMART model focuses on the changes in imports into a particular market when there is a change in trade policy. One of the advantages of using SMART is that the model can analyze the effects of trade policy reforms in the presence of imperfect substitutes and is more adequate than the homogenous good model while examining tariff preferences, as it avoids corner solutions. The model can be used to analyze the tariff effect of a single country market on disaggregated product lines. The theoretical framework of the model was first given by Laird and Yeats (1986).

The demand side of the market in SMART is based on the Armington assumption that commodities are differentiated by their country of origin. This assumption implies that, for a particular commodity, imports from one country are an imperfect substitute for imports from another country. Thus, even though an FTA entails preferential trade liberalization, import demand does not completely shift to a source from within the FTA. The SMART model also assumes that consumers' demand is decided in a two-stage optimization process involving allocating their spending by commodity and national variety. In the first stage, consumers decide how much to spend on the commodity, given changes in the price index of this commodity. The relationship between changes in the price index and the impact on import demand for this commodity is determined by a given import demand elasticity. In the second stage, the chosen level of spending for this commodity is allocated among the different national varieties, depending on the relative price of each variety. The substitution elasticity determines the extent of the between-variety response to a change in the relative price. Different countries compete to supply (export to) the market, and the model simulates changes in the composition and volume of imports into that market after a tariff reduction or another change in trade policy. The degree of responsiveness of each foreign exporter's supply to changes in price is known as the export supply elasticity. By default, The SMART model assumes that each foreign country's export supply elasticity is infinite, which implies that each foreign country can export as much of the good as possible at a specific price. This assumption may be appropriate for an importing country whose import quantity is too small to affect the prices of foreign exporters (i.e., the price-taker assumption). Suppose changes in the country's import quantity can affect the foreign exporter's price. SMART can operate with finite export supply elasticity, but the value of this parameter must be found and incorporated into the analysis.

SMART requires the following data, which can be extracted from WITS or imported from alternative sources of information, for the simulation of an FTA:

(i) Import value from each foreign partner

(ii) Tariff faced by each foreign partner

(iii) Import demand elasticity for the commodity

(iv) Export supply elasticity for the commodity

(v) Substitution elasticity between varieties of the commodity.

Kee, Nicita, and Olarreaga (2008) calculated the import demand elasticities provided in the model. SMART accepts just one import demand elasticity for the commodity, not one for each national variety. Moreover, the export supply elasticity must be the same for all foreign commodity exporters. SMART also expects that the substitution elasticity is the same for any pair of commodity varieties. Any preferential trade agreement results in trade creation and trade diversion effect (Viner,1950).

Trade creation

The trade creation effect is the increased demand in country j for commodity i from exporting country k resulting from the price decrease associated with the assumed complete transmission of price changes when tariff or non-tariff distortions are reduced or eliminated.

Trade diversion

The term trade diversion accounts for the tendency of importers to substitute goods from one source to another in response to a change in the import price of supplies from one source but not from the alternative source. Thus, if prices fall in one overseas country, there will be a tendency to purchase more goods from that country and less from countries with unchanged exports. Trade diversion can also occur not because of the change in the export price but because of the introduction or elimination of preferential treatment for goods from one (or more sources). In contrast, treatment for goods from other sources remains unchanged. Again, there could be a relative change in the treatment of the goods from different sources in the importing country by differential alterations in the treatment of different foreign suppliers.

Total trade effect

The total trade effect is obtained simply by adding the trade creation and trade diversion effects. Results can be summed for the imported across product groups and/or across sources of supply. Results can be summed across groups of importers for single products or groups of products and single supply sources or suppliers. Results can also be summed for suppliers across product groups. Finally, results can be summed for groups of suppliers either for individual products or across product groups.

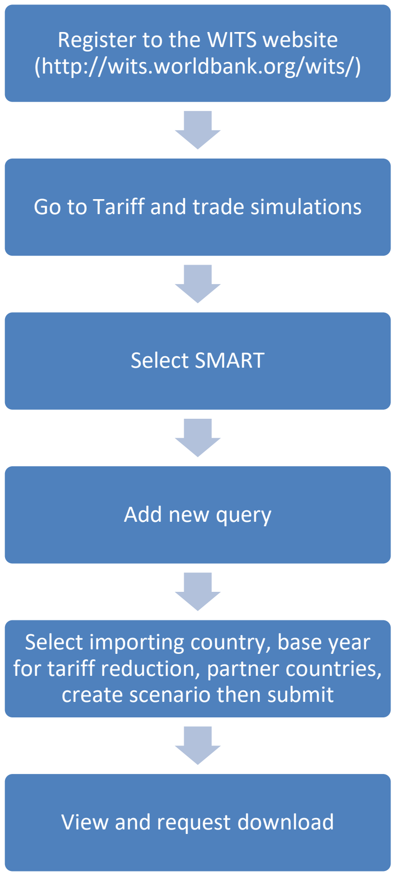
Welfare effect

The welfare effect implies that the benefits that consumers in the importing market derive from the lower domestic prices after the removal or reduction of tariffs or the ad valorem incidence of non-tariff distortions. The net welfare gain is usually estimated as the increase in import value times the average between the advalorem incidence of the trade barriers before and after their elimination. This welfare gain can also be considered an increase in consumer surplus.

Strengths and limitations of the SMART Model

The strengths of the SMART model are that it can be quickly learned and implemented with the WITS database, yields significant quantitative results on the trade and tariff revenue effects of an FTA, and the analysis can be performed at the most disaggregated level of trade data. However, the main limitation of the SMART model is that it is a partial equilibrium model, which means its results are limited to the direct effects of a trade policy change only in one market. In addition, SMART does not return results on an FTA's effects on domestic production, which may interest policymakers, nor does it consider the possibility of new foreign exporting countries serving the domestic market. Finally, SMART's results may be sensitive to the modeling assumptions and parameter values.

Steps in the SMART model simulation



References

Das, D. (2001) *Regional trading agreements and the global economy: an Asia-Pacific perspective*, Asian Development Bank, Manila.

Laird, S and Yeats, A (1986) The UNCTAD trade policy simulation model: a note on the methodology, data and uses, *UNCTAD Discussion Paper 19*, Geneva.

Kee, H L, Nicita, A and Olarreage, M. (2008) Import demand elasticities and trade distortions. *Review of Economics and Statistics*, **90** (4): 666-682.

Missios, P., Saggiy, K. and Yildizz, H.M. (2012) *Free Trade Agreements versus Customs Unions: Implications for Global Free Trade*, Department of Economics, Ryerson University, Canada.

Viner, J. (1950) *The customs union issue*, Oxford University Press, New York.

An Introduction to Systematic Reviews and Meta-Analysis

Praveen KV & Asha Devi S.S.
Division of Agriculture Economics, ICAR-Indian Agricultural Research Institute

Systematic reviews are a sort of literature review that utilizes systematic methods to gather secondary data and blend or synthesise the research evidence qualitatively or quantitatively. With the volume of research evidence on any topic growing at an ever-expanding rate, it is very difficult for individual researchers or policymakers to survey this tremendous amount of literature and arrive at the best decision on its basis. Following a systematic approach, systematic reviews help summarize the research knowledge on an intervention. It endeavours to gather all the empirical evidence that fits pre-determined eligibility criteria to answer to a particular research question. It utilizes systematic techniques that are chosen with the end goal of minimizing bias and hence giving more dependable findings from which conclusions can be drawn and choices made (Antman et al 1992, Oxman and Guyatt 1993).

Research questions

As in the case of any research, the first and most significant choice in setting up a systematic review is to decide its core interest. This is best done by framing the questions that the review looks to answer. Well-formulated questions will guide the systematic review procedure, including deciding eligibility criteria, literature search, gathering data from selected publications, organizing and presenting the findings (Cooper 1984, Hedges 1994, Oliver et al 2017). The FINER standards have been proposed to make life easy for a researcher while creating research questions. As per this strategy, questions ought to be Feasible, Interesting, Novel, Ethical, and Relevant (Cummings et al 2007). These measures raise key issues to be considered at the start of the review and ought to be borne as a primary concern when questions are framed.

A systematic review can address any research question that can be answered by primary research. Studies that compare interventions utilize the outcome of the participants to arrive at the impacts of various interventions. Statistical synthesis (for example meta-analysis) centres on comparison of a new intervention with the control. The differentiation between the outcomes of two groups treated contrastingly is known as the ‘effect’ or the 'treatment effect’. The primary objective of systematic reviews should be ideally framed in a single sentence. The objective can be structured as: 'To evaluate the impacts of [intervention or technology] for [income enhancement] in [types of individuals, region, and setting if specified]'. This may be trailed by at least one secondary targets, for instance identifying with various participant groups, varying comparison of interventions or diverse outcome measures. The detailing of review question(s) requires thought of a few key segments (Richardson et al 1995, Counsell 1997) which can be conceptualized by the 'PICO', an abbreviation for Population, Intervention, Comparison(s) and Outcome. The scope of the review should be just apt. It should not be too broad or narrow to be relevant.

Table 1. PICO formulation

Item	Example
Population	Farmers in developing countries Farmers involved in farmer groups or producer companies
Intervention	GM crops Integrated Pest Management
Comparator	Communities/famers not participating in FFS Farmers/communities receiving alternative interventions
Outcome	Yield Net revenue

Defining inclusion criteria

One of the highlights that differentiate a systematic review from a narrative review is that the authors of systematic review ought to pre-indicate criteria and standards for including and barring individual studies. When building up the protocol, one of the initial steps is to decide the components of the review question (the population, intervention(s), comparator(s) and outcome, or PICO components) and how the intervention, in the identified population, creates the outcomes. Eligibility criteria depend on the PICO components in addition to a specification of the kinds of studies that have addressed these inquiries. The population, intervention, and comparators in the review question can be usually translated into the inclusion criteria, but not always directly.

Literature search and study selection

Systematic reviews require a careful, objective, and reproducible search of a variety of sources to extract as many studies (eligible) as possible. The quest for studies should be as broad as possible to diminish the danger of reporting bias and to identify maximum evidence as possible.

Database determination ought to be guided by the survey theme. 'Grey literature' should also be considered. Authors ought to search for dissertations and conference abstracts also. They should also think about looking through the web, hand searching of journals and looking through full texts of journals electronically where accessible. They ought to inspect past reviews on a similar theme and check reference lists of included studies. Suitable search strategy should be formulated for searching in different databases. Choices about which studies to include for a review is among the most compelling choices that are made in the review procedure and they include judgment. Involvements of at least two individuals, working independently, are required to decide if each study meets the qualification standards. A PRISMA flow chart mentioning the selection of studies at each stage should be included in the report.

Table 2. List of databases to search

Sl No.	Database
1	Web of Science (Social science citation index)
2	CeRA
3	Google scholar
4	AgEcon search
5	Econlit
6	CAB abstract
7	Medline, Pubmed
8	ERIC

Coding and Data collection

Authors are urged to create layouts of tables and figures that will show up in the review to encourage the design of data collection forms. The way to effective data collection is to build simple-to-use forms and gather adequate and unambiguous information that present the source in an organized and structured way. Effort ought to be made to collect information required for meta-analysis. Data ought to be gathered and documented in a structure that permits future access and data sharing. Coding should provide for adding data in the following components:

- Study identification
- Intervention discriptives
- Process and implementation
- Context

- Popultion characteristics
- Research methods
- Effect size data
- Outcomes
- Subgroups

Effect measures

The kinds of outcome data that authors are probably going to experience are dichotomous data, continuous data, ordinal data, count or rate data and time-to-event data. The nature of the collected data determines the effect measures of intervention. Effect measures are statistical constructs that compare outcome data between two intervention groups. It is mainly of two distributed into two categories: ratio measures and difference measures. Estimates of effect describe the size of the intervention effect in terms of how diverse the outcome data were between the groups. For ratio effect measures, 1 indicates no distinction between the groups, while for difference measures, 0 indicates no distinction between the groups. Larger and smaller values than these 'null' values may suggest either benefit or harm of an intervention. The true effects of interventions very difficult to arrive at, and can only be assessed by the available studies. Estimates should thus be presented with uncertainty measures like confidence interval or standard error (SE). Examples of effect measures of dichotomous outcome data: Risk ratio, Odds ratio, Risk difference. Examples of effect measures of continuous outcome data: Mean difference, Standardised mean difference, Ratio of means

Meta-analysis

Meta-analysis can be considered as a key step in a systematic review. Meta-analysis involves deciding on the possibility of combining the results of selected studies. This procedure results in an overall statistic with a confidence interval that summarizes the effect of an intervention compared with the counterfactual. Meta-analysis is useful since they improve precision by including more information that smaller individual studies lack. To carry out a meta-analysis, at first, a summary statistic is computed for individual studies, to present the effect of the intervention in a uniform measure. Next, the individual study's intervention effects are statistically combined using a weighted average of the intervention effects estimated in the individual studies. Undertake random-effects meta-analysis if the studies are not all estimating the same intervention effect, but estimate intervention effects that follow a distribution across studies. On the other hand, if each study is estimating the same quantity, then a fixed-effect meta-analysis can be used. A confidence interval is derived that represents the precision of the summarized estimate. Meta-analysis can be carried out using two models:

- ***Fixed effect model***
  - Under the fixed-effect model we assume that all studies in the meta-analysis share a common (true) effect size.
  - Put another way, all factors that could influence the effect size are the same in all the studies, and therefore the true effect size is the same in all the studies.
  - Since all studies share the same true effect, it follows that the observed effect size varies from one study to the next only because of the random error inherent in each study.
  - If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect.
  - In practice, of course, the sample size in each study is not infinite, and so there is sampling error and the effect observed in the study is not the same as the true effect.
  - The observed effect for any study is given by the population mean plus the sampling error in that study.
- ***Random effects model***
  - There is no reason to assume that studies are identical in the sense that the true effect size is exactly the same in all the studies.
  - We might not have assessed these covariates in each study.
  - If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect for that study.
  - The sample size in any study is not infinite and therefore the sampling error is not zero. The observed effect for that study will be less than or greater than the true effect because of sampling error.
  - The distance between the overall mean and
  - the observed effect in any given study consists of two distinct parts: true variation in effect sizes (i) and sampling error
  - The observed effect for any study is given by the grand mean, the deviation of the

study's true effect from the grand mean, and the deviation of the study's observed effect from the study's true effect.

Meta-analysis: Demonstration (Example of meta-analysis of biofertilizer in India)

Setting the question

The effects of biofertilizer use in crop yields in India

- PICO
    - P- Experimental plots with biofertilizer application
    - I- Biofertilizer
    - C- Control plots
    - O- Yield

Search strategy for meta-analysis

A comprehensive literature search was undertaken from February to April 2019 in the google scholar, and CeRA (Consortium for e-resources in agriculture) to identify the studies to be included in the meta-analysis. The studies published between 2000 and 2019 were searched using the following search strings: "biofertilizer", "biofertiliser", "biofertilizer OR biofertiliser" AND "response" AND "India".

Screening, coding and data extraction

The studies were screened independently by authors to select the ones that meet the criteria to be included for the meta-analysis. The studies based on field trials, and that provide data for pairwise comparison of the yield effect of biofertilizer treated crop to that of the control are included. Full papers were reviewed to record the data on mean yields, standard deviations and the number of replications, and also other field-specific observations that would be required for the analysis. Out of the 16700 studies that appeared during the literature search, only 236 were selected after the preliminary screening to remove studies based on biofertilizer production technology, studies that are carried out in other countries, review studies, studies dealing with regulation and policies, and other aspects of biofertilizers that are not of our interest (these are termed as 'exclusion criteria). After removing the duplicate studies and the ones based on laboratory experiments, 86 studies were selected for full-text reviews. From this, only 18 articles were finally selected for the meta-analysis, as the others did not provide the information that we require for meta-analysis.

The flow of the search process is given in detail in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses(PRISMA) flow chart given in figure . The data from all the selected studies were then extracted and classified on the basis of types of biofertilizer. Nitrogen-fixing, phosphate solubilising, VAM, Combined biofertilizers, and others were the biofertilizer categories on the basis of which data extracted from the studies were grouped. Suitable predetermined codes were prepared in advance for this purpose. Example of coded sheet is given in the figure below. Further on the basis of crop groups, data were classified into that of cereals, legumes, vegetables and oilseeds. Thus from the 18 studies selected for meta-analysis, we were able to carry out 38 pairwise comparisons between biofertilizer treatment and control.

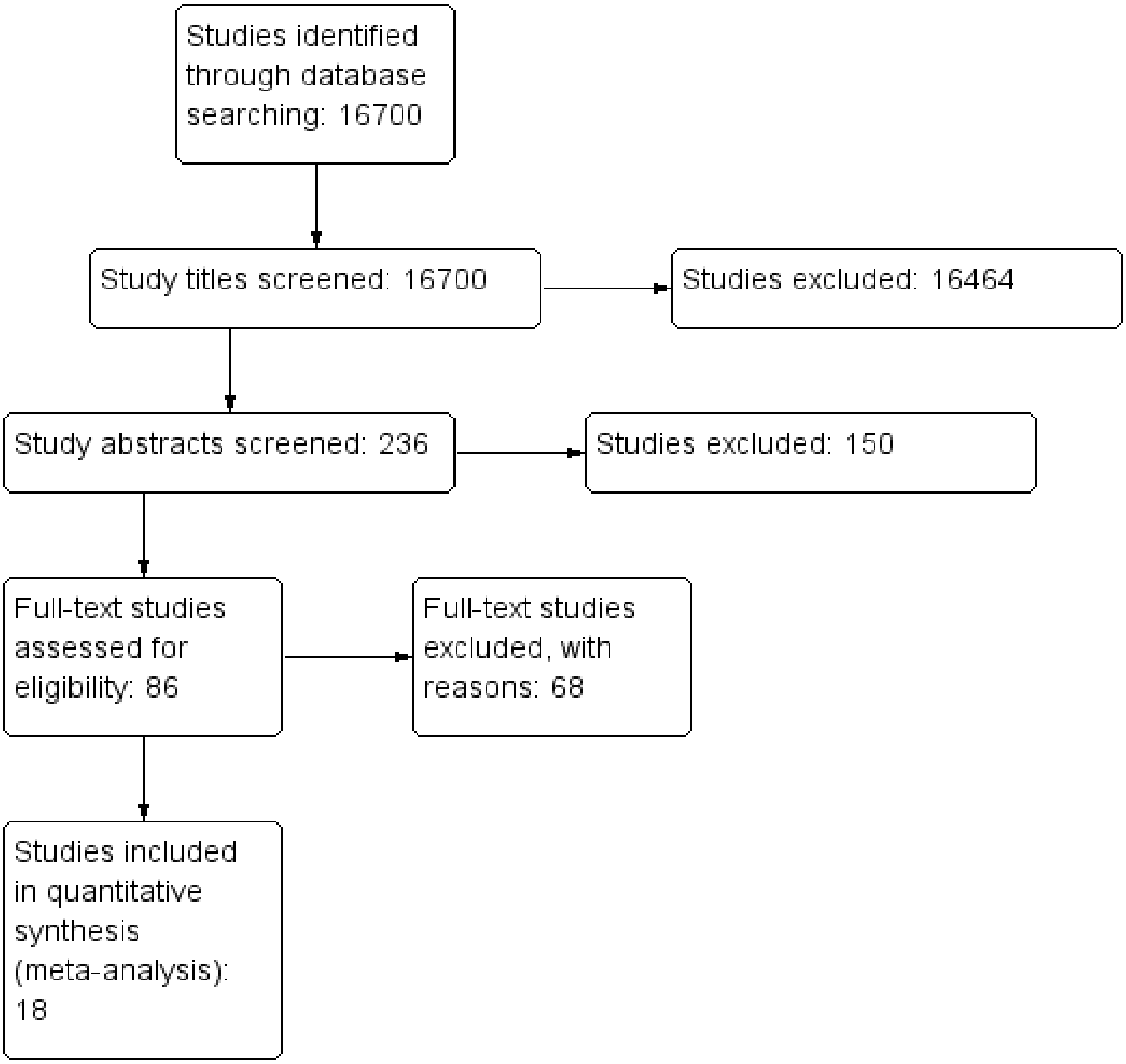


Figure 1. PRISMA flow chart

Paper no	Author	Effect size	Year	Crop	Location	Agro-ecological sub region (ICAR)	No of years of experiment	Soil	pH	Organic carbon %	Available N (kg/ha)	Available P(kg/ha)	Available K(kg/ha)	Biofertilizer species	Yield treatment (tonnes/ha)	Yield Control (tonnes.ha)	SD trt	SD control	Applied N (kg/ha)	Applied P(kg/ha)	Applied K(kg/ha)	Total N (available+applied) kg/ha	Total P (available+applied) kg/ha	Total K (available+applied) kg/ha	No. of replications	se
7	Upadhyay et al	1.62	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	Azospirillum	41.22	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
7	Upadhyay et al	1.85	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	PSM	41.88	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
7	Upadhyay et al	1.47	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	VAM	40.81	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
8	Yeptho	0.32	2012	Onion	Nagaland	Warm Pe	2	Sandy loa	4.5	2	212.3	10.5	173.2	Azotobacter	17.03	16.74	0.8515	0.837	104	32	152	316.3	42.5	325.2	6	0.35
10	Singh	1.29	2000	Potato	Meghalaya	Warm Pe	3	Sandy loa	5.4	1.7	172	8.2	235	Azotobacter	17	15.9	0.85	0.795	112	0	0	284	8.2	235	12	0.25
10	Singh	2.39	2000	Potato	Meghalaya	Warm Pe	3	Sandy loa	5.4	1.7	172	8.2	235	Phosphpbactrin	18	15.9	0.9	0.795	112	0	0	284	8.2	235	12	0.26
11	Gosh et al	0.55	2000	Potato	West Bengal	Hot Subh	2	sandy loa	6.2	1.2	165	13	122.5	Phosphert	17.24	15.75	2.49848	2.49848	120	44.5	83.5	285	57.5	206	6	1.02
16	Panwar	2.97	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azolla	42.64	36.27	2.132	1.8135	80	60	40	341.2	65.5	259.7	6	0.87
16	Panwar	1.61	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azospirillum	30.38	27.85	1.519	1.3925	0	0	0	261.2	5.5	219.7	6	0.62
16	Panwar	5.42	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azospirillum	41.44	30.72	2.072	1.536	60	45	30	321.2	50.5	249.7	6	0.85
18	Tagore et al	1.52	2013	Chickpea	Madhya Pradesh	Semi-arid	1	Clay loam	7.8	0.45	204	9.58	576	PSB	1.7	1.5	0.10219	0.10219	0	0	0	204	9.58	576	3	0.06
18	Tagore et al	3.2	2013	Chickpea	Madhya Pradesh	Semi-arid	1	Clay loam	7.8	0.45	204	9.58	576	Rhizobium	1.9	1.5	0.10219	0.10219	0	0	0	204	9.58	576	3	0.06
30	Kumar et al	6.09	2009	French b	Uttar Pradesh	Hot Sem	2	sandy loa	7.2	0.43	197.02	23.41	210	Biofertilizer	1.83	1.59	0.0915	0.0795	0	0	0	197.02	23.41	210	6	0.04
31	Kumawat et al	1.6	2010	Green gr	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	PSB	0.64	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
31	Kumawat et al	2	2010	Green gr	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	Rhizobium	0.66	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
31	Kumawat et al	5	2010	Green gr	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	Rhizobium+PSB	0.81	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
33	Singh et al	1.76	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	PSB	2.2	2	0.11	0.1	0	0	0	255.3	4.3	245	6	0.04
33	Singh et al	3.34	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	Rhizobium	2.4	2	0.12	0.1	0	0	0	255.3	4.3	245	6	0.05
33	Singh et al	3.98	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	Rhizobium+PSB	2.5	2	0.125	0.1	0	0	0	255.3	4.3	245	6	0.05
37	Sharma et al	0.11	2012	Pigeon p	Karnataka	Hot arid	3	Clay loam	8	0.5	180	25	350	Biofertilizer	0.014	0.013	0.009	0.009	25	50	0	205	75	350	9	0.00
39	Majumdar et al	2.27	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azospirillum	2.19	1.95	0.1095	0.0975	0	60	40	222.5	64.5	220	9	0.04
39	Majumdar et al	1.78	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azospirillum	3.38	3.08	0.169	0.154	60	60	40	282.5	64.5	220	9	0.06
39	Majumdar et al	3.03	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azotobacter	2.27	1.95	0.1135	0.0975	0	60	40	222.5	64.5	220	9	0.04
39	Majumdar et al	2.87	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azotobacter	3.58	3.08	0.179	0.154	60	60	40	282.5	64.5	220	9	0.06
43	Mathews et al	1.52	2006	Rice	Karnataka	Hot Hum	1	sandy loa	4.55	0.69	281	8.2	79	Azospirillum+PSB	5.71	4.53	0.62354	0.62354	0	0	0	281	8.2	79	3	0.36
43	Mathews et al	0.62	2006	Rice	Karnataka	Hot Hum	1	sandy loa	4.55	0.69	281	8.2	79	Azospirillum+PSB	8.88	8.4	0.62354	0.62354	75	75	90	356	83.2	169	3	0.36
45	Ghosh and Mohiuddin	1.05	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Bioplin	1.04	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
45	Ghosh and Mohiuddin	0.99	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Phosfert	1.02	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
45	Ghosh and Mohiuddin	0.98	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Vitormone	1.03	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
50	Behra and Rautaray	0.45	2009	Wheat	Madhya Pradesh	semi-arid	3	Clay loam	8.2	0.51	204	9.58	576	Azospirillum	4.78	4.67	0.239	0.2335	60	13.1	16.7	264	22.68	592.7	12	0.07

Figure 2. Coding

Meta-analysis

Mean difference was selected as the effect size. As per the results of the meta-analysis, application of biofertilizers resulted in an average yield increase of 0.36 tonnes per ha in India. The diamond shape gives the effect of subgroup and total biofertilizers. The size of the diamond shape gives the magnitude of the effect size and the edges represent the confidence interval (95% level). Meta-regression results suggest that only the combined biofertilizer application has a significant effect on yield improvement. The model, indicated significant yield increase due to biofertilizers in clay loam soil (in comparison to sandy loam), and soils with low K and high P content as well as low pH and low organic carbon content (in line with the findings of Schults, 2018). The variation in the performance of biofertilizers as per the agro-ecological conditions was also confirmed in this model. Most agro-ecological variables considered were significant. Among the crop groups, significant yield effects were detected in the case of cereals, legumes and vegetables (first model).

Study or Subgroup	Biofertilizer applied Mean	SD	Total	Biofertilizer not applied Mean	SD	Total	Weight	Mean Difference IV, Random, 95% CI
1.1.1 Nitrogen fixing biofertilizers								
Behra and Rautaray 2009	4.78	0.24	12	4.67	0.23	12	3.4%	0.11 [-0.08, 0.30]
Kumawat et al 2010	0.66	0.04	3	0.56	0.04	3	3.8%	0.10 [0.04, 0.16]
Majumdar et al 2007	3.38	0.17	9	3.08	0.15	9	3.5%	0.30 [0.15, 0.45]
Majumdar et al 2007	2.27	0.11	9	1.95	0.09	9	3.7%	0.32 [0.23, 0.41]
Majumdar et al 2007	2.19	0.11	9	1.95	0.09	9	3.7%	0.24 [0.15, 0.33]
Majumdar et al 2007	3.58	0.18	9	3.08	0.15	9	3.5%	0.50 [0.35, 0.65]
Malik et al 2009	3.29	0.07	6	2.41	0.07	6	3.7%	0.88 [0.80, 0.96]
Panwar 2014	42.64	2.13	6	36.27	1.81	6	0.2%	6.37 [4.13, 8.61]
Panwar 2014	30.38	1.51	6	27.85	1.39	6	0.4%	2.53 [0.89, 4.17]
Panwar 2014	41.44	2.07	6	30.72	1.54	6	0.2%	10.72 [8.66, 12.78]
Pathak and Godika 2010	1.19	0.06	3	1.15	0.06	3	3.7%	0.04 [-0.06, 0.14]
Singh 2000	17	0.85	12	15.9	0.79	12	1.5%	1.10 [0.44, 1.76]
Singh et al 2011	2.4	0.12	6	2	0.1	6	3.6%	0.40 [0.28, 0.52]
Tagore et al 2013	1.9	0.1	3	1.5	0.1	3	3.5%	0.40 [0.24, 0.56]
Upadhyay et al 2012	41.22	2.62	6	36.63	2.62	6	0.1%	4.59 [1.63, 7.55]
Yeptho 2012	17.03	0.85	6	16.74	0.84	6	0.9%	0.29 [-0.67, 1.25]
Subtotal (95% CI)			**111**			**111**	**39.3%**	**0.57 [0.35, 0.78]**
Heterogeneity: Tau² = 0.13; Chi² = 431.00, df = 15 (P < 0.00001); I² = 97%								
Test for overall effect: Z = 5.21 (P < 0.00001)								
1.1.2 Phoshate solubilising biofertilizers								
Behra and Rautaray 2009	4.76	0.24	12	4.67	0.23	12	3.4%	0.09 [-0.10, 0.28]
Ghosh and Mohiuddin 2000	1.02	0.14	6	0.87	0.14	6	3.5%	0.15 [-0.01, 0.31]
Ghosh et al 2000	17.24	2.49	6	15.75	2.49	6	0.1%	1.49 [-1.33, 4.31]
Kumawat et al 2010	0.64	0.04	3	0.56	0.04	3	3.8%	0.08 [0.02, 0.14]
Kumawat et al 2011	2.2	0.11	6	2	0.1	6	3.6%	0.20 [0.08, 0.32]
Pathak and Godika 2010	0.92	0.05	3	1.15	0.06	3	3.7%	-0.23 [-0.32, -0.14]
Singh 2000	18	0.9	12	15.9	0.79	12	1.4%	2.10 [1.42, 2.78]
Tagore et al 2013	1.7	0.11	3	1.5	0.1	3	3.5%	0.20 [0.03, 0.37]
Upadhyay et al 2012	41.88	2.62	6	36.63	2.62	6	0.1%	5.25 [2.29, 8.21]
Subtotal (95% CI)			**57**			**57**	**23.1%**	**0.22 [0.02, 0.42]**
Heterogeneity: Tau² = 0.06; Chi² = 98.22, df = 8 (P < 0.00001); I² = 92%								
Test for overall effect: Z = 2.20 (P = 0.03)								
1.1.3 VAM								
Behra and Rautaray 2009	4.78	0.24	12	4.67	0.23	12	3.4%	0.11 [-0.08, 0.30]
Upadhyay et al 2012	40.81	2.62	6	36.63	2.62	6	0.1%	4.18 [1.22, 7.14]
Subtotal (95% CI)			**18**			**18**	**3.5%**	**1.87 [-2.09, 5.82]**
Heterogeneity: Tau² = 7.13; Chi² = 7.21, df = 1 (P = 0.007); I² = 86%								
Test for overall effect: Z = 0.93 (P = 0.35)								
1.1.4 Combined biofertilizer application								
Behra et al 2007	4.39	0.26	20	4.19	0.26	20	3.5%	0.20 [0.04, 0.36]
Kumar et al 2009	3.36	0.17	6	2.39	0.12	6	3.5%	0.97 [0.80, 1.14]
Kumawat et al 2010	0.81	0.04	3	0.56	0.04	3	3.8%	0.25 [0.19, 0.31]
Kumawat et al 2011	2.5	0.13	6	2	0.1	6	3.6%	0.50 [0.37, 0.63]
Mathews et al 2006	8.88	0.62	3	8.4	0.62	3	0.8%	0.48 [-0.51, 1.47]
Mathews et al 2006	5.71	0.62	3	4.53	0.62	3	0.8%	1.18 [0.19, 2.17]
Pathak and Godika 2010	1.27	0.06	3	1.15	0.05	3	3.7%	0.12 [0.03, 0.21]
Subtotal (95% CI)			**44**			**44**	**19.7%**	**0.44 [0.21, 0.66]**
Heterogeneity: Tau² = 0.07; Chi² = 93.99, df = 6 (P < 0.00001); I² = 94%								
Test for overall effect: Z = 3.82 (P = 0.0001)								
1.1.5 Others								
Ghosh and Mohiuddin 2000	1.04	0.15	6	0.87	0.15	6	3.5%	0.17 [0.00, 0.34]
Ghosh and Mohiuddin 2000	1.03	0.15	6	0.87	0.15	6	3.5%	0.16 [-0.01, 0.33]
Kumar et al 2009	1.83	0.09	6	1.59	0.07	6	3.7%	0.24 [0.15, 0.33]
Sharma et al 2012	0.014	0.009	9	0.013	0.009	9	3.8%	0.00 [-0.01, 0.01]
Subtotal (95% CI)			**27**			**27**	**14.4%**	**0.14 [-0.02, 0.29]**
Heterogeneity: Tau² = 0.02; Chi² = 33.12, df = 3 (P < 0.00001); I² = 91%								
Test for overall effect: Z = 1.74 (P = 0.08)								
Total (95% CI)			**257**			**257**	**100.0%**	**0.36 [0.26, 0.46]**
Heterogeneity: Tau² = 0.07; Chi² = 1179.87, df = 37 (P < 0.00001); I² = 97%								
Test for overall effect: Z = 6.88 (P < 0.00001)								
Test for subgroup differences: Chi² = 13.01, df = 4 (P = 0.01), I² = 69.2%								

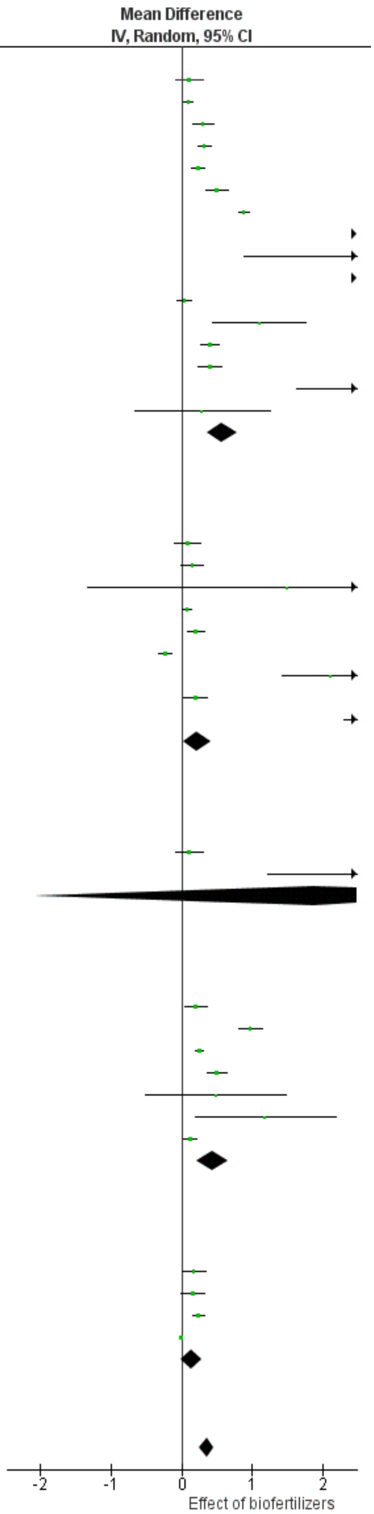


Figure 3. Forest plot

Table3. Meta-regression

Variables	Model with biofertilizer and agro-ecological groups	
	Coefficient	SE
Experiment duration	-6.99***	1.74
Ph	-2.50**	1.05
organic carbon	-2.03	1.76
Total N	0.00	0.01
Total P	0.03*	0.02
Total K	-0.04***	0.01
Replication number	1.35***	0.36
Clay loam	33.72***	8.42
VAM	-0.46	1.87
Combined biofertilizers	3.02**	1.25
Nitrogen fixers	0.51	1.11
Phosphate solubilizers	-0.38	1.01
Hot arid eco region	16.17**	5.81
Hot semi arid eco region	21.56***	4.65
Hot sub humid eco region	14.01***	3.50
Hot arid eco sub region	-8.67**	3.25
Northern plain	28.35***	5.00
Semi arid tropics	-1.18	1.51
Warm perhumid eco region	16.08***	3.66
Hot arid eco sub region	33.29***	6.84
Constant	16.54***	4.57
Observations	38	
R-squared adjusted	83.25	
F statistic	9.5	
Tau-sq	0.890	
I-sq	99.70	

References

Antman E, Lau J, Kupelnick B, Mosteller F, Chalmers T. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatment for myocardial infarction. JAMA 1992; 268: 240–248.

Cooper H. The problem formulation stage. In: Cooper H, editor. Integrating Research: A Guide for Literature Reviews. Newbury Park (CA) USA: Sage Publications; 1984.

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Annals of Internal Medicine 1997; 127: 380–387.

Cummings SR, Browner WS, Hulley SB. Conceiving the research question and developing the study plan. In: Hulley SB, Cummings SR, Browner WS, editors. Designing Clinical Research: An Epidemiological Approach. 4th ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2007. p. 14–22.

Hedges LV. Statistical considerations. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York (NY): USA: Russell Sage Foundation; 1994.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.

Oliver S, Dickson K, Bangpan M, Newman M. Getting started with a review. In: Gough D, Oliver S, Thomas J, editors. An Introduction to Systematic Reviews. London (UK): Sage Publications Ltd.; 2017.

Oxman A, Guyatt G. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125–133.

Praveen KV, Singh A. Realizing the potential of a low-cost technology to enhance crop yields: evidence from a meta-analysis of biofertilizers in India. Agricultural Economics Research Review. 2019;32(conf):77-91.

Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995; 123: A12 –13

Disclaimer: This chapter of the manual is prepared referring the sources listed in the reference section. This is prepared only for distribution to the participants of the training. Kindly cite the original sources while quoting the contents of this chapter. This chapter was earlier included in the training manual “Advances in Research Methodology for Social Sciences” released from the Division of Agricultural Economics.

Application of conjoint analysis for quantifying attribute preference

P. Venkatesh & Praveen.K.V.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction

Consumer behavioural analysis is the most important requirement for investors or entrepreneurs. Whenever the entrepreneur starts a new business or introducing a new product in the market, the first requirement is what is the demand for the product and what are all the quality attributes demanded by the consumers. The choice analysis methods are very useful for identifying attributes or traits of the products. Broadly there are two types of choice analysis namely, revealed preference (RP) techniques (eg.travel cost method, hedonic pricing analysis) and stated preference (SP) techniques (eg. contingent valuation methods, choice experiments, conjoint analysis). The RP techniques are based on the actual choice of the consumers, conversely, SP techniques are based on the hypothetical scenarios. For example, in a mobile shop, there are a number of brands (or same brand) with various attributes such as RAM, operating system, storage, camera quality and price. When we collect the sales data of mobiles along with it attributes and analyse and identifying the most preferred attributes then it is a type of revealed preference techniques, where we have observed actual purchasing behaviour of the consumers. Whereas, if we conduct the consumer survey with a hypothetical scenario of different attributes of mobiles phones and elicit the most preferred attributes of mobile phones, then it is a type of stated preference techniques.

Conjoint analysis

Conjoint analysis (CA) is one of the stated preference methods. It is used to understand the how consumers make complex choices among the various alternatives which has trade-offs. For example, one chooses low price mobile phones, then he (she) has to compromise OS and RAM etc. Everyone makes choices in day to day life like purchase of dress, mobiles, choosing restaurants for dinner etc which all involves mental conjoint analysis that contains multiple elements that lead us to our choice. CA is based on theory of demand (consumers derives utility or value from the attributes of the product) and theory of random utility (stochastic preference i.e. consumer may choose different choice from the same subset of alternatives at repeated presentation). Consumer choice decisions are based on the intrinsic and extrinsic cues of the product. The intrinsic cues are part of the physical properties of the product (eg. RAM. OS of mobile phone) and extrinsic cues are not part of the physical properties of the product (eg. price and brand). The traditional ranking method or rating survey cannot place the value for the attributes of the product. On the other hand, CA used to determine consumers preference by conjointly analysing their trade-offs between attributes. One of the advantages of the CA is that it provides relative importance of each attributes of the product (Lee et al, 2015).

Examples of conjoint analysis

CA mostly applied in market research analysis where to capture the consumer choice or preferences. Some of the examples for application of CA are as follows.

- Consumers preference for house: Consumers will be interested in location of the house like near to school, market, railway station, bus stations, hospital, size of the house, and other amenities and reprice of the house. Some consumers will be concentrated on price and some may be having preference for amnesties and some may be for locational advantage. The CA will identify the most preferred combination of attributes of the house as well as the importance of each attributes.
- Farmers preference for a variety: A variety may have various attributes like, duration, drought tolerance, pest and disease resistance, suitable for rainfed, yield, price of the seed, premium price in the market etc. The breeder cannot bring all the best attributes in single variety. Hence, he (she) wants to prioritize the breeding programme based on the farmers preference /need. CA can be useful tool to identify the most preferred attributes of the variety.

Steps in conjoint analysis

We will discuss the study on identification of preferred varietal attributes of pigeonpea variety by using hypothetical datasets.

1. **Identification of attributes and their levels:** Attributes are characteristics of the variety for example yield and it has three levels 10-15 q, 15-20 q and > 20 q. Similarly, other attributes ae and their levels are given below.

Sl. No.	Attributes	Levels		
1	Drought	Moderate resistant	Highly resistant	
2	Pod borer	Moderate resistant	Highly resistant	
3	Height	Short	Long	
4	Duration	Short (130 days)	Medium (13-150days)	Long (>150 days)
5	Yield (q/ha)	10-15 q	15 -20 q	> 20 q

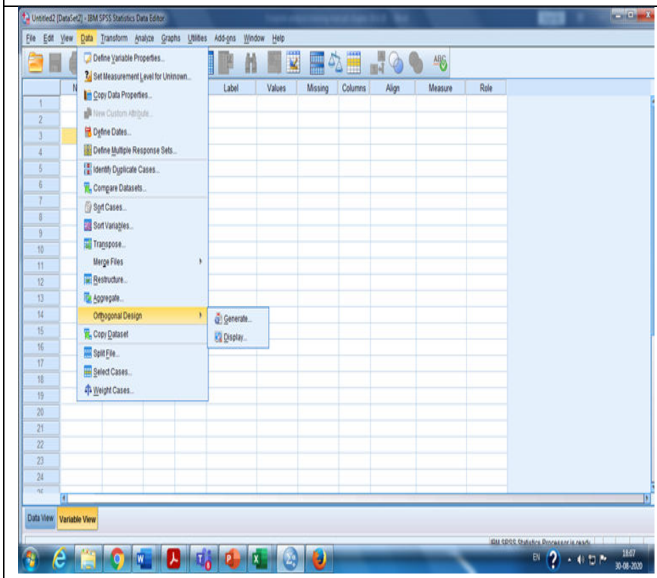
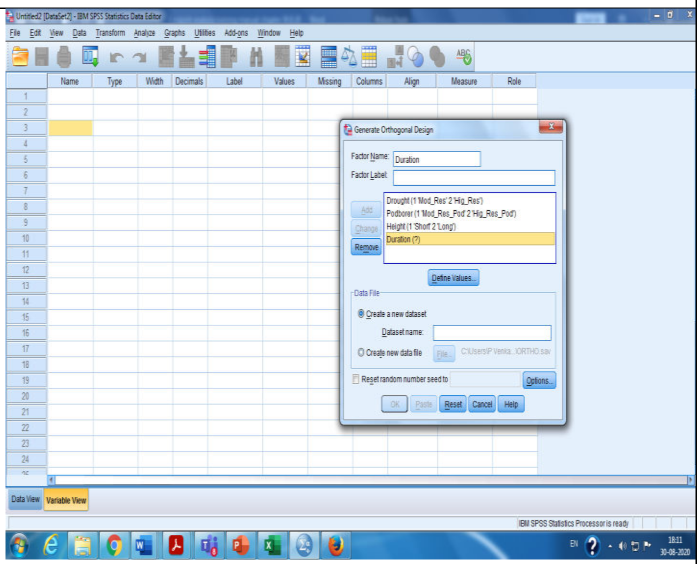
2. **Preparation of orthogonal design:** We have considered five attributes and each attribute is having different levels. In total (2x2x2x3x3=72) all possible combinations (sets) will be formed. However, it will be difficult for the respondents (farmers) to rank all these combinations. Therefore, by using orthogonal design, we can prepare a manageable number of combinations. Orthogonal Design procedure creates a reduced set of varietal combinations that is small enough to include in a survey but large enough to assess the relative importance of each attributes. By using SPSS, we can generate orthogonal design and plan file will be generated.

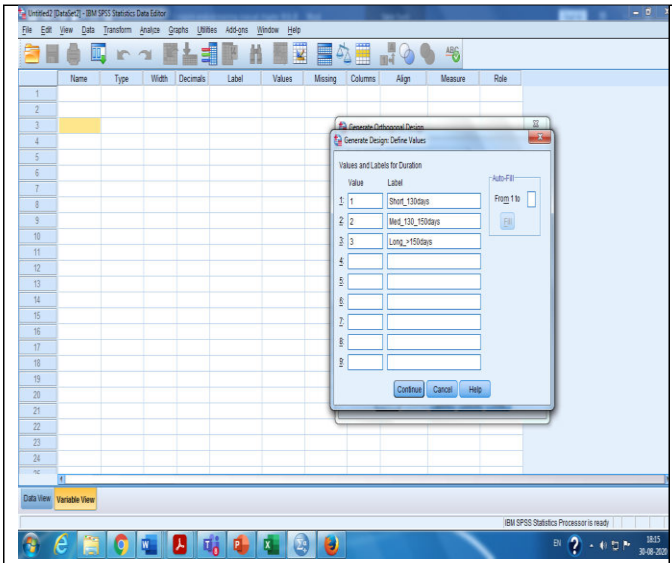
3. **Data collection:** A survey will be conducted among the farmers to rank the chosen level of combinations.

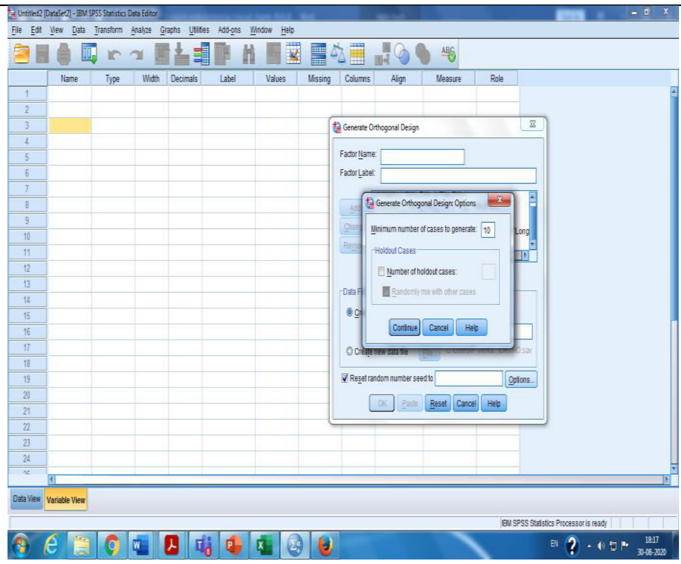
4. **Data analysis:** By using SPSS data can be analysed. Both plan file and data files are required for the analysis.

5. **Results:** SPSS will produce both utility files as well relative importance of the factors. By using utility values of each attributes, we can estimate the total utility value s of each combinations and we can find the most proffered combinations.

Analytical procedure in SPSS

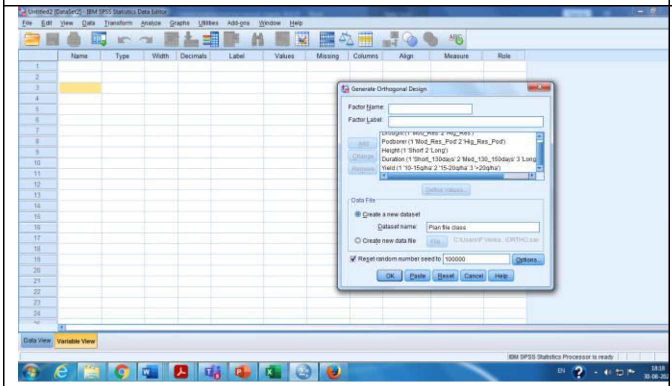
1.Orthogonal design create	**2.Defining factors**
	
3. Defining factor levels	**4.Setting minimum number of cases**

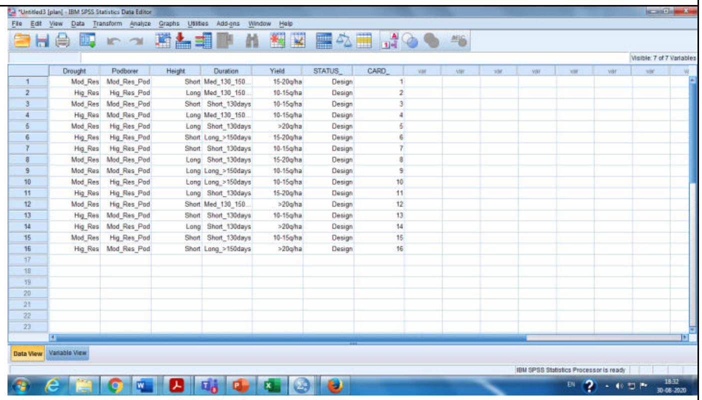




5.Setting seed value and defining file name

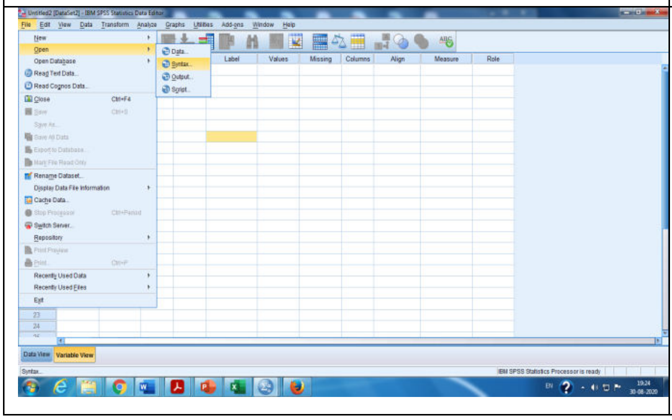
6.Plan file generated

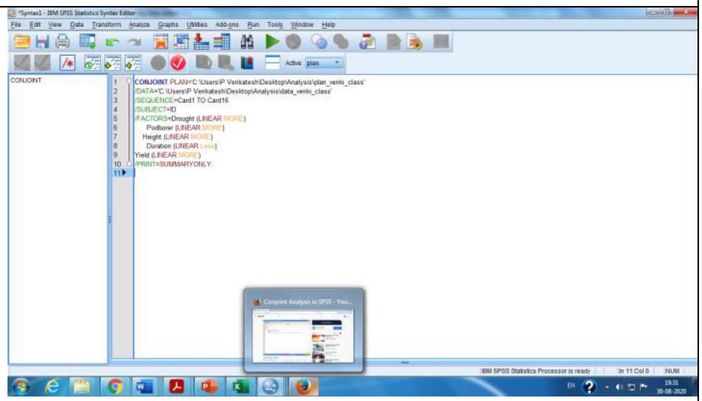




7.Opening of a syntax file

8.Syntax for analysis





Suggested readings

Lee P Y, Lusk K, Mirosa M, and Oey I (2015). An attribute prioritization-based segmentation of the Chinese consumer market for fruit juice. Food Quality and Preference, 46: 1–8.

IBM SPSS Conjoint 22 Available at: http://www.sussex.ac.uk/its/pdfs/SPSS_Conjoint_22.pdf.

Louviere, J J. (1991), "Analyzing Decision Making: Metric Conjoint Analysis", Sage University Paper Series on Quantitative Applications in Social Sciences, Series No. 07-067. Newbury Park, California.

Annunziata A and Vecchio R (2013) Consumer perception of functional foods: A conjoint analysis with probiotics. Food Quality and Preference, 28(1): 348-355

Praveen KV, Kuar S, Singh D R, Arya P, Chaudhary K and Kumar A (2013) A study on economic behaviour, perception and attitude of households towards traditional and modern food retailing formats in Kochi. Indian Journal of Agricultural marketing ,27 (2) 142-151

Disclaimer: This chapter of the manual is prepared referring the sources listed in the reference section. This is prepared only for distribution to the participants of the training. Kindly cite the original sources while quoting the contents of this chapter. This chapter was earlier included in the training manual "Advances in Research Methodology for Social Sciences" released from the Division of Agricultural Economics.

Effective Communication Skills for Effective Presentation

Girijesh Singh Mahra[1] & R. Roy Burman[2]

[1]Scientist, Division of Agricultural Extension, ICAR-IARI

[2]ADG (Agricultural Extension), ICAR

Communication

The word Communication is originated from latin word '*communis'* meaning 'common'. According to Leagans (1961), Communication is a process by which two or more people exchange ideas, facts, feelings or impressions in way that each gains a common understanding of meaning, intent and use of message. According to Roger and Shoemaker (1971) Communication is the process by which message are transferred from Source to receiver. Thus communication is a process of social interaction where we exchange ideas, facts, feelings or impressions to achieve common understanding. *Effective communication occurs when there is common understanding of meaning and intent*

Why effective communication skills?

The ability to communicate is the primary factor that distinguishes human beings from animals. And it is the ability to communicate well that distinguishes one individual from another. The fact, is that apart from the basic necessities, one needs to be equipped with habits for good communication skills, as this is what will make them a happy and successful social being. In order to develop these habits, one needs to first acknowledge the fact that they need to improve communication skills from time to time. They need to take stock of the way they interact and the direction in which their work and personal relations are going. The only constant in life is change, and the more one accepts one's strengths and works towards dealing with their shortcomings, especially in the area of communication skills, the better will be their interactions and the more their social popularity.

Every individual needs to be well equipped with the tools to communicate effectively, whether it is on the personal front, or at work. In fact, according to the management gurus, being a good communicator is half the battle won. After all, if one speaks and listens well, then there is little or no scope for misunderstanding. Thus, keeping this fact in mind, the primary reasons for misunderstanding is due to inability to speak well, or listen effectively.

Basic Communication Skills

Communication is essentially the transfer of ideas, messages or information from one person to another. It is effective when it gets the desired action or response. Basic communication skills are essential for continued success, whether personal or professional. At the very base one needs to understand the communication process.Thus, one may ask what are communication skills? To answer that simply - Basically, communicating is like a two-way street, which entails the relation between the sender and the receiver. In this process, a cycle of communicating messages is formed between the sender and the receiver. The sender is required to conceive the message he/she wishes to send, encode this message and then transmit. The receiver then is required to receive the message, decode is and clarify his/her understanding of the message.In order to maintain healthy communication, the two must go through this process, without bringing in other elements of intellectual thoughts and judgments, as they tend to harm the harmonious process of message passing and receiving.

From the sender's perspective one needs to have the following essential skills:

- Skills to compose the message
- Skills to send the message

From the receiver's perspective one needs to have the following essential skills:

The skill of receiving a message

- Without assumptions
- Placing biases aside
- Actively listening

Thus, the elements of effective communication are:

- Listening
- Verbal skills
- Non-verbal skills

Communication is generally classified into a couple of types. The classifications include: i) Verbal and non-verbal, ii) Technological and non-technological, iii) Mediated and non-mediated and iv) Participatory and non-participatory.

However, the commonly known types of communications are:

A. Intra-personal communication skill:

This implies individual reflection, contemplation, and meditation. One example of this is transcendental meditation. According to the experts this type of communication encompasses communicating with the divine and with spirits in the form of prayers and rites and rituals.

B. Interpersonal communication skill:

This is direct, face-to-face communication that occurs between two persons. It is essentially a dialogue or a conversation between two or more people. It is personal, direct, as well as intimate and permits maximum interaction through words and gestures. Interpersonal communications maybe:

1. **Focused Interactions:**

   This primarily results from an actual encounter between two persons. This implies that the two persons involved are completely aware of the communication happening between them.

2. **Unfocused interactions:**

   This occurs when one simply observes or listens to persons with whom one is not conversing. This usually occurs at stations and bus stops, as well as on the street, at restaurants, etc.

3. **Non verbal communication skills:**

   This includes aspects such as body language, gestures, facial expressions, eye contact, etc., which also become a part of the communicating process; as well as the written and typed modes of communications.

C. Mass communication:

This is generally identified with tools of modern mass media, which includes: books, the press, cinema, television, radio, etc. It is a means of conveying messages to an entire populace.

No matter what the different types of communication skills are, communicating is an ever-continuing process that is going on all the time. It is as important to human life as is day-to-day existence.

Effective Communication Skills

More often than never, most people consider themselves to be good and effective communicators simply because they feel they can speak fluently. While speaking fluently is an important aspect of communicating, yet it is not the only requirement. One should be able to listen effectively, speak fluently and clearly, write well and read in the language/s they are familiar with. Apart from these basic aspects of communications, one needs to keep in mind the non-verbal aspects too, in order to be considered adept in communication skills. The fact is that one needs to constantly work towards developing effective communication skills. And primarily they need to overcome the barriers to effective communication. And this can be done when they are aware of the barriers and shortcomings. This is in fact the first and foremost primary step to being good communicator. Given here are some of the barriers that occur in communicating effectively. Understanding these barriers will help one comprehend examples of communicating skills. After all breaking down barriers implies setting good examples...

The verbal barriers which need to be avoided are:

Attacking:

Interrogating, Criticizing, Blaming, Shaming

You messages:

Moralizing, Preaching, Advising, Diagnosing, Endorsing Power, Ordering, Threatening, Commanding, Directing

Shouting

Name-calling, Refusing to talk

The non-verbal barriers are:

Flashing eyes, Rolling eyes, Quick movements, Slow movements, Arms crossed, Legs crossed

Gestures out of exasperation

Slouching, Hunching, Lack of personal hygiene, Doodling, Avoiding eye contact, Staring at people, Over fidgeting

Needed Verbal Communication Skills

Everybody has interesting thoughts floating in their mind, however only a few are able to communicate them effectively, and bring about a resounding impact on their audience. This is because they have probably sharpened their verbal communication skills. Many feel that this skill does not need any training, as every individual is able to communicate. Yes, every individual can communicate, but the problem is that every individual cannot effectively communicate.

Then the common question that arises is: 'how to improve my communication skill'. Though the years, experts in the field of training have found innovative ways and have provided interesting tips and methods to improve your communication skills. **Given here are some interesting tips ways in which one can improve the way in which they communicate:**

A. **Be aware of the communication process:** One should be aware of every aspect of the present communication - the purpose, objective and needs. One needs to be aware of what is occurring within the self; aware of what the others present feel; aware of all that is occurring between the communicators and aware of all that is happening around the communicators.

B. **Digging deeper:** One should be able to dig below the surface and derive and understands each communicator's primary needs from the conversation taking place.

C. **Clarity of thought:** One needs to be clear and focused on the subject at hand and not beat around the bush and be ambiguous.

D. **Listening empathetically:** One should hone the skills of listening with understanding.

E. **Assert respectfully:** It is important that one develops speaking up assertive communication skills. This is because when one is assertive, they are proving that they are confident about what they need to convey.

F. **Conflict resolution:** One should be able to come to win-win solutions in order to solve all problems that may occur from time-to-time.

Based on the communication skills training programs conducted by known experts in the field, here are some tips to good communication skills:

1. **Maintain eye contact with the audience:** This is vital as it keeps all those present involved in the conversation. It keeps them interested and on the alert, during the course of the conversation.
2. **Body awareness:** One needs to be aware of all that their body is conveying to them, as well as others. For instance, if there is anxiety rising during the course of a conversation then one feels thirsty and there may be a slight body tremor. At that point one needs to pause and let someone else speak. A few deep breaths and some water works as the magic portion at this point.
3. **Gestures and expressions:** One needs to be aware of how to effectively use hand gestures and the way they need to posture their body to convey their messages effectively.

Sometimes it may happen that they verbally convey something, but their gestures and facial expressions have another story to tell.

4. **Convey one's thoughts:** It is important for one to courageously convey what they think. This is because when things are left unsaid, then what is being spoken is not as convincing as it should be. Then a lack of confidence develops.
5. **Practice effective communication skills:** One should practice speaking and listening skills as often as possible. In order to practice effective speaking skills one can read passages from a book aloud, in front of a mirror, or simply perform a free speech in front of the mirror. And where listening is concerned, one can try transcribing from the radio or television, etc. this helps in honing sharper listening skills.

The ability to communicate effectively is a trick learnt by many, but practiced perfectly by not too many. This is because for most communicating is simple process. However, it is not so, it a rather simple-complex-networking system that has varied undercurrents flowing between the speaker and listener/s.

Communications skills for Public speaking

Public speaking (sometimes termed oratory or oration) is the process or act of performing a presentation (a speech) focused around an individual directly speaking to a live audience in a structured, deliberate manner in order to inform, influence, or entertain them. Public speaking is commonly understood as the formal, face-to-face talking of a single person to a group of listeners. It is closely allied to "presenting", although the latter is more often associated with commercial activity. Most of the time, public speaking is to persuade the audience.In public speaking, as in any form of communication, there are five basic elements, often expressed as *"who is saying what to whom using what medium with what effects?"* The purpose of public speaking can range from simply transmitting information, to motivating people to act, to simply telling a story. Good orators should not only be able to engage their audience, but also able to read them. The power of a truly great presenter is the ability to change the emotions of their listeners, not just inform them. Public speaking can also be considered a discourse community. Interpersonal communication and public speaking have several components that embrace such things as motivational speaking, leadership/personal development, business, customer service, large group communication, and mass communication. Public speaking can be a powerful tool to use for purposes such as motivation, influence, persuasion, informing, and translation

Effective public Speaking

For becoming effective public speaker following points has to be kept in mind

1. **Know Your Audience**

   Before going to prepare your speech/content always investigate about educational background, beliefs, language, norms, and customs of your audience. Prepare the content in that language which is understandable by audience and relevant to their needs and have suitability with norms and mores of society. Knowing as much as you can about their knowledge, their age, the number of people you will be lecturing or speaking to. If you get some idea of what they hope to achieve from the speech, then that will help you to prepare in a more effective way.

2. **Know your purpose**

   Prepare the content according to the occasion. There are different styles of preparation of content for debate, declamation, teaching situation or speech. Be clear about your purpose.

3. **Command on Language and Subject Matter**

   A good speaker has rich and diverse vocabulary of the language in which he communicates. Always choose that language in which you and audience are comfortable.

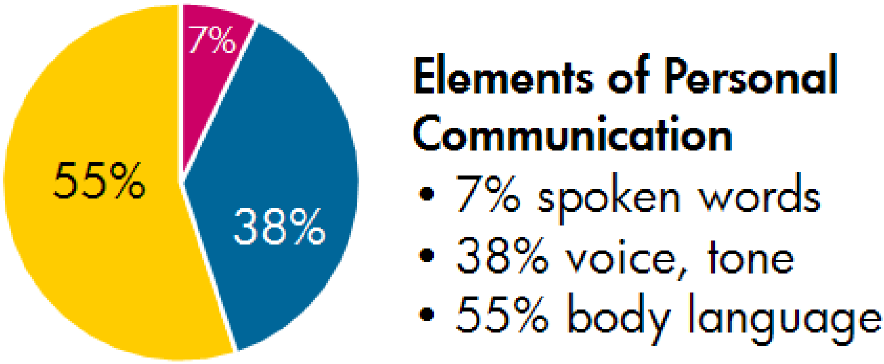
   Command over subject matter on which you are going to speak is the prerequisite of effective communication. Once you have well prepared your content then it will automatically give you inner confidence.

   Content should be Simple, understandable, attractive, free of jargons. Make message SMART (simple, measurable, authentic, reliable and timely).

4. **Verbal Communication (tone, pitch and modulation)**

   When your name is announced, from that very moment till you return after finishing your speech, communication goes on. Your starting/opening of speech be should be attractive and catchy. *Play with diversity of tone and pitch*. There should be balanced up and down of tone and pitch. Provide effects and punch at suitable places. Verbal Communication varies according to the speaking platform example for debates speaking needs to be argumentative, for declamation speech should contains both pro and cons of the topic. Adapt your voice according to the occasion. A large group lecture (50+) will be far more formal than a small group. For small groups you can include them in the speech, by asking questions or involving them. Be aware than any holes in your research will be noticed and questioned by someone in the audience, so to prevent any awkward moments make sure there are no holes, sweeping statements or uncertainties. Never assume "no one will notice" or "no one will ask about that". Always assume someone will ask. Try to think of all the questions you could be asked and make sure that you have an answer.

5. Non Verbal Communication



Maximum communication occurs through non verbal communication. Always wear cloths suitable to the occasion (formals with soothing color cloths). Gestures and postures plays a crucial role in making your audience understands. Use hand movements in balanced way and in synchronization with your verbal communication. Practice the speech at home, and time it.

6. The 6 I's of credibility for public speaking

Ideation	Be creative in presenting the idea
Information	Bring out new and decision driving facts
Influence	Be charismatic with show of confidence
Integrity	Be authentic and build a trust through the first half of the session
Impact	Identify and present a memorable delivery to root the message
Ignition	Call out to action, if required (E.g. Funding, Social Action etc.)

7. Timing

You will be given a time for your speech, sometimes these are flexible (between 10-15 minutes for example) which is good, as going over time is bad manners. However, if you are told you have 20 minutes make sure you stick to it, as you cannot rush your lecture to finish as this will affect delivery, and you do not want to be stopped before the end. If you have a time slot of 20 minutes, for example, the audience will know this, and will start fidgeting as you go over time, which is distracting for you, and you are certain they are not listening, but rather thinking about their lunch or tea-break.

Always read your notes before you arrive at the venue so it is fresh in your mind which will eliminate a total reliance on notes. Also remember that you are the only one who knows exactly what you intend to speak about, so if you miss something out no one else will know.

8. Listen to Great Orator

Try to develop a habit of listening great orators in History like Martin Luther King, M K Gandhi, Wiston Churchill, Swami Vivekananda etc.Learn new words daily and try to use then next time when you speak.

Useful References for public speaking

- http://www.wikihow.com/Be-an-Effective-Public-Speaker
- http://www.artofmanliness.com/2008/08/01/the-35-greatest-speeches-in-history/
- Swami Vivekananda Chicago Speech on 15th September, 1893 https://www.youtube.com/watch?v=TlwZNmgFBWM
- Mahtma Gandhi Speech London 1931 https://www.youtube.com/watch?v=_SakitCoNYc
- Martin Luther King, Jr. I Have A Dream Speech https://www.youtube.com/watch?v=3vDWWy4CMhE

Personality Development: Understanding inner and outer self

Girijesh Singh Mahra[1] & R. Roy Burman[2]
[1]Scientist, Division of Agricultural Extension, ICAR-IARI
[2]ADG (Agricultural Extension), ICAR, New Delhi

The Concept of Personality

Personality refers to the enduring characteristics and behavior that comprise a person's unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns. Various theories explain the structure and development of personality in different ways, but all agree that personality helps determine behavior. The field of personality psychology studies the nature and definition of personality as well as its development, structure and trait constructs, dynamic processes, variations (with emphasis on enduring and stable individual differences), and maladaptive forms.

Personality is that pattern of characteristic thoughts, feelings, and behaviours that distinguishes one person from another and that persists over time. It is the sum of biologically based and learnt behaviour which forms the person's unique responses to environmental stimuli.

Personality Development:

Personality development includes activities that improve awareness and identity, develop talents and potential, build human capital and facilitate employability, enhance quality of life and contribute to the realization of dreams and aspirations. When personal development takes place in the context of institutions, it refers to the methods, programs, tools, techniques, and assessment systems that support human development at the individual level in organizations.

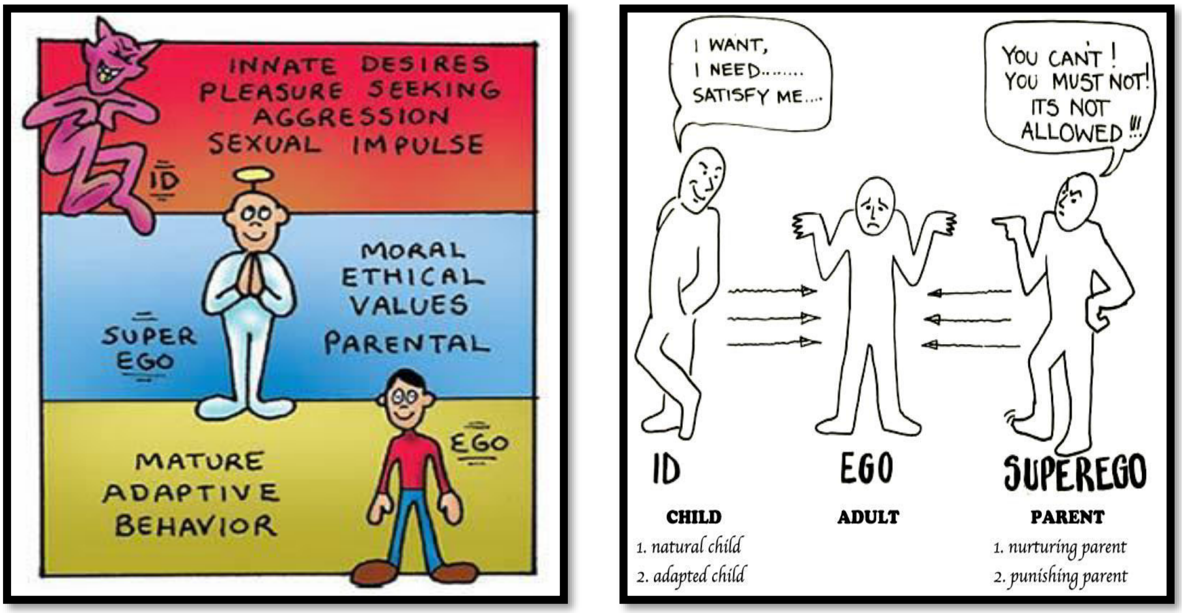
Personality development includes activities that develop talents, improve awareness, enhances potential and looks to improve the quality of life. It involves formal and informal activities that put people in the role of leaders, guides, teachers, and managers for helping them realize their full potential. Hence, it can be concluded that the process of improving or transforming the personality is called personality development.

Most people underestimate the importance of having a pleasing personality. Majority think it just means being born good-looking, that there isn't anything much to do about it. But this is not true. The scope of personality development is quite broad. It includes knowing how to dress well, social graces, grooming, speech and interpersonal skills. Whatever your career, these are very important skills that will promote your objectives. To better appreciate its importance, some of the key benefits of developing your personality include the following:

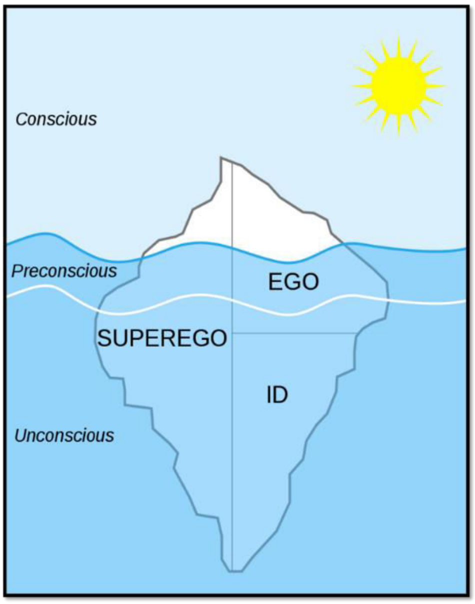
- **Confidence:** Personality development gives more confidence to people. When you know you are appropriately attired and groomed, this makes you less anxious when meeting a person. Knowing the right things to say and how to conduct yourself will increase your confidence.
- **Credibility:** Personality development makes people more credible. Despite the saying that you don't judge a book by its cover, people do tend to judge people by their clothing and how it is worn. This does not mean buying expensive clothes. We all know people who look shabby in expensive clothes. There are also people who look great even if their attire is inexpensive. Because of this, you must know what to wear and you must be aware of other aspects of enhancing your physical features.
- **Interaction:** Personality development encourages people to interact with others. Studies have consistently shown that people communicate more openly with people they are comfortable with. If your hygiene and social graces are unrefined, then expect to have a much harder time connecting with people.
- **Leading and Motivating:** Personality development enhances the capacity to lead and motivate. A person with a winning personality will be able to motivate better. People are less likely to get bored, and our ideas will have more credibility. We can lead better if we project an aura of confidence and credibility.
- **Curiosity:** A single wrong word can destroy a business relationship. Knowing the right things to say shows both respect and intellectual sophistication. This is especially the case if you are dealing with foreigners or if you conduct business outside the country. The right thing to do in our country could be horrible blunders in a different culture. These are the soft skills that may break or make a deal.
- **Communication skills:** It improves your communication skills. People are more receptive to what you say if they are impressed with your personality. Verbal communication skills are also part of personality development; improving your speech will strengthen the impact of your message. You cannot win by talent and hard work alone. Personality development is a crucial ingredient that you must obtain. Most of the people you see as models of great personality have taken a lot of effort in developing their natural features

Freud View of Personality: Freud theorized that personality contains three structures—the id, ego, and superego—and that the mind is like an iceberg, the unconscious making up 90% while the conscious (like the tip of the iceberg floating above water) makes only 10% of the mind. Freud suggested an analogy about the mind. He said that the mind is like an iceberg in the ocean, floating 10% above the water and 90% below. The unconscious, Freud proposed, makes up most of our mind. In Freud's view, only about 10% of our behaviours are caused by conscious awareness—about 90% are produced by unconscious factors. According to psychoanalytic theory, most of what controls our behaviours, thoughts, and feelings is unknown to our aware minds. Normally, the unconscious guides us. Freud said that the mind could be divided into three abstract categories. These are the id, the ego, and the superego.

- ***The id:*** Latin for the term "it," this division of the mind includes our basic instincts, inborn dispositions, and animalistic urges. Freud said that the id is totally unconscious, that we are unaware of its workings. The id is not rational; it imagines, dreams, and invents things to get us what we want. Freud said that the id operates according to the pleasure principle—it aims toward pleasurable things and away from painful things. The id aims to satisfy our biological urges and drives. It includes feelings of hunger, thirst, sex, and other natural body desires aimed at deriving pleasure.
- ***The ego:*** Greek and Latin for "I," this personality structure begins developing in childhood and can be interpreted as the "self." The ego is partly conscious and partly unconscious. The ego operates according to the reality principle; that is, it attempts to help the id get what it wants by judging the difference between real and imaginary. If a person is hungry, the id might begin to imagine food and even dream about food. (The id is not rational.) The ego, however, will try to determine how to get some real food. The ego helps a person satisfy needs through reality.
- ***The superego:*** This term means "above the ego," and includes the moral ideas that a person learns within the family and society. The superego gives people feelings of pride when they do something correct (the ego ideal) and feelings of guilt when they do something they consider to be morally wrong (the conscience). The superego, like the ego, is partly conscious and partly unconscious. The superego is a child's moral barometer, and it creates feelings of pride and guilt according to the beliefs that have been learned within the family and the culture.



How Id Ego and Superego affect us: Freud View of Personality



Freud View of Personality

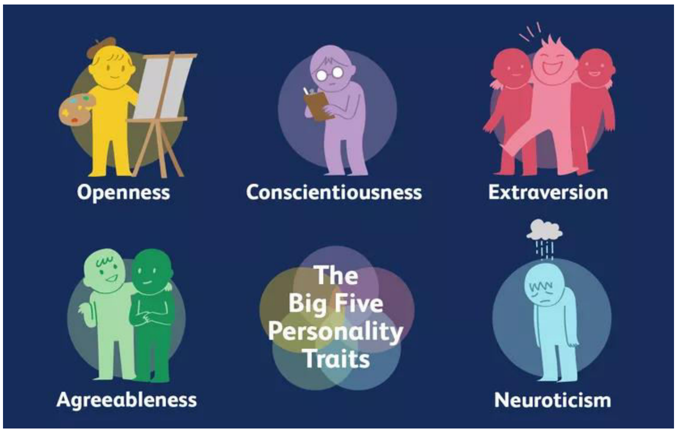
5-Factor Model of Personality (Big 5 personality Traits)

The Big 5 personality traits emerged are used to describe the broad traits that serve as building blocks of personality.

- ***Openness:*** Openness (also referred to as openness to experience) emphasizes imagination and insight the most out of all five personality traits. People who are high in openness tend to have a broad range of interests. They are curious about the world and other people and are eager to learn new things and enjoy new experiences.People who are high in this personality trait also tend to be more adventurous and creative.

Conversely, people low in this personality trait are often much more traditional and may struggle with abstract thinking.

- ***Conscientiousness:*** Among each of the personality traits, conscientiousness is one defined by high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines. Someone scoring lower in this primary personality trait is less structured and less organized. They may procrastinate to get things done, sometimes missing deadlines completely.
- ***Extraversion:*** Extraversion (or extroversion) is a personality trait characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness. People high in extraversion are outgoing and tend to gain energy in social situations. Being around others helps them feel energized and excited.People who are low in this personality trait or introverted tend to be more reserved. They have less energy to expend in social settings and social events can feel draining. Introverts often require a period of solitude and quiet in order to "recharge."
- ***Agreeableness:*** This personality trait includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative while those low in this personality trait tend to be more competitive and sometimes even manipulative.
- ***Neuroticism:*** Neuroticism is a personality trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in neuroticism tend to experience mood swings, anxiety, irritability, and sadness. Those low in this personality trait tend to be more stable and emotionally resilient.



Big 5 personality Traits

Johari window model and Personality

The Johari window model was developed by American psychologists Joseph Luft and Harry Ingham in the 1950s, while they were researching group dynamics. Today the Johari window model is especially relevant because of the modern emphasis on soft skills, behavior, empathy, cooperation, inter-group development and interpersonal development. Johari window model is sometimes called: a disclosure / feedback model of self-awareness. Basically, Johari window is an information processing tool. It helps enhance individual's perception on others. This model is based on two ideas that you can:

- *Acquire trust by revealing information about yourself to others, and*
- *Learn more about yourself from their feedback*

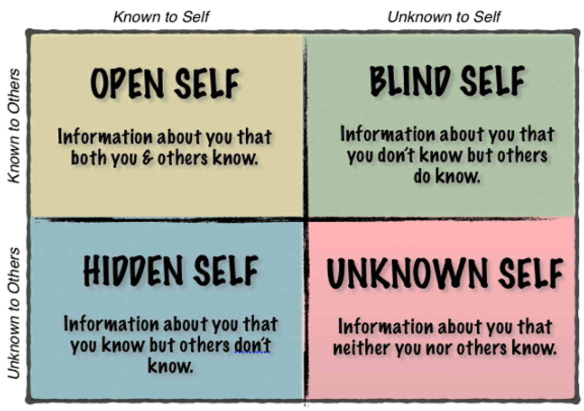
Johari window represents such information as feelings, experience, views, attitudes, skills, intentions, motivation, etc. from four perspectives. These perspectives capture assessments pertaining to a person or about that person's relation to their group.

The Johari window is generally used for teaching and considering and administering an understanding how individuals:

- Communicate with themselves and with others
- Present themselves to themselves and to others
- Perceive their place in the world

With a little consideration, Johari is also suitable for multiple usage as a:

- Coaching tool to to facilitate conversations around actions vs. perceived motivations
- Organizational Development tool to visualize the political and cultural issues that may be in or out of sync within a business
- Management tool to demonstrate team dynamics
- Self-development tool that helps to consider one's own behavior vs. reaction



The Johari window model

The four quadrants

Open: The open area is that part of our conscious self – our attitudes, behavior, motivation, values, way of life – that we are aware of and that is known to others. We move within this area with freedom. We are "open books".

Façade/hidden: Adjectives selected by the subject, but not by any of their peers, go in this quadrant. These are things the peers are either unaware of, or that are untrue but for the subject's claim.

Blind: Adjectives not selected by subjects, but only by their peers go here. These represent what others perceive but the subject does not.

Unknown: Adjectives that neither the subject nor the peers selected go here. They represent the subject's behaviors or motives that no one participating recognizes—either because they do not apply or because of collective ignorance of these traits.

How to develop your personality:

Become a Better Listener: If you are a good listener, you can learn a lot from your surroundings. It allows people to be more open to you and they will comfortably share any information with you. If someone listens to you intently, it makes you feel important. You provide the same feeling to other people by being a good listener. Try to instill this trait in your personality.

Expanding Your Interest: It is always good for the mental health to develop your interests. It will keep the mind fresh, and it will help in cultivating the new interests. It will also make you more attractive to others as you will always have new things to share and talk about.

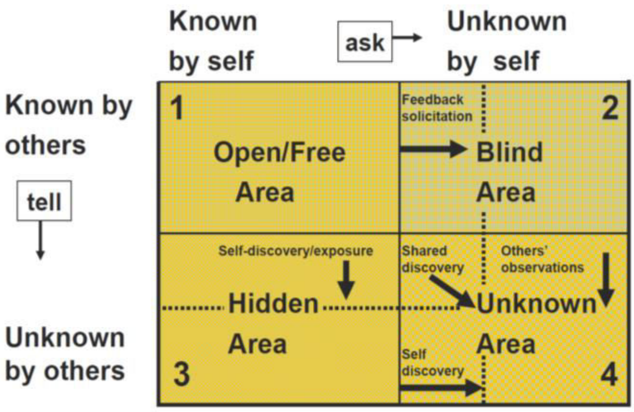
Meeting New People: Always try to meet new people. Engaging with people provides you with a lot of experience, and you can learn so much from different people.

Polishing Interpersonal Skills: Interpersonal skills are the core competencies for a successful life. These skills are useful for communicating and interacting daily. These skills are needed not just for individual interactions, but also in groups. You can work on developing good interpersonal skills. It is not just important in professional life; it is also vital for personal lives. These skills include listening, building connections and persuasion.

Developing Leadership Skills: You may have heard the saying that leaders are born, but it is not entirely accurate. If you want to achieve success in your professional life, it is important that you seek to develop and polish your leadership skills. Employees always look forward to improving them because it includes dealing with people and motivating them.

Presentation Skills: If you want to make sure that you can get your message across to other people, you need to have effective presentation skills. These skills are required in almost all fields. You need to have all the necessary skills like speaking and creativity to make the best use of your presentation skills.

Increasing open area in personality: By constant feedback from others and talking freely about yourself in group. The following figure illustrates how to increase open self.



Exercise on Johari Window

How self-aware are you?...

This questionnaire contains 20 pairs of statements which give different views on certain things. Using a five-point rating scale, circle the number that most closely reflects your view depending on the extent to which you can relate to the opposing statements. Please try not to 'sit on the fence', unless you genuinely feel you do not lean either way.

1	I find the comments of others helpful in learning how to do things	1	2	3	4	5	Most times I can learn to do things for myself	F	E
2	I usually keep my views to myself if I disagree with someone	1	2	3	4	5	I usually tell someone if I disagree with them	F	E
3	It's up to my manager to tell me how I'm doing	1	2	3	4	5	From time to time I ask my manager how I'm doing	F	E
4	When I don something new, I judge myself by my own standards	1	2	3	4	5	When I do something new, I like to be told how I've done	F	E
5	In team meetings I seek the views of others	1	2	3	4	5	In team meetings it is up to others to state their	F	E

6	I tend to control me behaviour when my teams are around	1	2	3	4	5	I tend to behave quite naturally in the presence of my team	F	E
7	I am interested in what others think of me	1	2	3	4	5	Other people's views are their concern	F	E
8	I tend to speak up for my view	1	2	3	4	5	I tend to listen to others	F	E
9	I like to seek the reactions of others to my work	1	2	3	4	5	At the end of the day, I'm paid to come up with my own ideas	F	E
10	I am generally quite self-sufficient	1	2	3	4	5	I like to know where I stand with others	F	E
11	Colleagues usually know where they stand with me	1	2	3	4	5	Sometimes my colleagues are uncertain about my position	F	E
12	Generally, I find it informative to hear what others say about me	1	2	3	4	5	I dislike to hear what others think of me	F	E
13	In relationships, I keep my feelings to myself	1	2	3	4	5	In relationships, I make my feelings known	F	E
14	Sometimes I openly express anger	1	2	3	4	5	People rarely see my angry	F	E
15	I tend to keep my shortcomings to myself	1	2	3	4	5	I sometimes talk to others about areas where I could improve	F	E

16	When a friend seeks my views, I usually give them	1	2	3	4	5	I am usually cautious about being too open, even with a friend	F	E
17	In relationships, there are some things people should keep to themselves	1	2	3	4	5	In relationships, it is best to be open and honest about everything	F	E
18	In a work group, I will disagree even if I oppose the majority	1	2	3	4	5	In a work group, I rarely oppose the common view	F	E
19	When writing something, I prefer to put my ideas down on paper first	1	2	3	4	5	When writing something, I prefer to bounce my ideas off someone	F	E
20	People I work with do not know my views on most things	1	2	3	4	5	People I work with know where I stand on most things	F	E

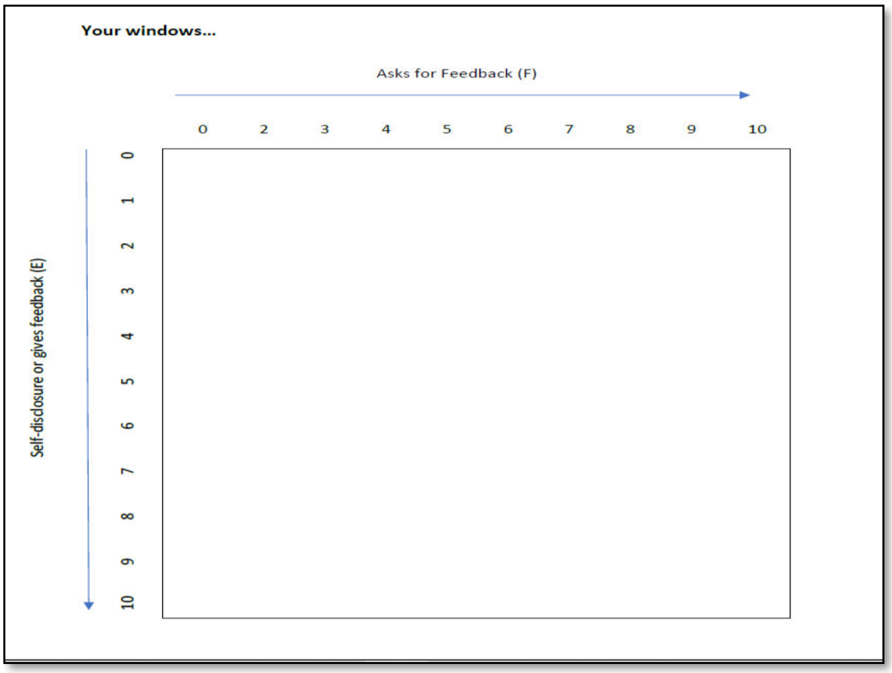
Scoring your Johari Window questionnaire

Please mark your questionnaire using the instructions below:

- If you have marked either a 1 or 2 on questions 1, 5, 7, 9 or 12, circle the F in the right-hand column
- If you have marked either 4 or 5 on questions 3, 4, 10, 17 or 19, circle the F in the right-hand column
- If you have marked either 1 or 2 on questions 8, 11, 14, 16 or 18, circle the E in the right-hand column
- If you have marked either a 4 or 5 on questions 2, 6, 13, 15 or 20 circle the E in the right hand column

- Now add up the number of Fs and Es you have circled. They should each total between 0-10

- Next plot your scores on the feedback model below. Draw a continuous vertical line down from your F score and a continuous horizontal line across from your E score, so that the model is divided into 4 quadrants. This gives and indication of your own 'windows'.



Introduction to spatial data analysis and its application in Social Science research

Sindhuja V Reddy, Alka Singh & Praveen K V

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction:

Despite significant progress achieved in the last two decades, global hunger and malnutrition remain big challenges. The number of undernourished people worldwide increased from 777 million to an estimated 815 million in 2016 (Mamun & Taylor, 2019). The problem of malnutrition is complex and persistent and continues to be one of the primary global health challenges affecting millions of people across countries. Sustainable Development Goal 2 focuses on preventing all forms of malnutrition by 2030. Child and maternal malnutrition are the leading risk factors for the burden of health problems in India (Sahu *et al.,* 2015; Simon *et al*., 2017).

The prevalence of stunting, wasting, and underweight among children during NFHS-5 is 36 percent, 19 percent, and 32 percent, respectively. Children's nutrition indicators have shown slight improvement when compared from NFHS-4 to NFHS-5 (Figure 1). India has not been successful to a large extent in the reduction of malnutrition (Joshi *et al*., 2020). Malnutrition in Indian children still remains a cause of concern. There are also large disparities at the sub-national and districts level in the nutritional and health status of children in India. Hence, there is necessary to understand the disease burden spatially in order to improve the nutrition indicators of the children.

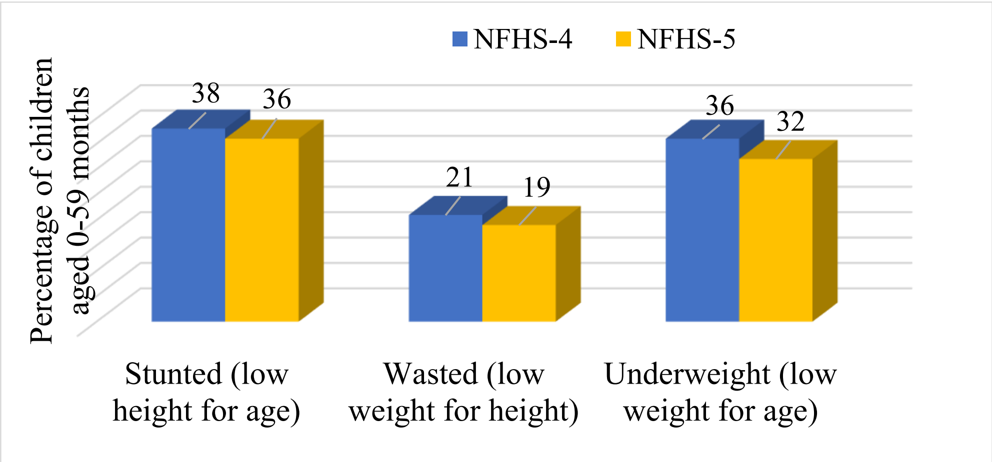


Figure 1: Trends in the nutritional status of children (under age five) in India

This chapter deal with how to calculate stunting among children under age five in India using the unit-level data of NFHS-5 in Stata. And also, how to perform spatial analysis of stunting among children under age five in India using the unit-level data of NFHS-5 in ArcGIS.

Data Source:

National Family Health Survey:

It is a large-scale, multi-round survey conducted on a representative sample of households throughout India (http://rchiips.org/nfhs/). The sample design followed by National Family Health Survey was a stratified, two-stage cluster design. NFHS-5 gathered information from 636,699 households, 724,115 women, and 101,839 men. It provides estimates for 707 districts, 28 states, eight union territories, and 30,198 PSUs in India, making it the largest in the world. Unit-level data of NFHS can be accessed on the DHS website (https://dhsprogam.com).

Steps to analyse DHS data:

- Select surveys for analysis
- Review questionnaires
- Register for the dataset access
- Download the datasets
- Open your datasets
- Get to know your variables
- Use sample weights

Spatial Analysis:

The geographical approach toward examining diseases, or spatial health research, primarily focuses on mapping diseases. (Douven & Scholten, 1995). Disease maps are used to highlight geographic areas with high and low prevalence, the incidence of a specific disease, and the variability of such rates over a spatial domain (Banerjeem, 2016). Spatial analysis is defined as the process of studying entities by examining, assessing, evaluating, and modeling spatial data features that reveal the geometric or geographic properties of data. Disease cluster analyses, spatial comparisons, and mapping remain important methods for disease prediction, prevention, and control (Anasuya Datta, 2020).

Spatial Distribution:

The spatial distribution of diseases refers to the geographical pattern in which diseases are found, indicating their prevalence and concentration across different regions or areas. A graphical display of a spatial distribution may summarize raw data directly or may reflect the outcome of a more sophisticated data analysis. e. g. It is used to find the spatial distribution of disease indicators of children (stunting) in India.

Spatial Autocorrelation:

Based on Tobler's law of Geography, "Everything is related to everything else, but near things are more related than distant things." Spatial autocorrelation analysis use to evaluate whether the spatial distribution of malnutrition indicators among children in India is random or not. Spatial autocorrelation not only clusters similar objects with other similar objects but also speaks about the degree of correlation or similarity. It is helpful in finding hidden patterns and relations. Moran's I is a spatial statistic used to measure spatial autocorrelation by taking the entire data set and producing a single value that ranges from –1 to + 1. Positive values indicate positive spatial autocorrelation (clustering). Negative values indicate negative spatial autocorrelation (dispersion). Values close to 0 indicate random spatial distribution.

Hot spot analysis:

Hot spot analysis (spatial clustering analysis) is a spatial analysis and mapping technique interested in the identification of significant clustering of spatial phenomena. These spatial phenomena are depicted as points in a map and refer to locations of events or objects. The hot spot analysis tool calculates the Getis-Ord Gi* statistic for each feature in a dataset. This tool works by looking at each feature within the context of neighboring features. The analysis result in a high GI* value means hot spot areas (high proportion of malnutrition) and a low GI* value means cold spot areas (low proportion of malnutrition).

Spatial Interpolation:

The spatial interpolation technique is used to predict the un-sampled/unmeasured value from sampled measurements. Ordinary Kriging prediction methods were used for this study to predict disease in unobserved areas of India.

Conclusion:

The spatial analysis results will highlight the most affected areas, the incidence of a specific disease, and the variability of such rates over a geographical domain. Public health interventions should be designed with a targeted approach to impact high-risk populations, which are vital to reducing malnutrition in most affected geographical regions of the country. So, it will help to build up strategies and localized interventions.

Qualitative Response Regression Models

Asha Devi S S, Praveen K V, Renjini V R & Chiranjit M

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction

In classical linear regression models, it is assumed implicitly that the dependent variable Y is quantitative whereas the explanatory variables are either quantitative or qualitative. But, many times the researcher may encounter with the dependent or response variables which are dichotomous or binary in nature, taking a 1 or 0 value or polychotomous or multiple categories. Suppose the research question is to find out the determinants of farmers' participation in farmer producer organizations (FPOs). Here, the dependent variable, is if the farmer is a member of FPO or not, which can take only two values; 1 if the person is a member of FPO and 0 if he or she is not. This dichotomous dependent variable can be modelled as a function of age, education, income, access to information etc., There are several examples where the dependent variable is dichotomous. For example, if someone wants to study to find why certain farmers are adopting a technology/ particular variety and some are not. Here also, the response variable can take only two values,1 for those who are adopting and 0 for non-adopters. Similarly, other examples can be ownership of a machinery, a family enroll for a development scheme or it does not, choice of a certain outcome over others, etc. A unique feature of all these examples is that the dependent variable is of the type which elicits a yes or no response. There are special estimation / inference problems associated with such models. The most commonly used approaches to estimating such models are the Linear Probability model, the logit model and the probit model.

1. The linear probability model

Consider the following regression model $Yi = \alpha + \delta Xi + \varepsilon h$ where Y takes value 1 if a farmer use own seed, and 0 otherwise; Xi includes farmer and other characteristics, γ and δ are vectors of parameters to be estimated, and ε is an error term.

The conditional expectation of the model E(Yi | Xi) can be interpreted as the conditional probability of Y.i.e.,

$$E(Yi \mid Xi) = \beta 1 + \beta 2Xi = Pi$$

There are certain problems associated with the estimation of Linear Probability Models such as: i) Non-normality of the Disturbances (U's)

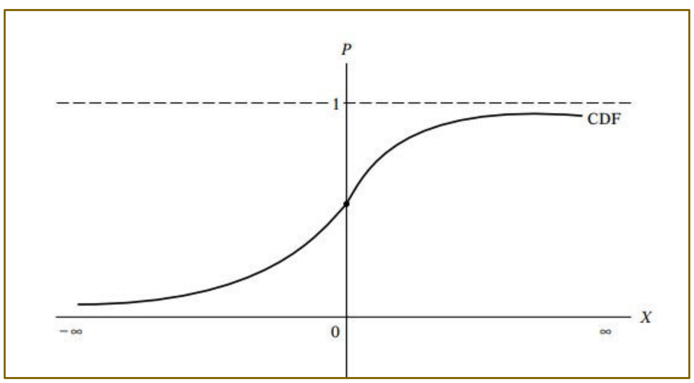
ii) Heteroscedastic variances of the disturbances

iii) Non-fulfillment of $0 \leq E(Y \mid X) \leq 1$ (Possibility of $\hat{Y}$ lying outside the 0-1 range)

iv) Questionable value of R2 as a measure of goodness of fit

The real problem with the OLS estimation of the LPM is that it assumes that Pi = E (Y =1| X) increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout. This seems sometimes very unrealistic. Therefore, there is a need of a probability model that has two features: (1) as X increases, Pi= E(Y = 1|X) increases but never steps outside the 0-1 interval, and (2) the relationship between Pi and Xi is non-linear, that is, approaches "one" which approaches zero at slower and slower rates as Xi gets small and approaches one at slower and slower rates as X gets very large as in Fig 1.

Fig 1. Cumulative Distribution Function



2. The Logit Model

Logit regression (logit) analysis solves the fundamental problem with the LPM and it allows the estimation of the probability that an event occurs or not, by modelling the probability as a non-linear function of the regressors and predicting the binary dependent outcome from a set of independent variables. In an example of seed source use where the dependent variable is farmers using own seed or not in relation to farm size, the linear probability Model depicted it as Pi = E(Y

= 1|Xi) = β1 + β2 Xi where X is the farm size and Y =1 means that the farmer using farm saved seeds.

Let us consider Pi =E (Y=1|Xi)= 1 /1+exp[−(β1 +β2 Xi)] = 1 /1+exp[−(Zi)] , where Zi = β1 + β2 Xi

This equation is known as the (cumulative) logistic distribution function. Here Zi ranges from - ∞ to + ∞; Pi ranges between 0 and 1; Pi is non-linearly related to Zi (i.e. Xi) thus satisfying the two conditions required for a probability model. In satisfying these requirements, an estimation problem has been created because Pi is nonlinear not only in X but also in the β's. This means that one cannot use OLS procedure to estimate the parameters.

Here, Pi is the probability of using farm saved seed and is given by 1/ 1+exp [−(Zi)] and (1- Pi), is the probability of using other source's seed, is (1-Pi) = 1 /1+exp[(Zi)] . Therefore, one can write Pi /1−Pi = 1+exp[−(Zi)]/ 1+exp[(Zi)]. Here, Pi/1-Pi is the odds ratio in favour of using farm saved seeds i.e; the ratio of the probability that a farmer will use farm saved seed to the probability that he uses some other source's seed.

Taking natural log of the previous equation, we obtain Li = ln [Pi / (1 - Pi)] = Zi = β1 + β2 Xi .That is, the log of the odds ratio is not only linear in X, but also linear in the parameters. L is called the Logit. If the data are available in grouped form, we can use OLS to estimate the parameters of the logit model, provided we take into account explicitly the heteroscedastic nature of the error term. If the data are available at the individual, or micro, level, nonlinear-in-the-parameter estimating procedures are called for.

Features of the Logit Model

1. As P goes from 0 to 1, the logit L goes from - ∞ to + ∞. That is, although the probabilities lie between 0 and 1, the logits are not so bounded.
2. Although L is linear in X, the probabilities themselves are not.
3. The interpretation of the logit model is as follows: β2 , the slope, measures the change in L for a unit change in X, i.e it tells how the log odds in favour of using farm saved seed change as farm size increase by a unit. The intercept β1 is the value of the log odds in favour of using farm saved seed when farm size is zero.
4. Given a certain land size, say X*, if we actually want to estimate not the odds in favour of farm-saved seed but the probability of using farm saved seed itself, this can be done directly (1) once the estimates of β1 and β2 are available.

5. The linear probability model assumes that Pi is linearly related to Xi, the logit model assumes that the log of odds ratio is linearly related to Xi.

Merits of Logit Model

1. Logit analysis produces statistically sound results. By allowing for the transformation of a dichotomous dependent variable to a continuous variable ranging from -∞ to + ∞, the problem of out-of-range estimates is avoided.
2. The logit analysis provides results which can be easily interpreted and the method is simple to analyse.
3. It gives parameter estimates which are asymptotically consistent, efficient and normal, so that the analogue of the regression t-test can be applied.

Demerits of Logit Model

1. As in the case of Linear Probability Model, the disturbance term in logit model is heteroscedastic and therefore, we should go for Weighted Least Squares.
2. The number of observations has to be fairly large for all Xi and hence in small sample; the estimated results should be interpreted carefully.
3. As in any other regression, there may be problem of multicollinearity if the explanatory variables are related among themselves.
4. As in Linear Probability Models, the conventionally measured R^2 is of limited value to judge the goodness of fit.

3. The Probit Model

In order to explain the behavior of a dichotomous dependent variable we have to use a suitably chosen Cumulative Distribution Function (CDF). The logit model uses the cumulative logistic distribution function. But this is not the only CDF that one can use. In some applications, the normal CDF has been found useful. The estimating model that emerges from the normal CDF is known as the Probit Model. Let us assume that in seed source use example, the decision of the i th farmer to use farm saved seed or not depends on unobservable utility index Ii, that is determined by the explanatory variables in such a way that the larger the value of index Ii , the greater the probability of the farmer using own seed.
The index Ii can be expressed as Ii =β1 +β2 Xi, where Xi is the farm size of the i^{th} household.

Assumption of Probit Model

For each household there is a critical or threshold level of the index (Ii*), such that if Ii exceeds Ii*, the family will use farm saved seed, otherwise not. But the threshold level Ii* is also not observable. If it is assumed that it is normally distributed with the same mean and variance, it is possible to estimate the parameters of the equation and thus get some information about the unobservable index itself. In Probit Analysis, the unobservable utility index (Ii) is known as normal equivalent deviate (n.e.d) or simply Normit. Since n.e.d. or Ii will be negative whenever Pi < 0.5, in practice the number 5 is added to the n.e.d. and the result so obtained is called the Probit i.e; Probit = n.e.d + 5 = Ii + 5 In order to estimate β1 and β2, the previous equation can be written as Ii =β1 +β2 Xi+ Ui

Steps Involved in Estimation of Probit Model

1. Estimate Pi from grouped data as in the case of Logit Model, i.e. $Pi\hat{} = ni\ Ni$
2. Using $Pi\hat{}$ obtain n.e.d (Ii) from the standard normal CDF, i.e. Ii=β1 +β2 Xi,
3. Add 5 to the estimated Ii to convert them into probits and use the probits thus obtained as the dependent variable
4. As in the case of Linear Probability Model and Logit Models, the disturbance term is heteroscedastic in Probit Model also. In order to get efficient estimates, one has to transform the model.
5. After transformation, estimate the final equation by OLS.

Choice between Logit and Probit Models

The chief difference between logit and probit is that logistic has slightly flatter tails i.e; the normal or probit curve approaches the axes more quickly than the logistic curve. Qualitatively, Logit and Probit Models give similar results; the estimates of parameters of the two models are not directly comparable. The choice between logit and probit model depends on the ease of computation. Probit model is mathematically a bit difficult as it involves integrals, which is not a serious problem with sophisticated statistical packages that are now readily available. But for all practical purposes, both logit and probit models give similar results.

References

Damodar, N, Gujarati. 2003. Basic econometrics (4th ed.), New York: Mc Graw Hill Publications

Extraction of NSSO data and its analysis in Stata: A case of non-farm income of agricultural households

Utkarsh Tiwari, Jagadeesh M. S. & Chiranjit Majumder
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute

Introduction:

The National Sample Surveys (NSS) are being conducted by the Government of India since 1950 to collect socio-economic data employing scientific sampling methods. These surveys provide invaluable insights into a wide range of domains related to living conditions, economic activities, and social characteristics. Handling & analyzing such big data is a tough task, but statistical software like Stata, SPSS, R, etc makes its handling easy and can help researchers, policymakers, and analysts make informed decisions and understand the changing dynamics of the Indian society and economy. Here we will deal with the NSSO 77th round dataset which covers the subjects 'Land and Livestock Holdings of Households and Situation Assessment of Agricultural Households' and 'Debt and Investment'.

This chapter will guide you through the process of analyzing NSSO data using Stata software. We will cover the steps involved in data extraction, renaming, data reshaping, data merging and analysis, focusing on the household level income of agricultural households and its component at all India level from the dataset of NSSO 77th-round survey. By the end of this chapter, you will have a solid understanding of how to harness the power of Stata to extract meaningful insights from NSSO data.

Data Extraction

For data extraction, we need to have 3 data files i.e.

i) Raw data files/ Text files
ii) Layout files
iii) Readme files (for checking the number of observations that we extracted, whether they are correct or not)

In Stata software, the NSSO data of 'Land and Livestock Holdings of Households and Situation Assessment of Agricultural Households' and 'Debt and Investment' will be extracted data of level 1 of visit-1 by using the command

infix str centre_code 1-3 str fsu 4-8 round 9-10 str schedule 11-13 sample1 14 sector 15 str nss_region 16-18 district 19-20 stratum 21-22 sub_stratum 23-24 str sub_round 25 str fod_sub_region 26-29 str sss 30 str hh_no 31-32 str visit_no 33 level1 34-35 filler 36-40 informant_sl_no 41-42 response_code 43 survey_code 44 substi_reason 45 str employee_code1 46-49 str employee_code2 50-53 str employee_code3 54-57 str survey_date 58-63 str dispatch_date 64-69 canvas_time 70-72 str investigator_no 73 rem_b_17a 74 rem_b_17b 75 remarks1 76 remarks2 77 blank 78-126 nsc 127-129 multiplier 130-139 using "path of the text file"

Similarly, by using the infix command, we can extract data for other levels of visit-1 as well as visit-2

Renaming of data:

After extraction, the second task is to rename the variable's name, as in the extracted dataset it is in the form of codes like (a1b1, a2b1, a3b1, etc.). So according to your convenience rename the variable name. Remember one thing here is that in Stata we cannot give space in between the name of variables eg. ~~centre code~~ is wrong, but centre_code is right. We should have use the underscore (_) sign while separating the two words. Command for renaming is

rename 'variable name' 'new variable name'

Generation of common ID:

After renaming generate a common or HHID for each of the observations by clubbing some common features together like here we club the details of fsu, sss, hh_no & visit_no. Stata command for this is

egen HHID=concat(fsu sss hh_no visit_no)

Generate weights:

After the generation of common id or HHID, generates weights for all the observations for this exercise we generate weights by using the command

gen weight= multiplier/100

Reshaping of data:

Reshaping data in Stata refers to the process of transforming data from a "long" format to a "wide" format or vice versa. This is often necessary when you have data organized in a way that doesn't suit your analysis needs, and you want to restructure it for better manipulation and analysis. Reshaping can be done using the **reshape** command in Stata. Command for reshaping NSSO 77th round data is

reshape wide variable - variable, ith variable jth variable

eg. **reshape wide** Crop_code- inputs_Expences_Rs, i(HHID) j(serial_number)string

As the NSSO 77th round situation assessment data is household-level data, therefore, there is one to more than one member in a household. So, in this dataset, they assigned serial numbers to different members of a household. This means they collected data from each member of the household. So, there is more than one entry for a household for the same variable (like non-farm income), non-farm income of household 'A' (if there are 3 members in household 'A' who are having non-farm income). Therefore, entry should be like non-farm income of A1, A2, and A3 for a single household, this is called the long format of data, therefore the number of observations is very high as there is repeated entry for the same household for different members. So, to make it easily understandable form the long form of data is converted into the wide form of data by reshaping. The data entry for different members of a household comes under a single unique household id or HHID i.e., non-farm income of HHID 'A' = 1+2+3. By reshaping the data in wide form there will be no repetition of the same household and all the entries will come under a common household id or HHID. The number of observations will be drastically reduced by doing this. For reshaping we have to critically identify the ith & jth variables ith variable is the common id we created, jth variable here is the serial numbers of the household members. Jth variable is that variable for which we want to reshape our data here we want to reshape our data for household members as this is repeated for the common households many times. Likewise, we have to calculate the income of a household from different crops, so here our jth variable will be the number of crops.

Additionally, it's important to ensure that your data is sorted appropriately before using the reshape command. For example, if you're reshaping long to wide, your data should be sorted by the identifier and the variable representing the time periods/categories. Reshaping data can be complex depending on your specific dataset and needs. Always make sure to back up your data before performing any data transformations and thoroughly test the results to ensure the reshaping was done correctly.

Merging:

After proper reshaping, we have to merge the two datasets in order to make available the data at one place, which is otherwise spreading in different levels or visits. For this in the Stata window we have to go into the 'Data' option on the left top of the window, within which we have to go in the option Combine datasets Merge two datasets One-to-one on key variables (One-to-one on key variables because we already reshaped the data in wide format, so there will no repetition of the common ID and since we defined the common ID or HHID for all the observations which will be same for the one household in all the levels as well as in both the visits). If we want to merge the long-format datasets, then we have to go for Many-to-many on key variables, if we already created the common ID or HHID for the datasets.

Analysis

After merging different levels data of visit 1 & visit 2, as well as visit 1 with visit 2 we have all the data at one place in a single Stata data file, from here we can start our analysis part. While we have categorical or qualitative data we can go for logit or probit model in Stata. But, if we have quantitative data we can go for simple regression or tobit regression model in which it is assumed that the dependent variable has a number of its values clustered at a limiting value, usually zero. The Tobit regression uses all observations, both those at the limit and those above it, to estimate a regression line, and it is to be preferred, in general, over alternative techniques that estimate a line only with the observations above the limit

Here we are focusing on tobit regression model/censored tobit regression model. Censored tobit regression is used when your dependent variable is having zero values for some or several observations. Example if we are calculating the determinants of non-farm income of agricultural households. So in this case, there are so many households which are having no non-farm income or zero income from non-farm sources. The general formulation for Tobit specification is given (Greene, 2003) as:

$$y_i^* = x_i'\beta + \varepsilon_i$$
$$y_i = 0 \quad if \; y_i \le 0$$
$$y_i = y_i^* \quad if \; y_i > 0$$

Where,

y_i* = a latent variable,

β = a parameter to be estimated,

x = a vector of explanatory variables and

ε = the error term

In case of tobit regression we are using Yi^* instead of Yi, because our dependent variable is truncated.

Stata command use for tobit regression is

```
tobit dependentvariable independentvariables [aw=weight], ll(0)
```

If independent variables are categorial variable then we have to prefix it with i.

eg. **tobit** Net_nonfarm_income HH_Size Age Education Total_Land_Acre i.Gender i.Religion i.Posses_KCC i.MNGREGA_Card [aw=weight], ll(0)

Here Net_nonfarm_income is our dependent variable. Gender, religion, KCC possession & MGNREGA card possession is categorial variables

ll(0) means, in our case the lower limit of non-farm income is zero i.e. non-farm income should be more than zero.

By using the above command we will get the outcome table like

Table. **Determinants of Non-Farm Income**

TOTAL NON-FARM INCOME	**Coef.**	**Std. Err.**	**P>t**	**TOTAL NON-FARM INCOME**	**Coef.**	**Std. Err.**	**P>t**
Total Farm Investment	0.033	0.12	0.785	**Social Group (ST = 0)**			
Total Non-Farm Investment*	0.304	0.06	**0**	SC (1) ***	-24580.06	14165.56	**0.083**
Household Size *	15879.8	1876.858	**0**	OBC (2) **	-25972.61	11812.36	**0.028**
Age *	5347.635	346.783	**0**	OTHERS (3) **	-38382.92	16564.11	**0.021**
Total Land Owned	-550.652	893.59	0.538	**Possess KCC = 0**			
Household classification based on major source of household income (Agril & Allied=0)				Do not possess KCC (1) *	31687.11	10141.32	**0.002**
Wage & Salary from Agri (1) *	123532	22264.91	**0**	**Possess MGNREGA Card = 0**			

Wage & Salary from Non-Agri (2) *	227905.4	10804.67	**0**	Do not possess MGNREGA Card (1) *	31091.95	11069.5	**0.005**
Non Agri Enterprises (3) *	214402.7	18065.71	**0**	**Member of FPO = 0**			
Others (4) *	96843.19	30059.91	**0.001**	Not Member of FPO (1) ***	53937.35	32397.01	**0.096**
Gender (Male=0)				_cons	-509300.9	41163.56	0
Female (1) *	-44589.27	18736.77	**0.017**	Var(e.TOTAL_NON FARM_INCOME)	38600000000	1410000000	
Education Group (up to Primary=0)							
Primary to High School (1) *	51929.63	11563.58	**0**				
High School to Higher Secondary or equivalent Diploma (2) *	163800.5	18750.95	**0**				
Graduate & Above (3) *	189924.8	19976.29	**0**				

Interpretation of this,

According to Tobit regression results, the non-farm income of agricultural households is significantly influenced by the factors like, 'total non-farm investment,' 'household size', 'age of household head,' 'major source of household's income,' 'gender of the household's head,' 'education of households head,' 'caste,' 'possession of KCC,' 'possession of MGNREGA card,' and 'membership of FPO'. An increase in non-farm investment of one rupee will result in an increase in non-farm income of 0.3 paise for agricultural households. Similarly, if the size of an agricultural household increases by one member, non-farm income will rise by ₹15880/- per household annually. The non-farm income of a household increases by ₹5348/- per year if the age of the head of the household rises by one year. In comparison to households whose primary source of income is agriculture and allied, those households whose primary source of income is farm wages and salaries, non-farm wages and salaries, income from the non-farm business, or having other sources of income, have higher non-farm income by ₹123532/-, ₹ 227905/-, ₹214403/-, and ₹96843/- per annum, respectively. Female-headed households have lower non-farm income i.e., by 44589 ₹ per annum as compared to male headed households and households with heads education levels above primary to high school, high school to higher secondary or equivalent diploma, and above graduation level have higher non-farm income by ₹ 51930/-, ₹ 163801/-, and ₹ 189925/- annually per household, respectively, than those with heads education levels below primary. All households other than S.T.'s households' have lower non-farm income (S.C.'s households' are lower by ₹ 24580/-, O.B.C.'s households' are lower by ₹ 25973/-, and Other households' are lower by ₹ 38383/- per annum) and in comparison to households in which any of the members own KCCs, households in which none of the members owns KCCs have higher non-farm income by ₹31687/- per annum. The non-farm income of households with no MGNREGA card holders is greater by ₹31092/- per annum compared to households with one or more MGNREGA card holders and in comparison, to households in which any of the members are having FPO membership, households in which none of the members are having FPO membership have higher non-farm income by ₹53937/- per annum.

Training Manual

Analytical Techniques for Empowering Social Science Research

August 7-18, 2023

Division of Agricultural Economics, ICAR-IARI, New Delhi

Course Director

Alka Singh
Professor and Head
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
Pusa Campus, Delhi – 110012
Email: asingh.eco@gmail.com

Coordinators

Praveen K V
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail: veenkv@gmail.com

Renjini V R
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail:renji608@gmail.com

Chiranjit Mazumder
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
E-mail: majumder.chira@gmail.com

Published By

NAHEP- Centre for Advanced Agriculture Science and Technology
ICAR- Indian Agriculture Research Institute, New Delhi
Website: **www.nahep-caast.iari.res.in**