

Quantitative Methods for SOCIAL SCIENCES

Edited by: Vinayak Nikam • Abimanyu Jhajhria • Suresh Pal

This reference book is designed keeping in mind the need for the application of advanced quantitative methods in social science research to enhance its accuracy. The chapters are written in such a way that social scientists can easily grasp the methods including their theoretical and practical aspects using statistical software. The book provides comprehensive coverage of multivariate techniques, forecasting methods, structural equations, optimization models, quantitative methods for impact assessment, growth models and other important methods used in social science research.

ABOUT THE EDITORS

Vinayak Nikam is working as Scientist (Senior Scale) at ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He did his PhD from ICAR-Indian Agricultural Research Institute (New Delhi) in agricultural extension. He also served ICAR-Central Soil Salinity Research Institute (Karnal) before joining NIAP in Nov 2015. He also holds faculty membership in the discipline of Agricultural Extension at ICAR-Indian Agriculture Research Institute (New Delhi). Currently, he is working on the performance and impact assessment of Agriculture Extension and Advisory Services as well as associated with technology impact assessment.

Abimanyu Jhajhria is a Scientist (Agricultural Economics) at ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He has received his doctorate in Agricultural Economics from the ICAR-Indian Agricultural Research Institute (New Delhi). His research interest includes markets, trade and institutions. He is involved in the projects on market reforms, agricultural commodity value chains and outlook models for agricultural commodities.

Suresh Pal is Director of ICAR-National Institute of Agricultural Economics and Policy Research (New Delhi). He has a PhD in agricultural economics from ICAR-Indian Agricultural Research Institute (New Delhi). He has published extensively on different aspects of Indian agriculture and guided doctoral and masters students. He has received awards for his contributions like best journal article awards, Norman E Borlaug International Science Fellowship, Fellow of the Indian Society of Agricultural Economics (Mumbai), and Fellow of the National Academy of Agricultural Sciences.



ICAR-NATIONAL INSTITUTE OF AGRICULTURAL ECONOMICS AND POLICY RESEARCH (NIAP)

D.P.S. Marg, Pusa, New Delhi - 110012. Phone: +91-11- 25847628, +91-11- 25846731
Fax: +91-11-25842684. E-mail: director.niap@icar.gov.in; Website: www.niap.res.in



Quantitative Methods for Social Sciences

Vinayak Nikam • Abimanyu Jhajhria • Suresh Pal

Vinayak Nikam • Abimanyu Jhajhria • Suresh Pal

Quantitative Methods for SOCIAL SCIENCES



ICAR-National Institute of Agricultural Economics and Policy Research

Reference book on

Quantitative Methods for

SOCIAL SCIENCES

Edited by

Vinayak Nikam
Abimanyu Jhahria
Suresh Pal



ICAR-National Institute of Agricultural Economics and Policy Research
New Delhi

Quantitative Methods for SOCIAL SCIENCES

Edited by
Vinayak Nikam, Abimanyu Jhajhria and Suresh Pal

© 2019 ICAR-National Institute of Agricultural Economics and Policy Research

Published by
Dr. Suresh Pal
Director, ICAR-National Institute of Agricultural Economics and Policy Research

ISBN: 978-81-940080-2-6

Printed at
Chandu Press: chandupress@gmail.com

Preface

The effectiveness of agricultural policies and programs is based on their robust design and timely implementation, which in turn, requires inputs based on a realistic assessment of farming systems and their empirical analysis. The disciplines of social science are better placed to conduct such an analysis. Some times greater emphasis is given on qualitative aspects and historical trends, neglecting the quantitative and futuristic analysis for suggesting the appropriate policy. For quality research output based on empirical analysis, one should employ a systematic approach and appropriate quantitative method and collate required data. This reference book on quantitative methods for social sciences is written to address practical needs for social science research to study various aspects of agriculture and farmers' behaviour.

The book is divided into six sections and contains 36 chapters on various quantitative methods. All the chapters are designed in a way that social science professionals can easily grasp the methods, including their theoretical as well as practical aspects. The volume begins with multivariate analysis and builds on the structural analytical framework for better understanding of the sector and analysis of various interventions. This is followed by the discussion on time-series analysis and forecasting methods for projecting likely behavior of variables like prices. Optimization methods have been of considerable interest to social scientists and these are explained using some real examples of the optimization problems. Most of the methods have been illustrated with examples and steps in estimation have been indicated along with software syntax.

This volume is different from other reference books on this topic. The quantitative methods covered in this book are based on actual research and have been contributed by the authors who are using these methods for their research work. It also covers both theory and practical application of different methods with illustrations using computer software. The book provides comprehensive coverage of multivariate techniques, forecasting methods, impact assessment methodologies, growth analysis and other important methods used in social science research. I take this opportunity to sincerely acknowledge the contributions of all the authors in the preparation of this book. I hope this volume will be useful for practitioners of social science research and valuable teaching material. Suggestions of the readers will be useful to improve further publications on this topic.

Suresh Pal
Director

Acknowledgments

This book is an outcome of the efforts of a large number of researchers and we acknowledge with all humility the help and support provided by them. First and foremost, we would like to express our sincere thanks to all the authors of the book chapters for their contributions within the stipulated time. The publication committee of the Institute has done internal review of the book and provided useful suggestions and comments to improve the quality of the book. The editors are grateful to each one of them. We are grateful to Ms. Aruna T Kumar for rigorous editing the book and useful suggestions to improve the presentation. Dr. Sapna Panwar has put great efforts in reading the book and her contribution is acknowledged with thanks. Summer School sanctioned by Indian Council of Agricultural Research (ICAR) was the trigger to start working on this reference book. We sincerely thank ICAR for financing the training program.

We take this opportunity to express our grateful thanks to our colleagues Subash S P and Balaji S J who have helped in fine-tuning some of the chapters. Special thanks are due to Ranjith P C who has worked tirelessly with us during the process of compiling and editing of different chapters. We express our sincere thanks to all the colleagues who have directly or indirectly contributed to the publication of this book.

**Vinayak Nikam
Abimanyu Jhajhria
Suresh Pal**

CONTENTS

<i>Preface</i>	iii
<i>Acknowledgements</i>	v
1. Overview of quantitative methods for social science research <i>Vinayak Nikam, Abimanyu Jhajhria and Suresh Pal</i>	1
Part I: Measures of interdependence of variables/cases	
2. Cluster analysis <i>Arpan Bhowmik, Sukanta Dash, Seema Jaggi and Sujit Sarkar</i>	7
3. Principal component analysis <i>Prem Chand, M. S. Raman and Vinita Kanwal</i>	19
4. Multidimensional scaling <i>Ramasubramanian V.</i>	31
5. Correspondence analysis <i>Deepak Singh, Raju Kumar, Ankur Biswas, R. S. Shekhawat and Abimanyu Jhajhria</i>	46
Part II: Regression analysis	
6. Linear and non-linear regression analysis <i>Ranjit Kumar Paul and L. M. Bhar</i>	59
7. Qualitative regression model (Logit, Probit, Tobit) <i>Shivaswamy G. P., K. N. Singh and Anuja A. R.</i>	70
8. Introduction to panel data regression models Ravindra Singh Shekhawat, K. N. Singh, Achal Lama and Bishal Gurung	78
9. Auto regressive and distributed lag models <i>Rajesh T., Harish Kumar H. V., Anuja A. R. and Shivaswamy G. P.</i>	88
10. Conjoint analysis <i>Sukanta Dash, Krishan Lal and Rajender Parsad</i>	96
11. Two stage least square simultaneous equation model <i>Shivendra Kumar Srivastava and Jaspal Singh</i>	110

12. Discriminant function analysis 121
Achal Lama, K. N. Singh, R. S. Shekhawat, Kanchan Sinha and Bishal Gurung

Part III: Time series analysis

13. Price forecasting using ARIMA model 129
Raka Saxena, Ranjit Kumar Paul and Rohit Kumar
14. Volatility models 142
Girish Kumar Jha and Achal Lama
15. Artificial neural network for time series modelling 155
Mrinmoy Ray, K. N. Singh, Kanchan Sinha and Shivaswamy G. P.
16. Hybrid time series models 163
Ranjeet Kumar Paul

Part IV: Impact assessment methods

17. Economic surplus approach 177
Vinayak Nikam, Jaiprakash Bishen, T. K. Immanuelraj, Shiv Kumar and Abimanyu Jhahhria
18. Introduction to causal inference 192
Arathy Ashok
19. Propensity score matching 199
K. S. Aditya and Subash S. P.
20. Difference-in-difference model 211
M. Balasubramanian and Gourav Kumar Vani
21. Regression discontinuity design 219
Subash S. P. and Aditya K. S.
22. Synthetic control method 230
Prabhat Kishore
23. Instrumental variable estimation 236
Anuja A. R., K. N. Singh, Shivaswamy G. P., Rajesh T. and Harish Kumar H. V.

Part V: Growth analysis

24. Computable general equilibrium models 245
Balaji S. J.

25.	Decomposition of total factor productivity: DEA approach <i>Dharam Raj Singh, Suresh Kumar, Venkatesh P. and Philip Kuriachen</i>	254
26.	Total factor productivity using stochastic production function <i>Shiv Kumar, Abdulla and Deepak Singh</i>	264
Part VI: Other methods		
27.	Linear programming <i>Harish Kumar H. V., Rajesh T., Shivaswamy G. P. and Anuja A. R.</i>	277
28.	Multi objective programming <i>Chandra Sen</i>	289
29.	Structural equation modelling <i>P. Sethuraman Sivakumar, N. Sivaramane and P. Adhiguru</i>	296
30.	Partial equilibrium model <i>Shinoj Parappurathu</i>	310
31.	Production function analysis <i>Suresh Kumar, Dharam Raj Singh and Girish Kumar Jha</i>	319
32.	Social network analysis <i>Subash S. P.</i>	335
33.	Construction of composite index <i>Prem Chand</i>	351
34.	Basic scaling techniques in social sciences <i>Sudipta Paul</i>	361
35.	Analytical hierarchy process: A multi-criteria decision making technique <i>Anirban Mukherjee, Mrinmoy Ray and Kumari Shubha</i>	371
36.	Artificial intelligence, machine learning and big data <i>Rajni Jain, Shabana Begam, Sapna Nigam and Vaijunath</i>	378
	List of contributors	395

Chapter 1

OVERVIEW OF QUANTITATIVE METHODS FOR SOCIAL SCIENCE RESEARCH

Vinayak Nikam, Abimanyu Jhajhria and Suresh Pal

Indian agriculture has witnessed large number of structural and institutional changes recently. Research institutions from National Agricultural Research System (NARS) and other government organizations (National Sample Survey Organisation (NSSO), various ministries etc.) are generating large number of data on agriculture and allied activities. Technological advancement in terms of satellite (remote sensing, geosynchronous) is also providing large volumes of data. ICT revolution has helped dissemination, gaining access and storage of large number of data. Therefore, decision making and policy formulation under these conditions demands that the recommendations are based on systematic, scientific and empirical research accompanied by rigorous quantitative methods. As large number of quantitative information is available and accessible, it is apparent and imperative that researchers and social scientists need appropriate working knowledge of procedures and techniques for analyzing the data. Therefore, this book is an attempt to acquaint the social scientists with basic and advanced quantitative methods.

Scientific knowledge in social science can be generated through logic (theory) and evidence (observations). For this we use both inductive (theory building) and deductive (theory testing) approach. Research can be exploratory (new area of enquiry), descriptive (careful observation and documentation of phenomenon) and explanatory research (explanation of observed phenomenon). Journey of scientific thoughts was started from the rationalism (around 3rd century BC), where emphasis was on understanding world through systematic logical reasoning. During 16th century, Francis Bacon proposed empiricism where emphasis was put on knowledge acquisition through observations. Natural philosophy combined the elements of empiricist and rationalism. In 18th century, Auguste Comte, brought the idea of positivism where emphasis was on verifications of theory through observations of things; by application of methods and practices of natural science into social science. Thus, it suggested more and more use of quantitative methods like experiment, quasi experiment and survey in social science.

Scientific research is characterized by systematic, step by step and logical approach. It starts with converting questions/situation into clearly stated research problem and the objectives of the research are specified. Formation of hypothesis is most important

step in quantitative research which gives tentative relationship between the variables. It is testable proposition about the possible outcome of the research study. Rigorous review of existing literature related to the topic is done in order to identify the current state of knowledge, gaps in knowledge and methodologies used to study the problem. Research design creates the blueprint for the activities to be undertaken to identify the answer of the research question. Many times limitation of randomization of subject in social science, we have to use the quasi experimental research or non-experimental research designs like survey method, case method, focused group method, action research etc. Collection of data is next important step in conducting research. For primary data collection, there should be sufficient care for the sampling, development of questionnaire, pilot testing and validating the questionnaire before collection of the data. Once data is collected, after data entry in excel or other software, one need to clean it properly to reduce the missing data and outliers. This also needed in case of secondary data. After data cleaning one can start the actual analysis using various methods specified in the book.

When relation between more than two variables is analyzed, it is called as multivariate analysis. Broadly, it is classified into two main categories like methods that measures dependence and interdependence. For measuring the interdependence, most common techniques are the factor analysis, principal component analysis, correspondence analysis, multidimensional scaling, cluster analysis, etc. In these techniques, there is no classification of variables into dependent and independent (explanatory) variables, but these are group of interrelated variables. To analyse the structure of variable, factors analysis is used, which extract the specific number of synthetic variables which are known as latent variables or factors. Cluster analysis is used to group cases or respondents in respect to structure such that object or cases in one cluster are more similar than other. In case of non-metric data, correspondence analysis is used which analyse simple two way or multi way tables containing some measures of correspondence between row and column. Multidimensional scaling is used for non-metric variable which is visual representation of distance or similarities between sets of objects.

Methods measuring dependence vary based on how many variables are predicted. To measure the multiple relationship among dependent and independent variables, structural equation modelling and simultaneous equations are used. Primary variable used in structural equation modelling is usually latent variable while primary variable in simultaneous equation is observed variable. Single relationship between several dependent variables if both dependent and independent variables are in metric, canonical correlation analysis is used, while for non-metric independent variable multi variate analysis of variance is used. In case of one dependent variable, if it is metric, multiple regression and conjoint analysis is used. For non-metric dependent variable, multiple discriminate analysis and linear probability models are used. Dependence can also be measured using the qualitative regression models like Logit, Probit and Tobit.

Forecasting in agriculture can be done with two basic approaches, namely structural modelling techniques and time series models. The structural models require demand and supply schedules and their intersection gives equilibrium price. These models require huge quantity and quality of data which usually are not available in developing countries. Consequently, researchers often rely on time series modelling as these techniques requires less data input for price forecasting. In time series modelling, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. Time-series models can be linear, non-linear and hybrid time series based on the nature and requirement of the data. Autoregressive Integrated Moving Average (ARIMA) model is one of the most important and widely used linear time-series technique for forecasting purpose due to its statistical properties. But one of the major limitations of these models is assumption of linearity which limits the application of ARIMA model to real time-series data. When the assumption of homoscedastic error variance is violated then non-linear time series models like Autoregressive Conditional Heteroskedasticity (ARCH) / Generalised Autoregressive Conditional Heteroscedastic (GARCH) are applied to capture the changes in the conditional variance of data. Artificial Neural Network (ANN) is a data-driven, self-adaptive, non-linear, non-parametric method of forecasting, unlike the traditional model-based methods. ANN models can be used to model many non-linear processes that have unknown functional relationship. Actual time-series data are rarely pure linear or non-linear in nature and sometimes contain both the trend in the data set. In this situation, a hybrid approach of combining the forecasts from a linear time-series model (ARIMA) and from a non-linear time-series model (GARCH, ANN) can give better forecast performance. The empirical evidences show that non-linear models are better for long-term forecasting and the linear models are more suitable for short range forecasting. Therefore, it is better to combine the linear and non-linear models to get accurate forecast.

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 the 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 the 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 the

impact. In this book, we have provided chapters on quasi-experimental methods such as propensity-score matching, difference in difference, regression discontinuity and synthetic control method. Economic surplus approach to measure the impact at macro level is also covered.

Quantitative analysis of development policies is critical for informed decision making. From understanding the contemporary performances in growth, investments, technologies, employment etc. to project future demand for food and processed commodities, measuring the impacts of rising wages, changes in land use, markets and other practices accordingly, quantitative methodologies play an important role. Notable among them are demand system estimation models such as Linear and Quadratic Almost Ideal Demand System (AIDS) models; Total Factor Productivity (TFP) estimation; decomposing them into various components such as technological change and efficiency change; other decomposition techniques that assess the relative role of area expansion, yield improvement and extent of diversification. Relatively complex are the partial and general equilibrium approaches. While the former makes a detailed model of inter-linkages in a particular crop/region/sector, general equilibrium approach models entire economic activities by tracing ‘real’ economic transactions among various agents like households, business enterprises, Government and external economies. They play a powerful role in taxation, trade, climate change and a variety of other policies.

Apart from above mentioned methods, many other methods are being used for analysis of data by social science researchers. Linear programming is used for optimum solution to a given problem; bridges gap between abstract economic theory and policy maker’s decision making in practice. Production function measures the functional relation between quantity of output and quantity of input (factors of production). It gives idea about maximum amount of output that can be obtained from a given number of inputs and indicates the increasing or decreasing returns to scale. Social network analysis is an effective tool in understanding the social relations and interactions of the individuals in the group in a specific social context. It has many applications in agriculture like understanding the adoption patterns, formation and sustenance of farmers producers organizations etc. Analytic hierarchy process is a multi-criteria decision making tool where many variables or criteria are considered in the prioritization and selection of alternatives in complex environments. Artificial intelligence, machine learning and big data analytics are the new tools that are being used in the analysis of large amount of quantitative data generated in agriculture. Scales and indexes are commonly used useful tools in social sciences. Scale helps in turning a series of qualitative facts or attributes into a quantitative series or variables. A composite index is a collection of large number of indicators or variables that are aggregated together to represent overall performance of sectors. Thus, this book is an attempt to cover all important methods and techniques used in the analysis of quantitative data in social science research.

PART I

**MEASURES OF INTERDEPENDENCE
OF
VARIABLES/CASES**

Chapter 2

CLUSTER ANALYSIS

Arpan Bhowmik, Sukanta Dash, Seema Jaggi and Sujit Sarkar

INTRODUCTION

Statistical science plays a major role in any scientific investigation. Use of appropriate statistical techniques for analyzing the data is very crucial to obtain a meaningful interpretation of the investigation. Throughout any scientific inquiry which is an iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables which leads to the study of multivariate analysis.

Cluster analysis is an important statistical tool with respect to multivariate exploratory data analysis. It involves intricate techniques, methods and algorithms which can be applied in various fields, including economics and other social research. The aim of cluster analysis is to identify groups of similar objects (e.g. countries, enterprises, households) according to selected variables (e.g. unemployment rate of men and women in different countries, deprivation indicators of households, etc.). Cluster analysis is typically used in the exploratory phase of research when the researcher does not have any pre-conceived hypotheses or prior knowledge regarding the similarity of the objects. It is commonly not only the statistical method used, but rather is done in the early stages of a project to help guide the rest of the analysis (Timm, 2002, Hair *et al.*, 2006).

Cluster analysis differs from other methods of classification such as Discriminant analysis where classification pertains to known number of groups and the operational objective is to assign new observations to one of these groups. Whereas cluster analysis is a more primitive tool as in that no assumptions are made about the number of groups or the group structure and the grouping is done based on similarities or distances (dissimilarities).

Cluster analysis is also an important tool for investigation in data mining. For example, in marketing research consumers can be grouped on the basis of their preferences. In short it is possible to find application of cluster analysis in any field of research.

CLUSTERING METHODS

The commonly used methods of clustering are divided into two categories (Johnson and Wichern, 2006).

- (i) Hierarchical and
- (ii) Non-Hierarchical.

Hierarchical Cluster Analysis

Hierarchical clustering techniques proceed by either a series of mergers or a series of successive divisions. Agglomerative hierarchical method starts with the individual objects, thus there are as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all sub groups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two sub groups such that the objects in one sub group are far from the objects in the others sub groups. These sub groups are then further divided into dissimilar sub groups. The process continues until there are as many sub groups as objects i.e., until each object form a group. The results of both agglomerative and divisive method may be displayed in the form of two dimensional diagram known as Dendrogram. It can be seen that the Dendrogram illustrate the mergers or divisions that have been made at successive levels.

Linkage methods are suitable for clustering items, as well as variables. This is not true for all hierarchical agglomerative procedures. The following linkage are now discussed:

- (i) single linkage (minimum distance or nearest neighbour)
- (ii) complete linkage (maximum distance or farthest neighbour)
- (iii) average linkage (average distances)

Other methods of hierarchical clustering techniques like Ward's method and Centroid method are also available in literature.

Steps of agglomeration in Hierarchical cluster analysis

The following are the steps in the agglomerative hierarchical clustering algorithm for groups of N objects (items or variables).

- i. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distance (or similarities) $D = \{d_{ik}\}$.
- ii. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between most similar clusters U and V be d_{uv} .
- iii. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters.

- iv. Repeat steps (ii) and (iii) a total of $N-1$ times (all objects will be in a single cluster after the algorithm terminates). Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

The basic ideas behind the cluster analysis are now shown by presenting the algorithm components of linkage methods.

Non Hierarchical Clustering Method

Non Hierarchical clustering techniques are designed to group items, rather than variables, into a collection of K clusters. The number of clusters, K , may either be specified in advance or determined as part of the clustering procedure. Because a matrix of distance does not have to be determined and the basic data do not have to be stored during the computer run. Non hierarchical methods can be applied to much larger data sets than can hierarchical techniques. Non hierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points which will form nuclei of the cluster.

K Means Clustering

The K means clustering is a popular non hierarchical clustering technique. For a specified number of clusters K the basic algorithm proceeds in the following steps (Afifi, Clark and Marg, 2004).

- i. Divide the data into K initial cluster. The number of these clusters may be specified by the user or may be selected by the program according to an arbitrary procedure.
- ii. Calculate the means or centroid of the K clusters.
- iii. For a given case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster whose centroid is closest to it.
- iv. Repeat step (iii) for each case.
- v. Repeat steps (ii), (iii), and (iv) until no cases are reassigned.

Dendrogram

Dendrogram, also called hierarchical tree diagram or plot, shows the relative size of the proximity coefficients at which cases are combined. The bigger the distance coefficient or the smaller the similarity coefficient, the more clustering involved combining unlike entities, which may be undesirable. Trees are usually depicted horizontally, not vertically, with each row representing a case on the Y axis, while the X axis is a rescaled version of the proximity coefficients. Cases with low distance/ high similarity are close together. Cases showing low distance are close, with a line linking them a short distance

from the left of the Dendrogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating likeness. When, the linking line is to the right of the Dendrogram the linkage occurs at a high distance coefficient, indicating the cases/ clusters are agglomerated even though much less alike. If a similarity measure is used rather than a distance measure, the rescaling of the X axis still produces a diagram with linkages involving high likeness to the left and low likeness to the right.

Distance measures

Given two objects X and Y in a 'p' dimensional space, a dissimilarity measure satisfies the following conditions:

1. $d(X,Y) \geq 0$ for all objects X and Y
2. $d(X,Y) = 0$ $X = Y$
3. $d(X,Y) = d(Y,X)$

Condition (3) implies that the measure is symmetric so that the dissimilarity measure that compares X and Y is same as the comparison for object Y verses X. Condition (2) requires the measures to be zero, when ever object X equals to object Y. The objects are identical if $d(X, Y) = 0$. Finally, Condition (1) implies that the measure is never negative.

Some dissimilarity measures are as follows.

Euclidian distance

This is probably the most commonly chosen type of distance. It is simply the geometric distance in the multidimensional space. It is computed as,

$$d(X,Y) = \left\{ \sum_{i=1}^p (X_i - Y_i)^2 \right\}^{1/2} \quad \text{or}$$

In matrix form

$$d(X,Y) = \sqrt{(X-Y)'(X-Y)}$$

Where $X = (X_1, X_2, \dots, X_p)$

$$Y = (Y_1, Y_2, \dots, Y_p)$$

The statistical distance between the same two observations is of the form

$$d(X,Y) = \sqrt{(X-Y)' A (X-Y)},$$

where $A = S^{-1}$ and S contains the sample variances and covariances.

Euclidian and square Euclidian distances are usually computed from raw data and not from standardized data.

Square euclidean distance

Square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as:

$$d^2(X,Y) = \left\{ \sum_{i=1}^p |X_i - Y_i|^m \right\}^{\frac{1}{m}}$$

or in matrix form

$$d^2(X,Y) = (X - Y)' (X - Y)$$

Minkowski metric

When there is no idea about prior knowledge of the distance group then one goes for Minkowski metric. This can be computed as given below:

$$d(X,Y) = \left\{ \sum_{i=1}^p |X_i - Y_i|^m \right\}^{\frac{1}{m}}$$

For $m = 1$, $d(X,Y)$ measures the city block distance between two points in p dimensions.

For $m = 2$, $d(X,Y)$ becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

City-block (Manhattan) distance

This distance is simply the average difference across dimensions. In most cases, this distance measure yields result similar to the simple Euclidean distance. This can be computed as :

$$d(X,Y) = \sum_{i=1}^p |X_i - Y_i|$$

Chebychev distance

This distance measure may be appropriate in case when we want to define the objects as different if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$d(X,Y) = \text{maximum } |X_i - Y_i|$$

Two additional popular measures of distance or dissimilarity are given by the Canberra metric and the Czekanowski coefficient. Both of these measures are defined for non negative variables only. We have

Canberra metric:
$$d(X, Y) = \sum_{i=1}^p \frac{|X_i - Y_i|}{(X_i + Y_i)}$$

Czekanowski coefficient =
$$1 - \frac{2 \sum_{i=1}^p \min(X_i, Y_i)}{\sum_{i=1}^p (X_i + Y_i)}$$

ILLUSTRATION

Given below is food nutrient data on calories, protein, fat, calcium and iron. The objective of the study is to identify suitable clusters of food nutrient data based on the five variables (Chatfield and Collins, 1990).

Table 1: Food nutrient data on calories, protein, fat, calcium and iron

Food items	Calories	Protein	Fat	Calcium	Iron
1	340	20	28	9	2.6
2	245	21	17	9	2.7
3	420	15	39	7	2
4	375	19	32	9	2.6
5	180	22	10	17	3.7
6	115	20	3	8	1.4
7	170	25	7	12	1.5
8	160	26	5	14	5.9
9	265	20	20	9	2.6
10	300	18	25	9	2.3
11	340	20	28	9	2.5
12	340	19	29	9	2.5
13	355	19	30	9	2.4
14	205	18	14	7	2.5
15	185	23	9	9	2.7
16	135	22	4	25	0.6
17	70	11	1	82	6
18	45	7	1	74	5.4
19	90	14	2	38	0.8
20	135	16	5	15	0.5
21	200	19	13	5	1
22	155	16	9	157	1.8
23	195	16	11	14	1.3
24	120	17	5	159	0.7
25	180	22	9	367	2.5
26	170	25	7	7	1.2
27	170	23	1	98	2.6

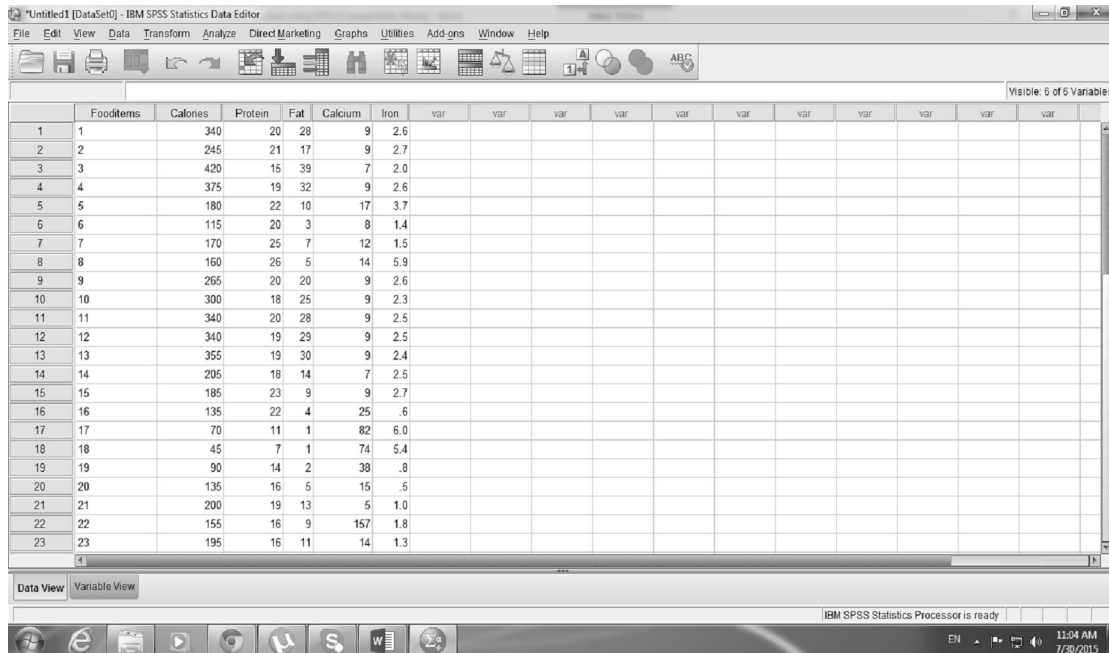
Analysis using SPSS

Start by entering the datasheet into SPSS using the steps below.

Step: Go to file→ open→ browse the datasheet→ click open or

Enter all the data in the data editor as shown in Figure1.

Cluster Analysis



	Fooditems	Calories	Protein	Fat	Calcium	Iron	var	var	var	var	var	var	var	var	var	var
1	1	340	20	28	9	2.6										
2	2	245	21	17	9	2.7										
3	3	420	15	39	7	2.0										
4	4	375	19	32	9	2.6										
5	5	180	22	10	17	3.7										
6	6	115	20	3	8	1.4										
7	7	170	25	7	12	1.5										
8	8	160	26	5	14	5.9										
9	9	265	20	20	9	2.6										
10	10	300	18	25	9	2.3										
11	11	340	20	28	9	2.5										
12	12	340	19	29	9	2.5										
13	13	355	19	30	9	2.4										
14	14	205	18	14	7	2.5										
15	15	185	23	9	9	2.7										
16	16	135	22	4	25	.6										
17	17	70	11	1	82	6.0										
18	18	45	7	1	74	5.4										
19	19	90	14	2	38	.8										
20	20	135	16	5	15	.5										
21	21	200	19	13	5	1.0										
22	22	155	16	9	157	1.8										
23	23	195	16	11	14	1.3										

Fig 1: Screen shot after entering the data in data editor

Now click Analyze→ Classify→ Hierarchical Cluster as shown in Figure2.

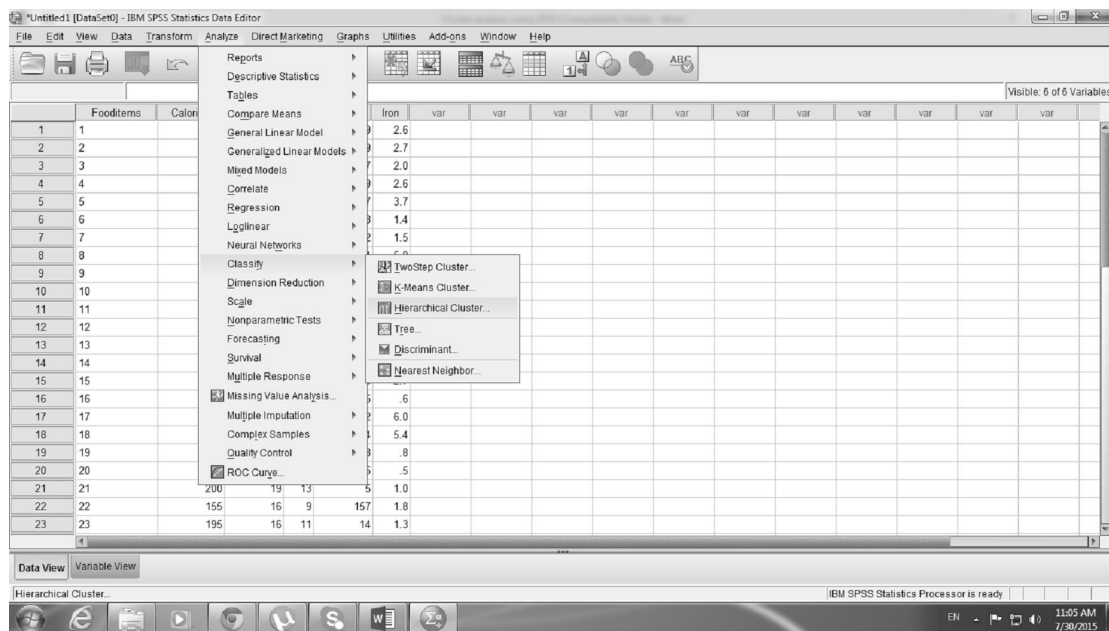


Fig 2: Screen shot of selecting the analysis procedure

Then Identify Name as the variable by which to label cases and Calories, Protein, Fat, Calcium, and Iron as the variables. Indicate that you want to cluster cases rather than variables and want to display both statistics and plots as shown in Fig 3.



Fig 3: Cluster cases rather than variables and want to display both statistics and plots

Click Statistics and indicate that you want to see an Agglomeration schedule with 2, 3,

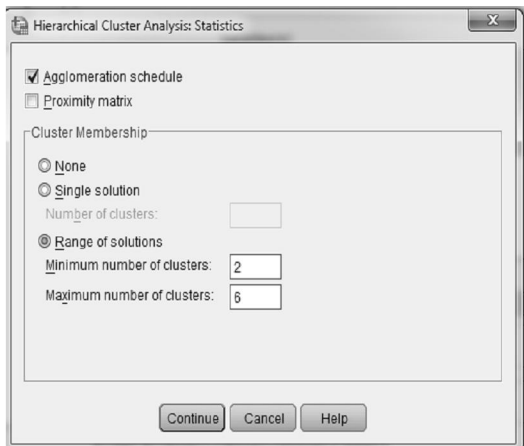


Fig 4: Hierarchical cluster analysis statistics

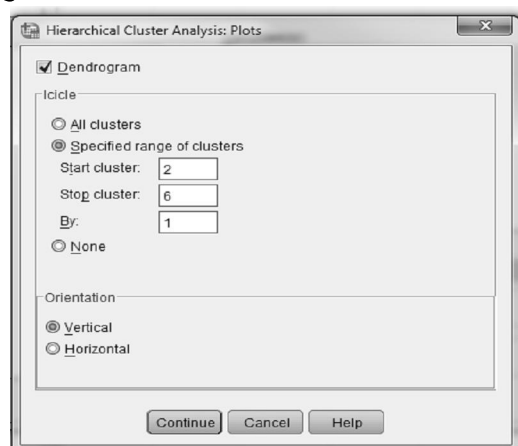


Fig 5: Hierarchical cluster analysis plot

4, and 5 cluster solutions. Click Continue as shown in Fig 4

Click plots and indicate that you want a Dendogram and a verticle Icicle plot with 2, 3, and 4 cluster solutions. Click Continue as shown in Fig 5

Click Method and indicate that you want to use the Between-groups linkage method of clustering, squared Euclidian distances, and variables standardized to z scores (so each variable contributes equally). Click Continue as shown in Fig 6.

Cluster Analysis

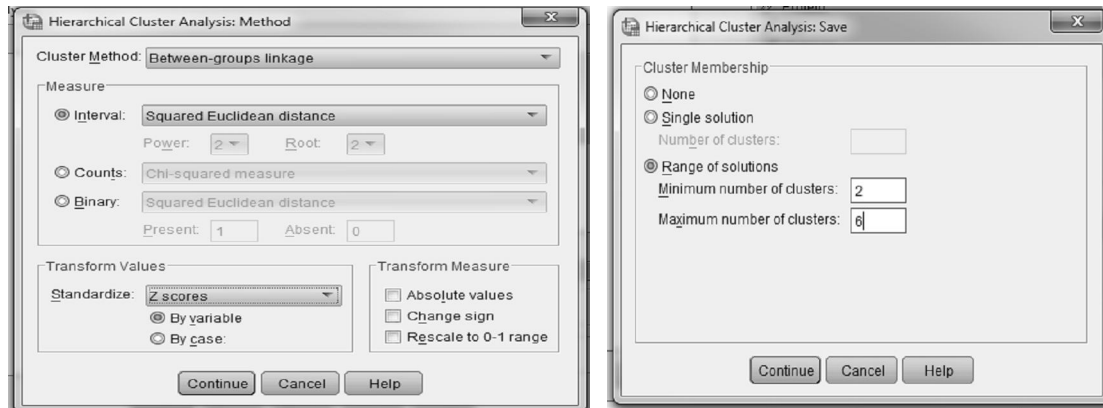
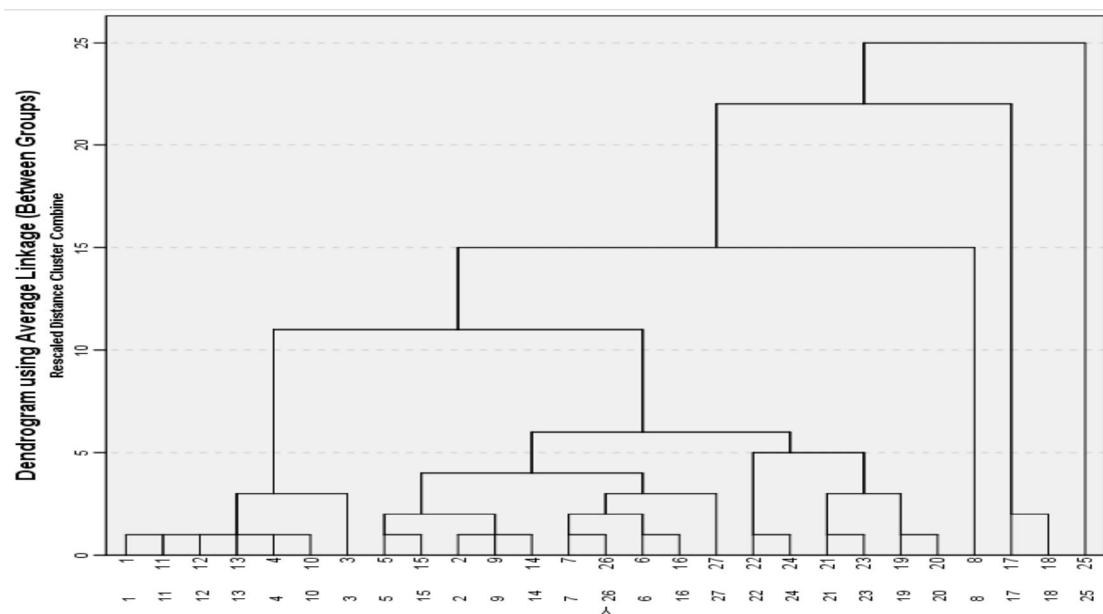


Fig 6: Hierarchical cluster analysis method

Click Save and indicate that you want to save, for each case, the cluster to which the case is assigned for 2, 3, 4, 5 and 6 cluster solutions. Click Continue, OK as shown in Fig 7

SPSS starts by standardizing all of the variables to mean 0, variance 1. This results in all the variables being on the same scale and being equally weighted.

Dendrogram



Interpretation

The main objective of our analysis is to group the food items on the basis of their nutrient content based on the five variables such that food items within the groups are homogeneous and between the groups are heterogeneous.

Table 2: Interpretation

Number of groups	Food items
Two groups	Group-1 (1,11,12,...,18) Group-2 (25)
Three groups	Group-1 (1,11,...,8) Group-2 (17,18) Group-3 (25)
Four groups	Group-1 (1,11,...,20) Group-2 (8) Group-3 (17,18) Group-4 (25)
Five groups	Group-1 (1,11,...,3) Group-2 (5,15,...,20) Group-3 (8) Group-4 (17,18) Group-5 (25)
Six groups	Group-1 (1,11,...,3) Group-2 (5,15,...,27) Group-3 (22,24,...,20) Group-4 (8) Group-5 (17,18) Group-6 (25)

Illustration (Using survey data from social science)

Given below is a part of the data based on a study which was conducted to understand the socio-economic implication of climate and vulnerability of farmers in arid ecosystem of Rajasthan by Sarkar (2014). Two districts Jodhpur and Jaisalmer were selected from arid ecosystem and 100 farmers were selected randomly for the present study. However, for the present chapter, in order to demonstrate the similarity in terms of adaptive behaviour of the farmers, the cluster analysis was performed by considering variables like awareness, attitude towards climate change, egalitarianism, risk perception w.r.t. 20 farmers.

Table 3: Illustration

Farmers' ID	Awareness	Attitude	Egalitarianism	Risk perception
1	26	60	37	60
2	18	43	25	58
3	25	67	40	65
4	23	53	34	57
5	20	41	37	41
6	16	37	38	50
7	23	60	38	65

Cluster Analysis

Farmers' ID	Awareness	Attitude	Egalitarianism	Risk perception
8	19	41	27	50
9	23	41	26	64
10	26	61	37	60
11	18	48	25	59
12	26	67	40	67
13	23	53	35	57
14	20	41	37	41
15	16	37	38	48
16	25	59	38	66
17	19	40	27	50
18	23	42	27	64
19	26	68	36	61
20	16	42	25	58

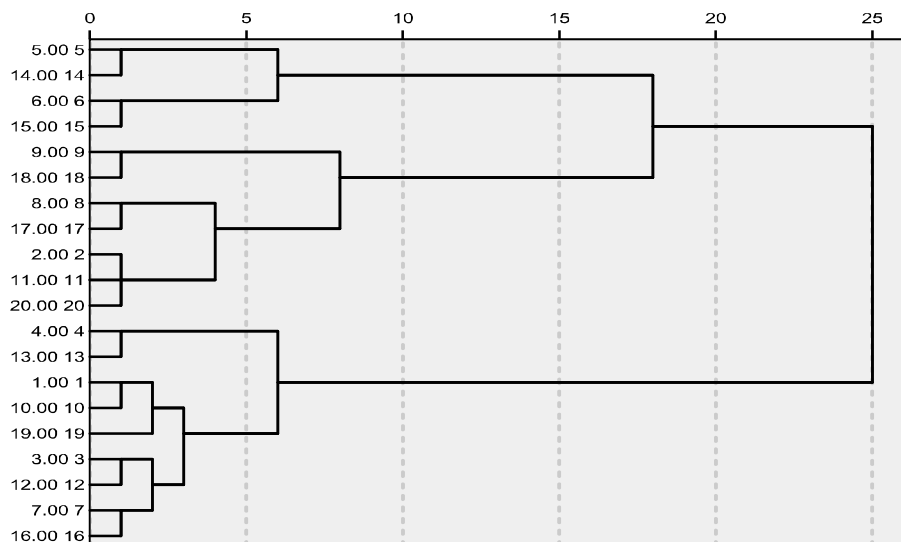
Here, the purpose of the cluster analysis is to group the farmer based on their adaptive behaviour so that appropriate action can be suggested for the farmers who are lagging behind. Two groups were formed viz. adopters and non-adopters. The results are summarized as follows:

Table 4: Adopters and non-adopters

Groups	Farmers' ID
Adopters	1,3,4,7,10,12,13,16 and 19
Non-adopters	2,5,6,8,9,11,14,15,17,18,20

Dendrogram

Dendrogram using Average Linkage (Between Groups) Rescaled Distance Cluster Combine



REFERENCES

- Afi, A., V. A. Clark and S. Marg (2004), Computer Aided Multivariate Analysis. CRC Press, USA.
- Chatfield, C. and A. J. Collins (1990), Introduction to Multivariate Analysis. Chapman and Hall Publications.
- Johnson, R. A. and D. W. Wichern (2006), Applied Multivariate Statistical Analysis. 5th Edn., London, Inc. Pearson Prentice Hall.
- Sarkar, S. (2014), Assessment of Climate Change Led Vulnerability and Simulating the Adaptive Behaviour of Farmers in the Himalayan and Arid Ecosystems. Ph.D. Thesis, IARI, New Delhi.

Chapter 3

PRINCIPAL COMPONENT ANALYSIS

Prem Chand, M. S. Raman and Vinita Kanwal

INTRODUCTION

Principal component analysis (PCA) is one among techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower-dimensional form, without losing too much information. It was invented by Pearson (1901) and Hotelling (1933) and first applied in ecology by Goodall (1954) under the name “Factor Analysis”. During 1970 PCA was considered as the ordination method of choice for community data. Further, simulation studies made by Swan (1970), Austin and Noy-Meir (1971) demonstrated the horseshoe effect and showed that the linear assumption of PCA was not compatible with the non-linear structure of community data. Recently, it has stimulated the search for more appropriate ordination method and is most widely used as well as well known of the “standard” multivariate methods.

PCA is one of the simplest and most robust ways of dimensionality reduction. It is also one of the oldest methods, and has been rediscovered many times in many fields, so it is also known as the Karhunen-Loève transformation, the Hotelling transformation, the method of empirical orthogonal functions, and singular value decomposition. PCA is concerned with explaining the variance covariance structure of a set of variables through a few linear combinations of these variables. Mathematically it is orthogonal linear transformation of data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component (PC)), the second greatest variance on the second coordinate, and so on. PCA is an intermediate step for further treatment of data that includes regression analysis, indexing, assigning weights, etc.

For example, suppose we want to investigate how previous year’s prices of rice and wheat affect the acreage under these crops. We have data pertaining to 100 farmers, collected on prices received by them for three varieties of rice and two varieties of wheat respectively. In this case, we have 100 observations and 5 variables (Table 1).

Table 1: Prices of different varieties of rice and wheat received by farmers during 2016-17 (Rs./q)

Household ID	Rice			Wheat	
	Variety-1 (X1)	Variety-2 (X2)	Variety-3 (X3)	Variety-1 (X4)	Variety-2 (X5)
1	1550	1350	1150	1738	2326
2	1587	995	1404	1471	1538
3	1094	1053	1012	1804	1449
4	1296	1113	930	1747	1299
5	1564	1511	1458	1405	1825
6	1356	1291	1226	1161	2043
7	1168	2500	3832	5164	1625
8	1452	1145	838	1531	1648
9	1291	1352	1413	1474	1548
10	1270	1351	1432	1512	1453
...
100	1296	1347	1399	1450	1074

Now suppose we transform these data by taking average prices of rice and wheat (Table 2). Let us call these variables as T1 and T2. Here T1 is average price of rice and T2 is average price of wheat or we can say that T1 is the linear combination of X1, X2 and X3 ($1/3 \cdot X1 + 1/3 \cdot X2 + 1/3 \cdot X3$) and T2 is linear combination of X4 and X5 ($1/2 \cdot X4 + 1/2 \cdot X5$). Here we have used the data of all five variables and reduced the dimensions to two by converting into the new variable. This transformation of converting linearly correlated variables (X1, X2, X3, X4 & X5) into reduced number of linearly independent variables (T1 and T2) is PC criteria. However, here we are not sure that this dimensionality reduction has retained the variability or not. In fact, in this case we have lost the variability in the data.

Table 2: Average prices of rice and wheat received by farmers during 2016-17 (Rs./q)

Household ID	Average prices of rice (T1)	Average prices of wheat (T2)
1	1350	2032
2	1329	1504
3	1053	1626
4	1113	1523
5	1511	1615
6	1291	1602

Principle Component Analysis

Household ID	Average prices of rice (T1)	Average prices of wheat (T2)
7	2500	3395
8	1145	1590
9	1352	1511
10	1351	1483
...
100	1347	1262

PCA ensures that the variability is captured as much as possible. Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number m of the PCs. If so, there is almost as much information in the m components as there is in original p variables. The p variable can be replaced by m PCs. The original data set consisting of N observations on p variables is reduced to a data set consisting of N observations on m PCs (Bose and Dey, 2011).

ASSUMPTIONS AND LIMITATIONS

Sample size: Correlation coefficients tend to be reliable when estimated from large samples. With increase in sample sizes the sampling distribution become narrower implying that in normal run we get more precise estimates. Therefore, it is important that sample size is large so enough that correlations area could be reliably estimated. Though there is no scientific answer to exact sample size for running of PCA, some arbitrary rules of thumbs say that there should be at least 10 observations for each variable. Hutcheson and Sofroniou (1999) recommend that the minimum sample size should be at least 150- 300. For a few highly correlated variables sample size closer to 150 is sufficient. Some suggest that number of variable determines the size of sample. Bryant and Yarnold (1995), Nunnally (1978) and Gorsuch (1983) recommend that the ideal sample to variable ratio should be at least five. The rule of significance states that there should be 51 extra observations than the number of variables, to support chi-square testing (Lawley and Maxwell, 1971). These rules are not mutually exclusive.

Normality: PCA is a generally a non-parametric analysis. If variables are normally distributed, the solution is enhanced. To the extent normality fails, the solution is degraded by may still be worthwhile.

Linearity: The analysis is degraded when linearity fails, because correlation measures linear relationship and does not reflect non-linear relationship

Orthogonality: Principal components (PCs) are orthogonal. That means the second PC will be perpendicular to first PC and subsequently the third PC will be perpendicular to second PC.

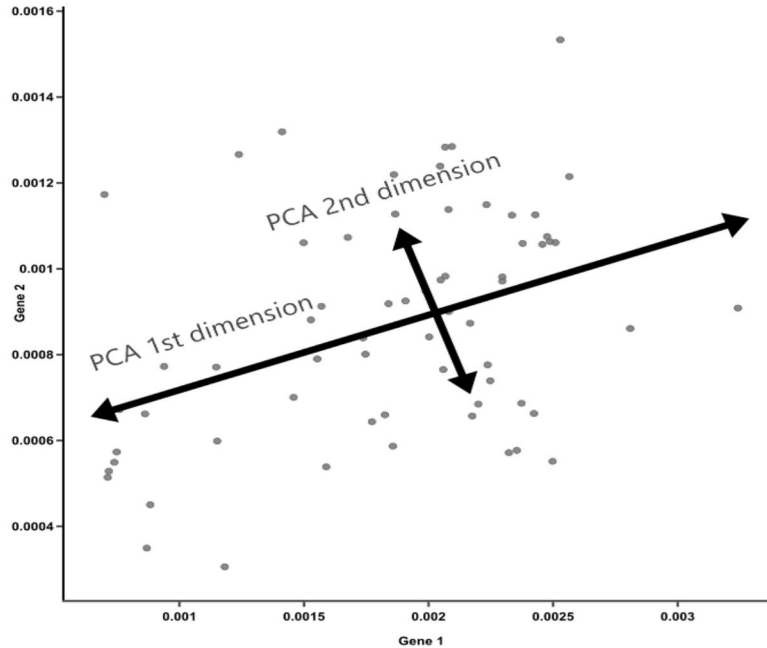


Fig 1: Orthogonality in the principal components

Linear function of PCA

PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called PCs. PCA aims at reducing a large set of variables to a small set that still contains most of the information in the large set. The first PC accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Each PC is a linear function of all the variables. That is, assuming K variables, the K PCs can be written as

$$P_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \dots \alpha_{1K}X_K$$

$$P_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \dots \alpha_{2K}X_K$$

$$\vdots \quad \vdots \quad \vdots$$

Objectives:

$$P_K = \alpha_{K1}X_1 + \alpha_{K2}X_2 + \dots \alpha_{KK}X_K$$

1. Dimension reduction without much loss of information
2. Scale development
3. Evaluation of psychometric quality of a measure
4. Dimensionality assessment of variables

Analysis of PCs frequently serves as intermediate steps in much larger investigations. For example, PCs may be inputs to be multiple regression, factors analysis and discriminant analysis.

STEPS IN PRINCIPAL COMPONENT ANALYSIS

There are five major steps in doing PCA analysis as listed below.

Step 1: Preparation of data: The first and foremost step is preparation of data set. The preparation of data includes finding correlation and, centring and normalizing data. It is worth mentioning that PCA cannot always reduce a large number of original variables to a small number of transformed variables. If the original variables are uncorrelated, the PCA is of no use. Contrary to this, a significant reduction is possible if the original variables are highly correlated. If the variables are significantly correlated, we go for further analysis, i.e. centering and scaling the data. The centring of data produces a data set whose mean is zero. The centered data matrix also called as ‘X’ data matrix is done by subtracting the mean from each variable. If the variances of the variables in the data are significantly different, it is better to scale the data to unit variance. This is achieved by dividing each variable by its standard deviation.

Step 2: Computation of sample variance/covariance matrix C: The formula for calculation of covariance is as follows.

$$C = \frac{1}{N-1} (X - \bar{X}')' (X - \bar{X}')$$

Where, C is covariance matrix, N is sample size, X’ is transpose of centred data matrix X. Since we have centred the dataset to zero mean, the mean vector \bar{X}' will be zero. Therefore, covariance formula will be as follow.

$$C = \frac{1}{N-1} XX'$$

In order to prevent undue influence of any variables of different scale or unit on the PCs, it is common to standardise these so as to have zero means and unit variances. The co-variance matrix C then takes the form of the correlation matrix.

Step 3: Finding eigenvalues and eigenvectors: The third step in PCA is computation of eigenvalues (λ_i) or latent root/unit or normalised eigenvectors (e_i) of A. Here A denotes covariance matrix. An eigenvalue of a matrix A is a scalar (λ) if there is a non-zero vector x satisfies that $Ax = \lambda x$. The eigenvalues of matrix A can be found by solving the characteristics equation, $\det (A - \lambda I) = 0$, where, det is determinant and I is an identity matrix.

However, this is possible only if the number of variables is small. If there are too many variables, solving for λ is non-trivial and we use other methods (Gentle *et al.*, 2004; Golub and van der Vorst, 2000). An important property of the eigenvalues is that they add up to the sum of the diagonal elements of A, i.e., sum of variances of PCs is equal to sum of variances of original variables. Once the eigenvalues of a matrix (A) have been found, we can find the eigenvectors by Gaussian Elimination.

After computing eigenvalues, arrange the variable in descending order based on per cent variation explained by them. Eigenvalues determine the radius of “ellipse”. By this

process, we will be able to extract lines that characterise the data. The first eigenvector will go through the middle of the data point, as if it is the line of best fit. The second eigenvector will give us the other, less important, pattern of the data and so on. The number of chosen eigenvectors will be the number of dimensions of the new data set. The components of lesser significance can be ignored, so as to reduce the dimensions of the data set.

Step 4: Selecting PCs: There are different methodologies for choosing the number of PCs; including both heuristic and statistical. The common methodologies used for selecting PCAs are explained below.

Selection based on proportion of variance that components explain: The cumulative proportion can be used to determine the amount of variance that PCs explain. The PCs are retained that explain an acceptable level of variance. As mentioned above the acceptable level depends on the application. For example, in evocative purposes, you may only need 80% of the variance explained while in other analysis one may want to have at least 90% of the variance explained by PCs.

Selection based on eigenvalues: Size of the eigenvalue can also be used to determine the number of PCs. Generally, the PCs with eigenvalue greater than one are retained for further analysis.

Scree plot method of PC selection: The scree plot orders the eigenvalues from largest to smallest. The ideal pattern is a steep curve, followed by a bend, and then a straight line. The “elbow” location (PC 4 in Figure 2) in a scree plot might indicate a good number of PCs to retain. A more precise method to “detect the elbow” is to start at the right-hand side of the scree plot and look at the points that roughly lie on a straight line. The leftmost point along this line indicates the number of components to be retained. For example, in Figure 2, components 4–7 are almost forming trend line, therefore components 1–4 would sufficiently explain the total variance. This method was proposed by Cattell (1966) and later revised by Cattell and Jaspers (1967).

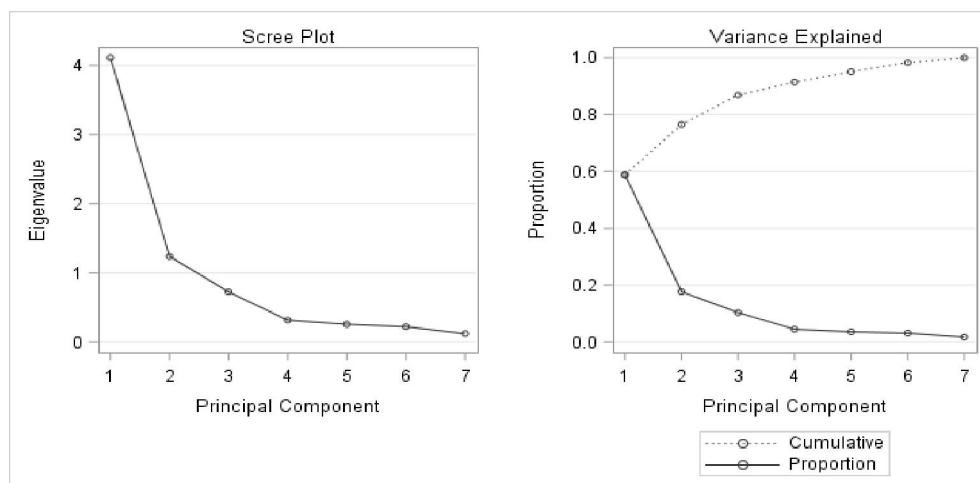


Fig 2: Scree plot to determine number of PCs

Broken-stick method: Broken-stick method is one of the better methods for choosing the number of PCs because this method offers a good combination of simplicity of calculation and accurate evaluation of dimensionality as compared to other statistical approaches (Cangelosi and Goriely, 2007 and Jackson, 1993). The broken-stick model retains components that explain more variance than would be expected by randomly dividing the variance into p parts. While randomly dividing a quantity into p parts, the expected proportion of the k th largest portion is $1/p \sum_{i=k}^p 1/i$, where the summation is over the values $i=k..p$. For example, if $p=5$ then $E1 = (1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6 + 1/7) / 5 = 0.37$; $E2 = (1/2 + 1/3 + 1/4 + 1/5 + 1/6 + 1/7) / 5 = 0.228$, $E3 = (1/3 + 1/4 + 1/5 + 1/6 + 1/7) / 5 = 0.156$, and so on. Eigenvalues of correlation matrix are plotted against the broken-stick proportions, the PCs for which the observed proportions higher than the expected proportions are retained. For example, in Fig 3, as per the broken-stick model only one PC is retained because only the first observed proportion of variance is higher than the corresponding broken-stick proportion.

Step 5: Deriving new dataset

After computation of matrix of eigenvectors and deciding PCs, the final step in PCA is deriving new dataset. The new dataset is derived by taking $Y = XV$, where Y , matrix of transformed data also called as matrix of PC scores. X is original dataset; and V , transformation matrix (eigenvector matrix). Basically we have transformed our dataset so that it is expressed in terms of pattern between them, where the pattern is the lines that most closely describe the relationship between the data.

Rotation of principal components

The interpretation of the components (which is governed by the loadings—the correlations of the original variables with the newly created components) can be enhanced by “rotation” which could be thought of a set of coordinated adjustments of the vectors on a bi-plot. There is no single optimal way of doing rotations, but probably the most common approach is “varimax” rotation in which the components are adjusted in a way that makes the loadings either high positive (or negative) or zero, while keeping the components uncorrelated or orthogonal. Varimax rotation assumes that there is no intercorrelations between components. Varimax rotation (also called Kaiser-Varimax rotation) maximizes the sum of the variance of the squared loadings, where ‘loadings’ means correlations between variables and factors. This usually results in high factor loadings for a smaller number of variables and low factor loadings for the rest. Remaining components all have eigenvalues of more than one. In simple terms, the result is a small number of important variables are highlighted, which makes it easier to interpret your results. One side-product of rotation is that the first, or principal components is no longer optimal or the most efficient single-variable summary of the data set, but losing that property is often worth the increase in interpretability.

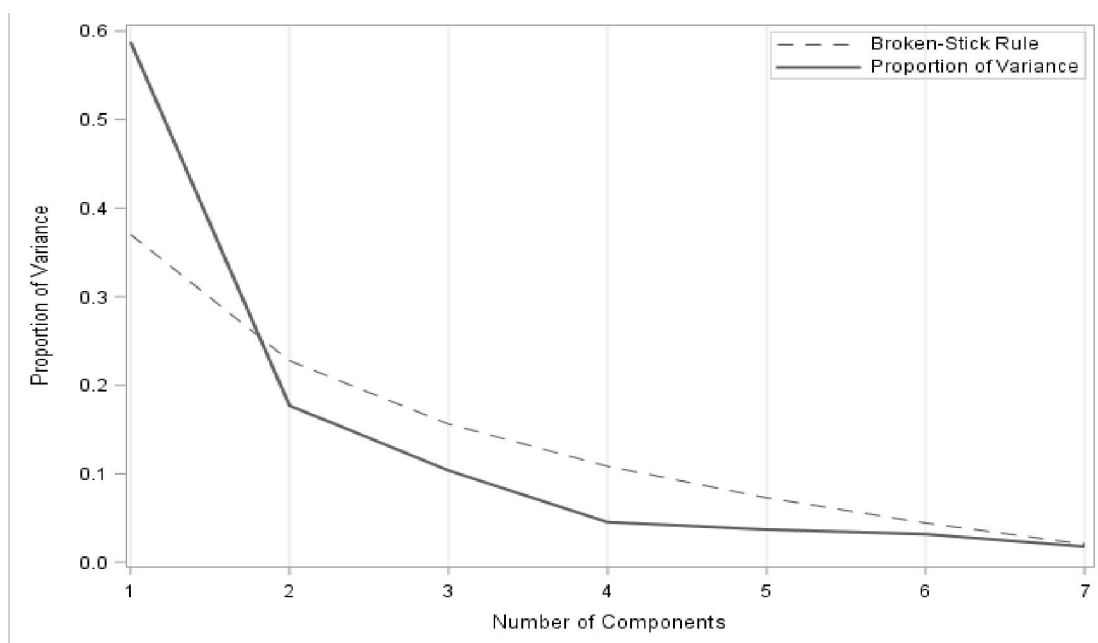


Fig 3: Broken Stick method for retaining PCs

5.1 Analysis using SPSS

Start by entering the datasheet into SPSS using the steps below.

Step: Go to file → open → browse the datasheet → click open or

Enter all the data in the data editor as shown in Fig 4.

	District	Surplus	Pasture	Institution	CPR growth	Literacy	Productivity stock	Productivity Labour	Calories	var	var	var	var	var	var	var	var	var	var
1	AJMER	735089.90	35.10	21.90	.90	49.10	4205.60	32823.50	421.10										
2	ALWAR	836702.30	20.80	33.40	4.90	44.00	3733.60	17729.20	570.30										
3	BANSW	901632.70	54.20	39.70	-.40	27.90	1543.60	8324.30	265.90										
4	BAREN	401941.50	91.00	14.90	.20	42.20	1580.70	9012.50	237.10										
5	BAREM	59939.80	16.70	4.90	.10	43.90	2054.40	15335.40	317.80										
6	BHARA	633119.80	10.00	48.20	-.60	44.10	3383.10	19188.80	531.10										
7	BHILW	792650.20	53.00	29.60	.60	33.50	2014.70	13961.00	325.10										
8	BIKAN	2652181.20	12.80	5.10	1.30	42.60	2539.10	27159.40	461.50										
9	BUNDI	359643.50	71.00	19.40	.40	37.80	2230.30	14761.40	369.00										
10	CHITT	813799.70	67.90	23.10	3.00	36.50	1342.60	7859.10	274.10										
11	CHURU	1331503.30	4.70	8.70	-.60	53.90	3176.30	6433.60	267.60										
12	DAUSA	362200.30	24.00	38.20	-.80	43.20	4484.70	25781.90	574.40										
13	DHOLP	339495.20	30.40	28.30	2.30	42.40	2835.20	21616.00	462.00										
14	DUNGA	753025.20	80.80	62.60	-.40	31.20	1149.80	7192.50	206.00										
15	GANGA	702910.00	9.40	18.00	-.40	52.70	4603.80	33102.60	510.70										
16	HANUM	1229934.70	2.80	12.40	3.50	52.70	3713.10	22406.90	574.90										
17	JAIPU	844807.70	24.90	35.10	-.70	56.20	5039.10	40799.50	398.00										
18	JAISA	961922.10	37.30	1.80	.20	32.30	1194.00	28454.20	360.40										
19	JALOR	401006.70	11.00	14.50	-.10	27.50	3053.50	18145.20	578.50										
20	JHALA	631734.80	56.60	18.70	.30	40.40	2189.20	12443.10	287.10										
21	JHUNA	162244.50	20.30	38.30	-.10	60.10	3965.80	20351.60	433.30										
22	JODHP	154284.40	14.00	8.60	.20	39.20	3459.80	24073.70	292.00										
23	KARAU	431092.80	111.70	20.60	.60	45.40	3448.80	21583.90	501.80										
24	KOTA	344632.00	55.40	21.70	-.60	61.30	2360.10	17909.80	155.40										
25	NAGAU	126958.50	8.20	13.10	-.10	40.50	2593.80	15992.40	285.80										
26	PALI	740304.90	33.90	14.50	.10	36.70	2603.10	24523.90	391.30										
27	RAISA	560349.20	89.80	46.10	.00	37.90	2544.20	25507.70	448.30										
28	SAWAI	398931.30	43.30	21.80	-.50	35.40	2370.90	12792.00	268.20										

Fig 4: Screen shot after entering the data in data editor

Principle Component Analysis

Now click Analyze→ Dimension reduction→ Factor as shown in Fig 5.

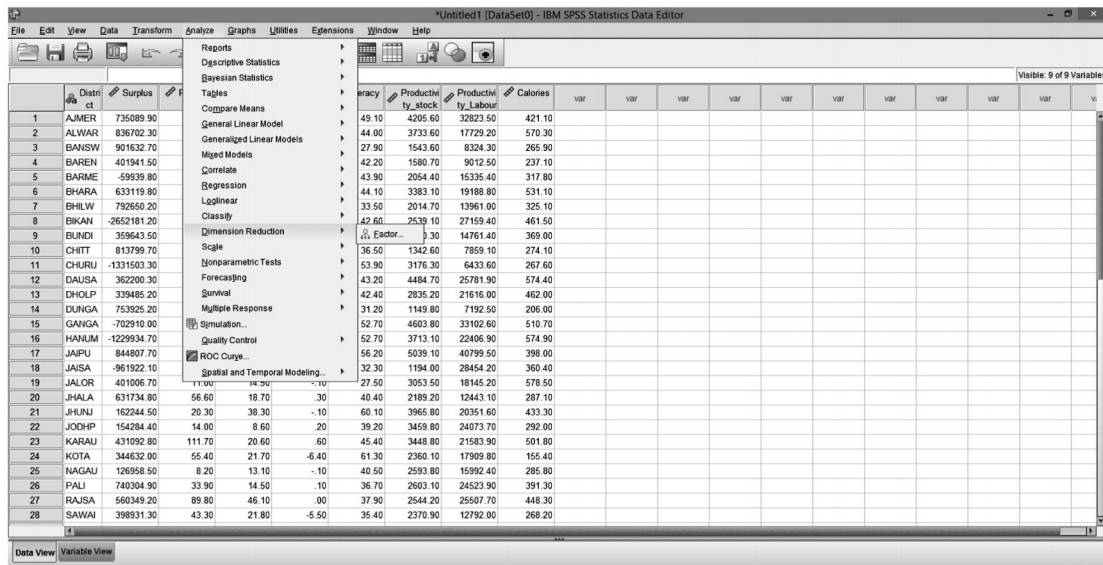


Fig 5: Screen shot of selecting the analysis procedure

Then Identify Name as the variables to which we want to reduce. In this dataset surplus, pastures, institutions, CPR, productivity of stock and productivity of labours are the variables we suspect to be highly correlated and want to reduce in dimension (Fig 6). Select descriptive→ check univariate descriptive initial solutions; Determinants→ continue

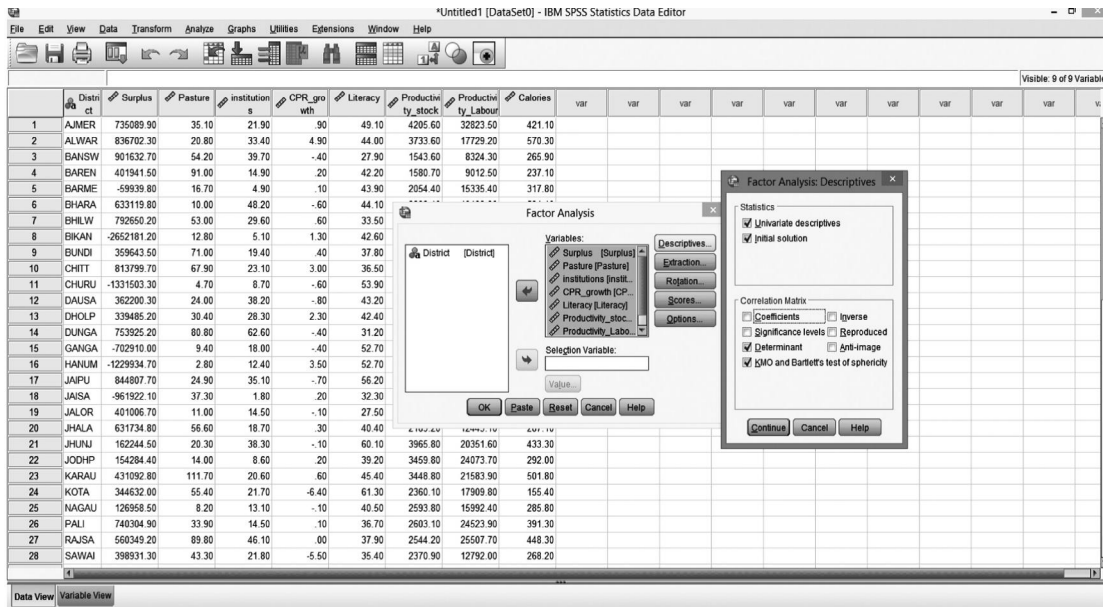


Fig 6: Screen shot of selecting the variables

Select extraction → Method: check principal components, correlation matrix, unrotated factor solution, Scree plot; extract→based on eigen value → continue (Fig 7)

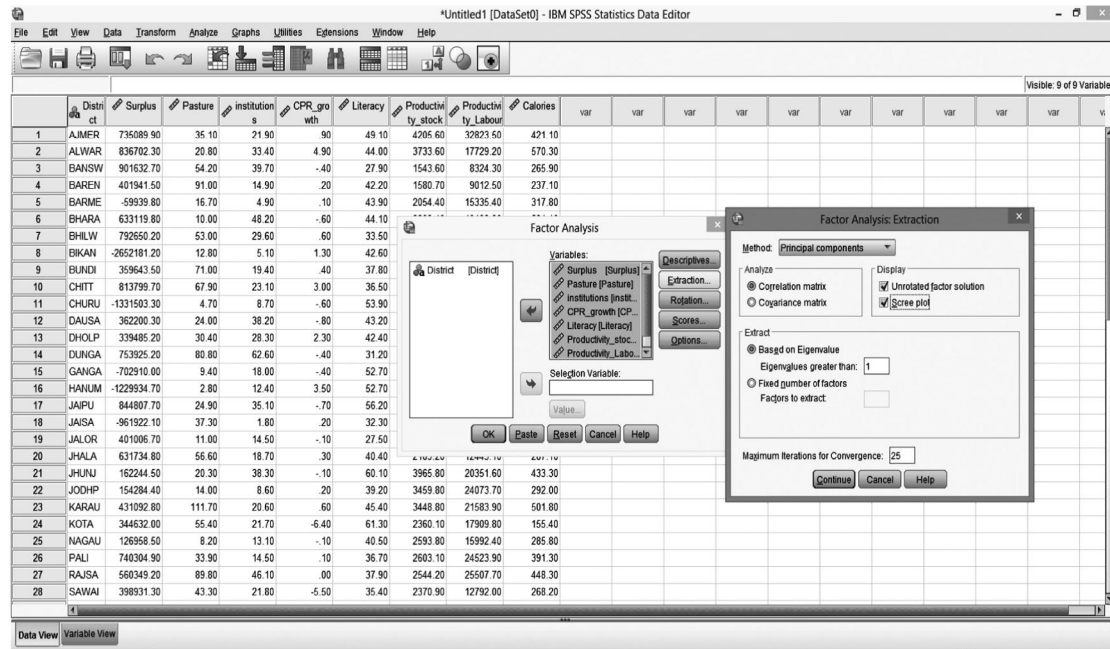


Fig 7: Screen shot showing detailing for principal component

Select Rotation→ check varimax→ continue (Fig 8)

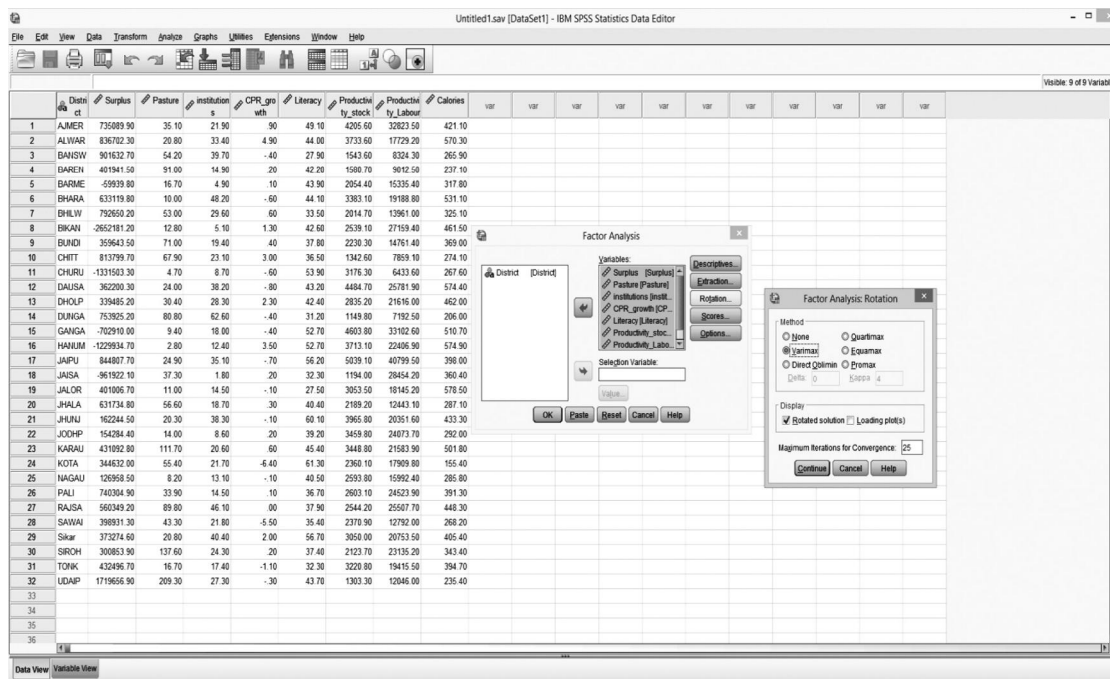


Fig 8: Factor analysis rotation using Varimax method

Fig 9 depicts the result obtained from dummy exercise conducted by the authors. Out of the total variance explained and Scree plot; any one can be chosen to reduce the dimensions of the data. The results indicate that first three eigen values explains 72% of the variance out of total in given data set. Rotated component matrix or loadings obtained at the end, is the key output of principal components analysis. It contains estimates of the correlations between each of the variables and the estimated components.

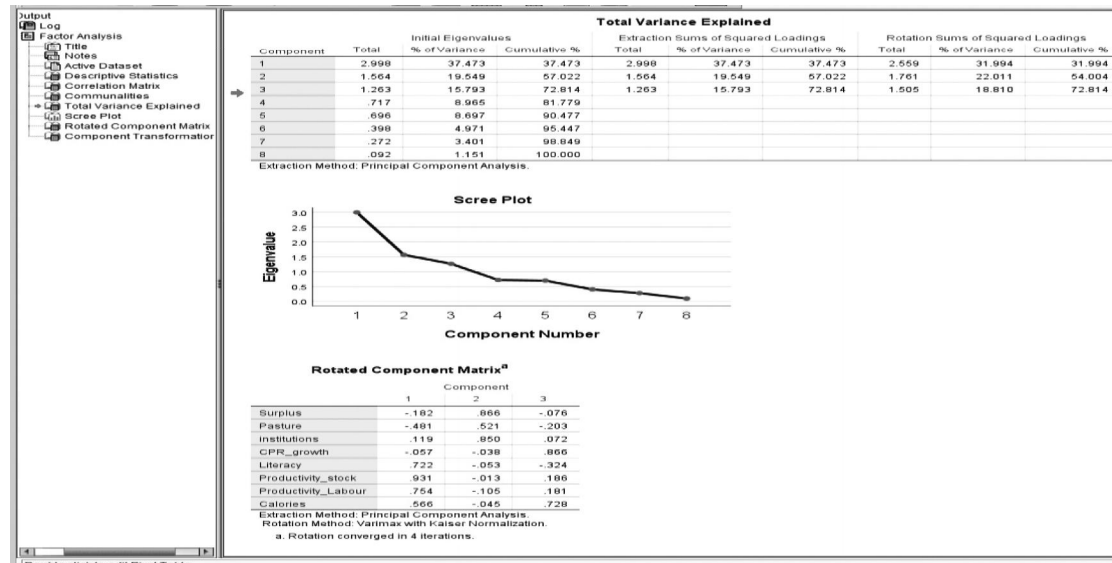


Fig 9: Results of principle component analysis

REFERENCES

- Austin, M. P. and I. Noy-Meir (1971), The problem of non-linearity in ordination: experiments with two gradient models. *Journal of Ecology*, 59: 763–773.
- Bose, A. and A. Dey (2011), The wonderful world of eigenvalues. In: R. Sujatha, H. N. Ramaswamy, C. S. Yogananda. (Eds.) *Math Unlimited: Essays in Mathematics*. CRC Press, Boca Raton.
- Bryant, F. B. and P. R. Yarnold (1995), *Principal components analysis and exploratory and confirmatory factor analysis*. In Grimm and Yarnold, *Reading and understanding multivariate analysis*. American Psychological Association Books.
- Cattell, R. B. (1966), The scree test for the number of factors. *Multivariate behavioral Research*, 1(2): 245-276.
- Cattell, R. B. and J. Jaspers (1967), A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*.
- Cangelosi, R. and A. Goriely (2007), Component retention in principal component analysis with application to cDNA microarray data. *Biology Direct*, 2 (2): 1-21.

- Gentle, J. E., Härdle W. and Mori, Yuichi (eds.) (2004), *Handbook of Computational Statistics: Concepts and Methods*, Springer.
- Golub, G. H. and H.A. van der Vorst (2000), Eigenvalue computation in the 20th century, *Journal of Computational and Applied Mathematics*, 123 (1-2): 35-65.
- Goodall, D. W. (1954). Objective methods for the classification of vegetation: An essay in the use of factor analysis. *Australian Journal of Botany*, 2: 304-324.
- Gorsuch, R. L. (1983), *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum. Orig. ed. 1974.
- Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* 24: 417-441, 498-520.
- Hutcheson, G. and N. Sofroniou (1999), *The multivariate social scientist: Introductory statistics using generalized linear models*, Thousand Oaks, CA: Sage Publications.
- Jackson, D. A. (1993), Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches, *Ecology*, 74(8).
- Lawley, D. N. and A. E. Maxwell (1971), *Factor analysis as a statistical method*, London: Butterworth and Co.
- Nunnally, J. (1978), *Psychometric theory*, New York: McGraw-Hill.
- Pearson, K. (1901), On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.
- Swan, J. M. A. (1970), An examination of some ordination problems by the use of simulated vegetational data. *Ecology*, 51: 89–102.

Chapter 4

MULTIDIMENSIONAL SCALING

Ramasubramanian V.

INTRODUCTION

Multi-dimensional scaling (MDS) is basically a data visualization method that helps the analyst to uncover the hidden structure in multidimensional data. While a number of MDS methods exist with a plethora of different algorithms to arrive at MDS plots, each has been designed to arrive at an optimal low-dimensional configuration (usually in two or three dimensions) for the data under consideration. Some early applications for which MDS has been used were to generate a perceptual map given the (perceived) preferences/ similarities/ dissimilarities of objects subjected to pairwise comparison by individuals and also to reconstruct an original map given a two-way table of (objective) proximities/ distances of places (say, cities), instrumental in paving a way for a greater understanding of the power of MDS. Needless to say, over time, MDS has been extensively used in various application areas such as psychology, marketing, molecular biology and bioinformatics, agriculture etc. In these applications, in some cases, MDS has been used to identify key dimensions underlying respondents' evaluations of objects. From another standpoint, MDS has been used for identifying clusters of points, with points within a particular cluster viewed as being 'closer' to the other points in that cluster than to points present in other clusters. The theoretical concept of MDS, its various methods (as it is not a single procedure), elaboration of the analytical methods involved are discussed along with applications of the different types of MDS employed illustrated with the help of examples. In this chapter, these MDS procedures have been demonstrated on various datasets by analysing them in R software and hence the R codes and corresponding output with interpretation are also given (along with datasets used).

A detailed discussion regarding various MDS techniques can be found in many books (Kruskal and Wish, 1978; Chatfield and Collins, 1980; Coxon, 1982; Hair *et al.*, 1995; Cox and Cox, 2001; Borg and Groenen, 2005; Izenman, 2008). de Leeuw and Mair (2009) have given a good account on MDS discussing the various versions of MDS in a lucid manner and also R software MDS package named SMACOF i.e. *Scaling by MAjorising a COmplicated Function* in which all the known MDS procedures are embedded. Practical applications of MDS has been done by many researchers. Vishwanath and Chen (2006) have examined empirically the composition of technology clusters of several technology concepts and the differences in these clusters formed by adopters and non-adopters using the Galileo system of MDS and the associational diffusion framework. Ramasubramanian *et al.* (2014), while

envisioning future technological needs for plant breeding and genetics subdomain of Indian agriculture, employed MDS using information obtained from experts for identifying key agricultural dimensions emerging out from the factors responsible relating to prioritizing new crop varieties and found that the factors seem to cluster together into two genetic and environmental groups.

TYPES AND METHODS OF MDS

The way MDS is done has evolved over years. Usually ‘objects’ (they could be tangible or intangible) are compared pairwise (‘all pairs design’) by different ‘subjects’ (they could be persons rating the objects or these could be repeated observations of same objects). Again the situations under which the data are collected could differ such as experimental conditions, subjects, stimuli etc. Accordingly, the two broad types of MDS are given as follows:

- (i) One way or multi-way (i.e. K) MDS: In case of K (>1) way MDS, each pair of objects has K dissimilarity measures one each from different ‘replications’ like repeated measures, multiple raters etc.
- (ii) One mode or multi-mode MDS: Here the K (>1) modes have their dissimilarities qualitatively different like experimental conditions, subjects, stimuli etc.

Also note that each of the above versions has metric and non-metric versions. An example of the K-way MDS is the INDSCAL i.e. INDividual Differences SCALing wherein K separate $n \times n$ symmetric dissimilarity matrices are there from the K judges for n objects. For more details, see Borg and Groenen (2005).

Classical MDS

Depending on the way the data are collected, various proximity (read similarity)/dissimilarity measures can be taken into consideration. In this section, let us discuss the classical MDS under metric scaling set up. The conventional form of performing MDS is nothing but classical MDS wherein the dissimilarities (always it is better to convert similarities into dissimilarities for convenient representation purposes) can be taken as (say, Euclidean) distances without any additional transformation. Almost any standard book or chapter that attempts to give an exposition on MDS usually starts with a certain ‘golden oldie’ problem which is discussed subsequently. It is obvious that given a map, with cities marked in that, a matrix of pairwise distances (in MDS parlance, ‘distances’ are likened to ‘dissimilarities’) between these cities can always be formed. But suppose, the problem is looked the other way round, i.e. only the pairwise distances between these cities are available and the map has to be constructed. A ‘perceptual map’ solution is possible by the classical MDS approach under metric scaling (as the distances here are actual measurements and not on any gradation on, say, a five point scale, in which case it becomes non-metric; the latter case is encountered when there is a perception/ preference/ dis/similarity matrix)

and sometimes also known as ‘principal coordinates’ analysis. For this, consider the following matrix of distances (in kilometres) among 21 European cities (given in R software itself) Table 1.

Table 1: Distances of European cities in kilometres

↓ City →	1.Athens	2	3	4	5	6	7	8	9	10
2.Barcelona	3313									
3.Brussels	2963	1318								
4.Calais	3175	1326	204							
5.Cherbourg	3339	1294	583	460						
6.Cologne	2762	1498	206	409	785					
7.Copenhagen	3276	2218	966	1136	1545	760				
8.Geneva	2610	803	677	747	853	1662	1418			
9.Gibraltar	4485	1172	2256	2224	2047	2436	3196	1975		
10.Hamburg	2977	2018	597	714	1115	460	460	1118	2897	
11.Hook of Holland	3030	1490	172	330	731	269	269	895	2428	550
12.Lisbon	4532	1305	2084	2052	1827	2290	2971	1936	676	2671
13.Lyons	2753	645	690	739	789	714	1458	158	1817	1159
14.Madrid	3949	636	1558	1550	1347	1764	2498	1439	698	2198
15.Marseilles	2865	521	1011	1059	1101	1035	1778	425	1693	1479
16.Milan	2282	1014	925	1077	1209	911	1537	328	2185	1238
17.Munich	2179	1365	747	977	1160	583	1104	591	2565	805
18.Paris	3000	1033	285	280	340	465	1176	513	1971	877
19.Rome	817	1460	1511	1662	1794	1497	2050	995	2631	1751
20.Stockholm	3927	2868	1616	1786	2196	1403	650	2068	3886	949
21.Vienna	1991	1802	1175	1381	1588	937	1455	1019	2974	1155
↓ City →	11	12	13	14	15	16	17	18	19	20
12.Lisbon	2280									
13.Lyons	863	1178								
14.Madrid	1730	668	1281							
15.Marseilles	1183	1762	320	1157						
16.Milan	1098	2250	328	1724	618					
17.Munich	851	2507	724	2010	1109	331				
18.Paris	457	1799	471	1273	792	856	821			
19.Rome	1683	2700	1048	2007	1011	586	946	1476		
20.Stockholm	1500	3231	2108	3188	2428	2187	1754	1827	2707	
21.Vienna	1205	2937	1157	2409	1363	898	428	1249	1209	2105

Using the above data with the following R code, the output obtained is given in Fig 1.

In the following set of R codes, the statements after ‘#’ are comments

loc <- cmdscale(eurodist,k=2,) #’classical MDS’ with the dataset in ‘eurodist’ (distances #between European cities within datasets package of R software and ‘loc’ is name given from the user’s side for storing the result output in R

x <- loc[, 1] # the first coordinate values for the 21 cities

```
y <- -loc[, 2] # the second coordinate values for the 21 cities; a 'reflection' is done by
adding a negative sign so that North direction is at the top in the map
```

```
plot(x, y, type = "n", xlab = "", ylab = "", asp = 1, axes = FALSE,
```

```
main = "cmdscale(eurodist)") ## note asp = 1, to ensure aspect ratio of y/x is 1 so that
Euclidean distances are represented correctly
```

```
text(x, y, rownames(loc), cex = 1.0)
```

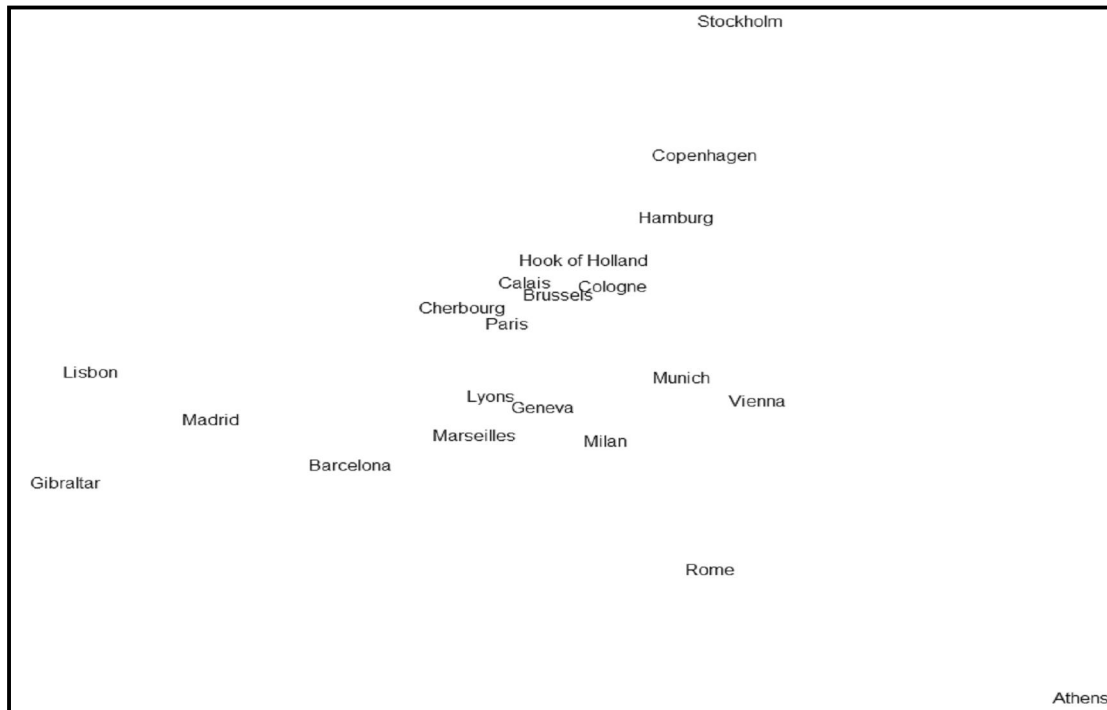


Fig 1: The MDS map constructed using the distances among 21 European cities

The MDS map in Fig 1 can be compared with the actual map given in Fig 2. It can be seen that, by and large, the map has been reconstructed. However, one should also take into account the reflection and rotation of the map. The line distances between cities are represented in the MDS plot while the map involves the curvature represented in a plane. Thus, some distortions can be noticed, for example, the positions of the two cities Madrid and Lisbon in the MDS plot as compared to the original map. Such errors would arise because the distances between cities would in reality need to be represented in more than two dimensions. Nevertheless, essentially what MDS has tried to generate is to find the coordinates (x,y) so that the same can be represented in a map. It is also noted here that Anonymous (2007) has attempted to construct the same using built-in Solver Add-in utility of MS Excel by first principle approach with the steps involved explained in a lucid manner. For this, a starting configuration for the 'n' objects ($n=21$ 'cities' in the above discussion) in the two dimensions i.e. coordinates (x, y) were arbitrarily selected for each object. The Euclidean distances (d_{ij} 's) between the

Now, let us consider another example. Consider the case of the matrix of distances (Table 2) among cities of United States (given in R software itself) and MDS produces a map as shown in Fig 3 which can be compared with the actual map given in Fig 4.

Table 2: Distance between places of United States (in mile)

	Boston	NY	DC	Miami	Chicago	Seattle	SF	LA	Denver
Boston	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
Miami	1504	1308	1075	0	1329	3273	3053	2687	2037
Chicago	963	802	671	1329	0	2013	2142	2054	996
Seattle	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
Denver	1949	1771	1616	2037	996	1307	1235	1059	0

The necessary R codes are given below:

```
df <- read.csv(file.choose(), header = TRUE)
row.names(df) <- df[, 1]
df <- df[, -1]
fit <- cmdscale(df, eig = TRUE, k = 2)
x <- fit$points[, 1]
y <- fit$points[, 2]
plot(x, y, pch = 19, xlim = range(x) + c(0, 600))
city.names <- c("BOSTON", "NY", "DC", "MIAMI", "CHICAGO", "SEATTLE", "SF",
               "LA", "DENVER")
text(x, y, pos = 4, labels = city.names)
x <- 0 - x
y <- 0 - y
plot(x, y, pch = 19, xlim = range(x) + c(0, 600))
text(x, y, pos = 4, labels = city.names)
```

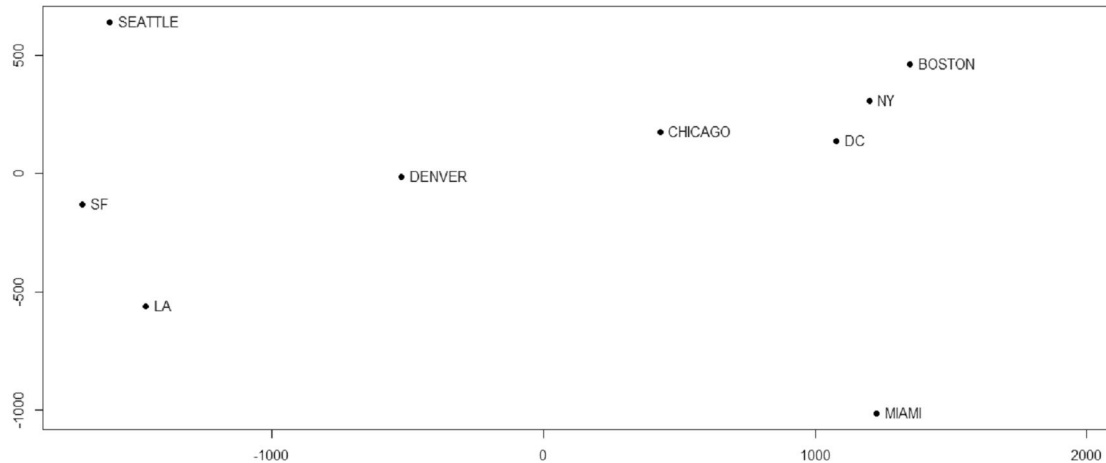


Fig 3: The MDS map of the nine places of United States considered

Normally, MDS is used to provide a visual representation of a complex set of relationships that can be scanned at a glance. Since maps on paper are two-dimensional objects, this translates technically to finding an optimal configuration of points in 2-dimensional space. However, the best possible configuration in two dimensions may be a very poor, highly distorted, representation of your data. When this happens, you can increase the number of dimensions, if required. There are two difficulties with increasing the number of dimensions. Firstly, even three dimensions are difficult to display on paper and are significantly more difficult to comprehend. Four or more dimensions' render MDS virtually useless as a method of making complex data more accessible to the human mind.



Fig 4: The actual map of the nine places of United States considered
(Source: Google)

Secondly, with increasing dimensions, you must estimate an increasing number of parameters to obtain a decreasing improvement in stress. The result is model of the data that is nearly as complex as the data itself.

Non-metric MDS scaling

When the data are not exact distances (dissimilarities) but in terms of perceptions or dis/similarities, then it is non-metric. Let us consider two such case studies, the R codes used for producing MDS plots, the plots and their interpretation.

The first case study relates to correlations of crime rates over 50 states of USA (Borg and Groenen, 2005) given in Table 3. Let us proceed to construct a non-metric MDS using this data. The U.S. Statistical Abstract 1970 issued by the Bureau of the Census provides statistics on the rate of different crimes in the 50 U.S. states. One question that can be asked about these data is to what extent one can predict a high crime rate of murder, say, by knowing that the crime rate of burglary is high. A partial answer to this question is provided by computing the correlations of the crime rates over the 50 U.S. states (Table 3). But even in such a fairly small correlation matrix, it is not easy to understand the structure of these coefficients. This task is made much simpler by representing the correlations in the form of a “picture” of a two-dimensional MDS representation given in Fig 5 where each crime is shown as a point. The points are arranged in such a way that their distances correspond to the correlations. That is, two points are close together (such as murder and assault) if their corresponding crime rates are highly correlated. Conversely, two points are far apart if their crime rates are not correlated that highly (such as assault and larceny).

Table 3: Correlations of crime rates over 50 states of USA

	Murder	Rape	Robbery	Assault	Burglary	Larseny	Autotheft
Murder	1	0.52	0.34	0.81	0.28	0.06	0.11
Rape	0.52	1	0.55	0.7	0.68	0.6	0.44
Robbery	0.34	0.55	1	0.56	0.62	0.44	0.62
Assault	0.81	0.7	0.56	1	0.52	0.32	0.33
Burglary	0.28	0.68	0.62	0.52	1	0.8	0.7
Larseny	0.06	0.6	0.44	0.32	0.8	1	0.55
Autotheft	0.11	0.44	0.62	0.33	0.7	0.55	1

The necessary R codes are given below:

```
df <- read.csv(file.choose(),header = TRUE)
```

```
row.names(df) <- df[, 1]
```

```
df <- 1-df[, -1]
```

```
fit <- cmdscale(df,add=TRUE)
x <- fit$points[, 1]
y <- fit$points[, 2]
plot(x, y,asp = 1, axes = FALSE)
crimes.names <- ("MURDER", "RAPE", "ROBBERY", "ASSAULT", "BURGLARY",
"LARSENY", "AUTOTHEFT")
text(x, y, labels = crimes.names, cex=1.2)
```

Once the MDS plot is obtained, by visual inspection, the dimensions of the perceptual map can be labelled as given in Fig 5.

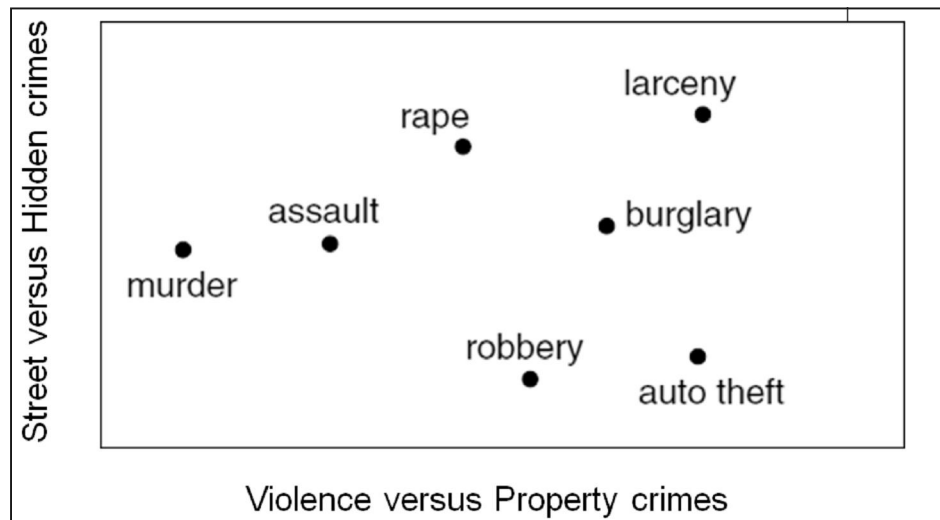


Fig 5: MDS map of crimes over 50 states of USA

The second case study relates to Similarity ratings for 12 nations (rated by individuals) given in Table 4. The data are from a pilot study on perceptions of nations (Kruskal and Wish, 1978). Each of the 18 students (in a psychological measurement course taught by Wish) participating in the study rated the degree of overall similarity between twelve nations on a scale ranging from 1 for “very different” to 9 for “very similar.” There were no instructions concerning the characteristics on which these similarity judgments were to be made; this was information to discover rather than to impose. The first step of the data analysis was to compute the mean similarity rating for each of the 66 pairs (all combinations of the 12 nations i.e. $_{12}C_2$). Thus, for example, USSR and Yugoslavia were perceived to be more similar to each other (mean = 6.67) than any other pair of nations, while China-Brazil and USA-Congo were judged to be the most dissimilar pairs (mean = 2.39). The corresponding MDS plot is given in Fig 6) which also have the dimensions labeled by the researcher.

Table 4: Similarity ratings for 12 nations (rated by individuals)

	Brazil	Congo	Cuba	Egypt	France	India	Israel	Japan	China	USSR	USA	Yugosl.
Brazil	9.00	4.83	5.28	3.44	4.72	4.50	3.83	3.50	2.39	3.06	5.39	3.17
Congo	4.83	9.00	4.56	5.00	4.00	4.83	3.33	3.39	4.00	3.39	2.39	3.50
Cuba	5.28	4.56	9.00	5.17	4.11	4.00	3.61	2.94	5.50	5.44	3.17	5.11
Egypt	3.44	5.00	5.17	9.00	4.78	5.83	4.67	3.83	4.39	4.39	3.33	4.28
France	4.72	4.00	4.11	4.78	9.00	3.44	4.00	4.22	3.67	5.06	5.94	4.72
India	4.50	4.83	4.00	5.83	3.44	9.00	4.11	4.50	4.11	4.50	4.28	4.00
Israel	3.83	3.33	3.61	4.67	4.00	4.11	9.00	4.83	3.00	4.17	5.94	4.44
Japan	3.50	3.39	2.94	3.83	4.22	4.50	4.83	9.00	4.17	4.61	6.06	4.28
China	2.39	4.00	5.50	4.39	3.67	4.11	3.00	4.17	9.00	5.72	2.56	5.06
USSR	3.06	3.39	5.44	4.39	5.06	4.50	4.17	4.61	5.72	9.00	5.00	6.67
USA	5.39	2.39	3.17	3.33	5.94	4.28	5.94	6.06	2.56	5.00	9.00	3.56
Yugosl.	3.17	3.50	5.11	4.28	4.72	4.00	4.44	4.28	5.06	6.67	3.56	9.00

The R codes for constructing the MDS map are given subsequently.

```
df<- read.csv(file.choose(),header = TRUE)
row.names(df) <- df[, 1]
df<- 9-df[, -1]
fit <- cmdscale(df, add=TRUE)
x <- fit$points[, 1]
y <- fit$points[, 2]
plot(x, y,asp = 1, axes = FALSE)
country.names<- c("Brazil","Congo","Cuba","Egypt","France","India","Israel",
"Japan","China","USSR","USA","Yugoslavia")
text(x, y, labels = country.names, cex=1)
```



Fig 6: MDS map for similarity ratings for 12 nations (rated by individuals)

Unfolding

One more variant of MDS is the ‘unfolding’ method which requires rectangular MDS input matrices of order $(n_1 \times n_2)$ wherein n_1 judges rate n_2 objects. Suppose $n_1 = 2$ judges (1 and 2) consider a set of $n_2 = 5$ objects (say research papers A, B, C, D, E) and individually rank them as follows:

	1 st	2 nd	3 rd	4 th	5 th
Judge 1	B	C	A	E	D
Judge 2	A	B	C	E	D

The above observations can be arranged in the following manner:

D	E		C		1	B		2	A
---	---	--	---	--	---	---	--	---	---

In the following arrangement, it can be seen that the distances from judge 1 to the five ‘objects’ have the same ranking as his original ranking of the objects; similarly for judge 2.

1	B		C	A	E	D
2	A	B		C	E	D

For each judge the line can be folded at the judge’s position and their original rankings are observed. Alternatively, looking at the situation in reverse, the judges’ rankings when placed on a line can be ‘unfolded’ to obtain the “common” ordering of the objects.

Thus unfolding tries to model preferential choice by assuming that different individuals perceive the various objects of choice in the same way but differ with respect to what they consider an ideal combination of the objects’ attributes. In unfolding, the data

are preference scores such as rank-orders of different individuals for a set of choice objects. These data can be conceived as proximities between the elements of two sets, individuals and objects. In a way, unfolding can be seen as a special case of MDS where the within-sets proximities are missing. Individuals are represented as ‘ideal’ points in MDS plot so that the distances from each ideal point to the object points correspond to the preference scores. Borg and Groenen (2005) has introduced the basic notions of unfolding models by means of an example. They considered data wherein 42 individuals were asked to rank-order 15 breakfast items (A=toast pop-up; B=buttered toast; C=English muffin and margarine; D=jelly donut; E=cinnamon toast; F=blueberry muffin and margarine; G=hard rolls and butter; H=toast and marmalade; I=buttered toast and jelly; J=toast and margarine; K=cinnamon bun; L=Danish pastry; M=glazed donut; N=coffee cake; O=corn muffin and butter) from 1 (= most preferred) to 15 (= least preferred) which are given in Table 5. In Table 5, each row contains the ranking numbers assigned to breakfast items A, ..., O by individual i (=1 to 42).

Fig 7 presents an unfolding solution to the above data. The resulting configuration consists of 57 points, 42 for the individuals (shown as numbers) and 15 for the breakfast items (shown as their short names). Every individual is represented by an ideal point. The closer an object point lies to an ideal point, the more the object is preferred by the respective individual.

Table 5: Individual rank-order for breakfast items (Unfolding example)

	Breakfast items														
Individual↓	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1.	13	12	7	3	5	4	8	11	10	15	2	1	6	9	14
2.	15	11	6	3	10	5	14	8	9	12	7	1	4	2	13
3.	15	10	12	14	3	2	9	8	7	11	1	6	4	5	13
4.	6	14	11	3	7	8	12	10	9	15	4	1	2	5	13
5.	15	9	6	14	13	2	12	8	7	10	11	1	4	3	5
6.	9	11	14	4	7	6	15	10	8	12	5	2	3	1	13
7.	9	14	5	6	8	4	13	11	12	15	7	2	1	3	10
8.	15	10	12	6	9	2	13	8	7	11	3	1	5	4	14
9.	15	12	2	4	5	8	10	11	3	13	7	9	6	1	14
10.	15	13	10	7	6	4	9	12	11	14	5	2	8	1	3
11.	9	2	4	15	8	5	1	10	6	7	11	13	14	12	3
12.	11	1	2	15	12	3	4	8	7	14	10	9	13	5	6
13.	12	1	14	4	5	6	11	13	2	15	10	3	9	8	7
14.	13	11	14	5	4	12	10	8	7	15	3	2	6	1	9
15.	12	11	8	1	4	7	14	10	9	13	5	2	6	3	15
16.	15	12	4	14	5	3	11	9	7	13	6	8	1	2	10

Multidimensional Scaling

	Breakfast items														
Individual↓	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
17.	7	10	8	3	13	6	15	12	11	9	5	1	4	2	14
18.	7	12	6	4	10	1	15	9	8	13	5	3	14	2	11
19.	2	9	8	5	15	12	7	10	6	11	1	3	4	13	14
20.	10	11	15	6	9	4	14	2	13	12	8	1	3	7	5
21.	12	1	2	10	3	15	5	6	4	13	7	11	8	9	14
22.	14	12	10	1	11	5	15	8	7	13	2	6	4	3	9
23.	14	6	1	13	2	5	15	8	4	12	7	10	9	3	11
24.	10	11	9	15	5	6	12	1	3	13	8	2	14	4	7
25.	15	8	7	5	9	10	13	3	11	6	2	1	12	4	14
26.	15	13	8	5	10	7	14	12	11	6	4	1	3	2	9
27.	11	3	6	14	1	7	9	4	2	5	10	15	13	12	8
28.	6	15	3	11	8	2	13	9	10	14	5	7	12	1	4
29.	15	7	10	2	12	9	13	8	5	6	11	1	3	4	14
30.	15	10	7	2	9	6	14	12	8	11	5	3	1	4	13
31.	11	4	9	10	15	8	6	5	1	13	14	2	12	3	7
32.	9	3	10	13	14	11	1	2	4	5	15	6	7	8	12
33.	15	8	1	11	10	2	4	13	14	9	6	5	12	3	7
34.	15	8	3	11	10	2	4	13	14	9	6	5	12	1	7
35.	15	6	10	14	12	8	2	4	3	5	11	1	13	7	9
36.	12	2	13	11	9	15	3	1	4	5	6	8	10	7	14
37.	5	1	6	11	12	10	7	4	3	2	13	9	8	14	15
38.	15	11	7	13	4	6	9	14	8	12	1	10	3	2	5
39.	6	1	12	5	15	9	2	7	11	3	8	10	4	14	13
40.	14	1	5	15	4	6	3	8	9	2	12	11	13	10	7
41.	10	3	2	14	9	1	8	12	13	4	11	5	15	6	7
42.	13	3	1	14	4	10	5	15	6	2	11	7	12	8	9

The R codes for constructing the MDS map are given subsequently. It is noted here that the SMACOF package of R has to be installed and loaded before proceeding for running these codes.

```
res <- unfolding(breakfast)
```

```
res
```

```
plot(res)
```

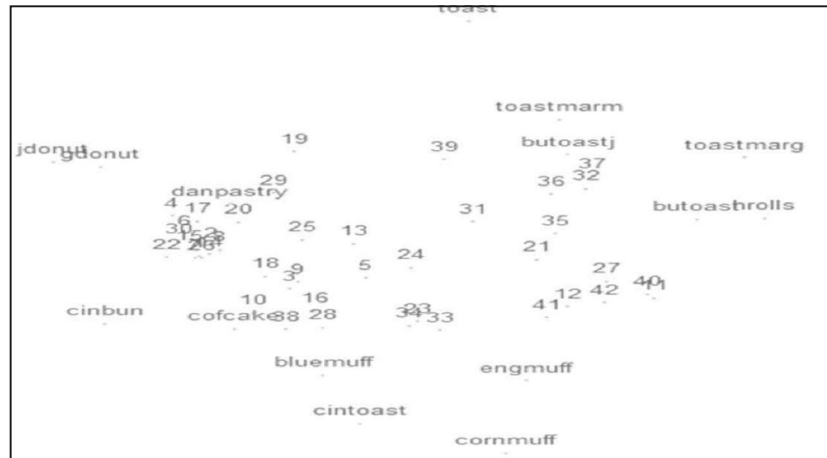


Fig 7: MDS unfolding of individual rank-ordering for breakfast items

Assume that Fig 7 was printed on a thin handkerchief. If this handkerchief is picked up with two fingers at the point representing individual i , say, y_i on the MDS plot and then pulled through the other hand, we have folded it: point y_i is on top, and the farther down the object points, the less preferred the objects they represent. The order of the points in the vertical direction corresponds (if we folded a perfect representation) to how individual i ordered these objects in terms of preference. Picking up the handkerchief in this way at any individual's ideal point yields that individual's empirical rank-order. The MDS process, then, is the inverse of the folding, that is, the unfolding of the given rank-orders into the distances. For example, individual 4 prefers L=Danish pastry (written 'danpastry' in the plot) the most, because the object points of these breakfast items are closest to this individual's ideal point, followed by M=glazed donut (gdonut) and D=jelly donut (jdonut), and then K=cinnamon bun ("cinbun") and N=coffee cake ("cofcake"). It is noted that for individual 4, J=toast and margarine ("toastmarg") is least preferred. Somewhat less preferred is the coffee and cake breakfast (N), whereas A, B, F, I, H, C, G and O are more or less equally disliked. Now, one can confirm this plot ranking by referring to item ranking in the Table 5 against individual 4. In a similar manner, the ranking preferences of breakfast items for other individuals are also unfolded.

In this way, MDS can be used to understand and also unravel the information in any underlying data. While no claim is made to have discussed all the methods under MDS, the essential methods have been truthfully covered.

REFERENCES

Anonymous-Solver-MSExcel (2007), Constructing Perceptual Maps with the Aid of SOLVER in Office. <http://cw.routledge.com/textbooks/9780415458160/instructorresources/MDS%20with%20spreadsheet.docx>, accessed on 01 September, 2019.

- Borg, I. and P. J. F. Groenen (2005), Modern Multidimensional Scaling: Theory and Applications, Second edition, *Springer-Verlag*, New York.
- Chatfield, C. and A. J. Collins (1980), Introduction to Multivariate Analysis, Chapman and Hall, London.
- Cox, T.F. and M. A. A. Cox (2001), Multidimensional Scaling, Second edition. Chapman & Hall/CRC, Boca Raton.
- Coxon, A. P. M. (1982), The User's guide to Multidimensional Scaling, Heinemann Educational Books, Great Britain.
- de Leeuw, J. and P. Mair (2009), Multidimensional scaling using majorization: SMACOF in R, *Journal of Statistical Software*, 31(3): 1-30.
- Hair, J. F., R. E. Anderson, R.L. Tatha, and W. C. Black (1995), Multivariate Data Analysis, 4th Edition, Prentice Hall, New Jersey.
- Izenman, A. J. (2008), Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Springer, New York.
- Kruskal, J. B. and M. Wish (1978), Multidimensional Scaling, Series: Quantitative applications in the Social Sciences, Sage University Press, California.
- Manly, B. F. J. and J. A. N. Alberto (2017), Multivariate statistical methods: A Primer, Fourth Edition, CRC Press, Taylor and Francis Group, Boca Raton, Florida.
- Ramasubramanian, V., A. Kumar, K. V. Prabhu, V. K. Bhatia and P. Ramasundaram (2014), Forecasting technological needs and prioritizing factors in agriculture from plant breeding and genetics domain perspective: A review, *Indian Journal of Agricultural Sciences*, 84 (3): 311-316.
- Vishwanath, A. and H. Chen (2006), Technology clusters: using multidimensional scaling to evaluate and structure technology clusters, *Journal of the American Society for Information Science and Technology*, 57(11): 1451-1460

Chapter 5

CORRESPONDENCE ANALYSIS

Deepak Singh, Raju Kumar, Ankur Biswas, R. S. Shekhawat
and Abimanyu Jhahria

INTRODUCTION

Correspondence Analysis (CA) is an exploratory-multivariate-graphical technique representing non-metric information arranged in two-way contingency tables in correspondence maps. A contingency table is a two way cross-classification, which contains the frequency of items (counts) in each cell having the information about the joint distribution of categorical variables. Categorical variables are variables, which are either nominal or ordinal in nature like counts or frequency. Correspondence analysis has unique capability to represent both linear and non-linear relationships. This technique transforms the categorical data into metric data, applies dimensional reduction technique and represents the information into correspondence maps. In multidimensional reduction, the singular value decomposition is used in which the orthogonal components are extracted in decreasing order of importance so that the maximum information can be presented in two or three-dimensional correspondence maps. The correspondence map of correspondence analysis is the geometric approach having ability to represent the interaction of the two categorical variables graphically.

In 1960's Jean-Paul Benzecri and his colleagues developed this technique and named the technique as “analyse factorielle des correspondances” but later shortened this to “analyse des correspondances” which is “correspondence analysis” in english translation. It is also referred as reciprocal averaging, dual scaling, optimal scaling or scoring, homogeneity analysis, canonical analysis of contingency tables, categorical discriminant analysis, and multivariate quantification of qualitative data. When the contingency tables are three way or higher order then the multiple correspondence analysis is used which is the extension of correspondence analysis and the analytical procedure is similar to correspondence analysis.

Differences from other multivariate techniques

Conceptually it is similar to Principal Component and Factor analysis dealing with summarizing the data in graphical form through singular value decomposition but being assumption free it is a wonderful approach to work with categorical data where Principal Component and Factor analysis fails. Both multidimensional scaling and correspondence analysis can deal with nominal data but the compositional (Attribute – Based) approach of the correspondence analysis distinguishes it from the multidimensional scaling which is decompositional (Attribute – Free) in nature.

Uniqueness of correspondence analysis

All the multivariate techniques deal with either continuous data or main effects of categorical data but this multivariate technique has the ability to deal with main and interaction effects of the categorical data. Therefore, it has the unique features like

- Capability to use categorical data and therefore assumption free
- Compositional in nature
- Ability to represent the relationship of the rows/objects to each other
- Ability to represent the relationship of the columns/variables to each other
- Ability to represent the interaction of the objects variables to each other
- Interdependence of categorical variables
- Dimension reduction and multivariate in nature
- Capable to represent both linear and non-linear relationships.
- Correspondence mapping of categorical variables simultaneously which is the explicit objective of correspondence analysis

Limitations

Correspondence analysis performs poor when cell frequencies in the contingency table are very small or zero, in which case it is usually recommended to combine two or more categories to increase the frequencies of the cells. All data should be non-negative under same scale of correspondence analysis to be applicable and the method treats rows and columns equivalently.

Workflow

The workflow for the correspondence analysis is explained in Fig 1.

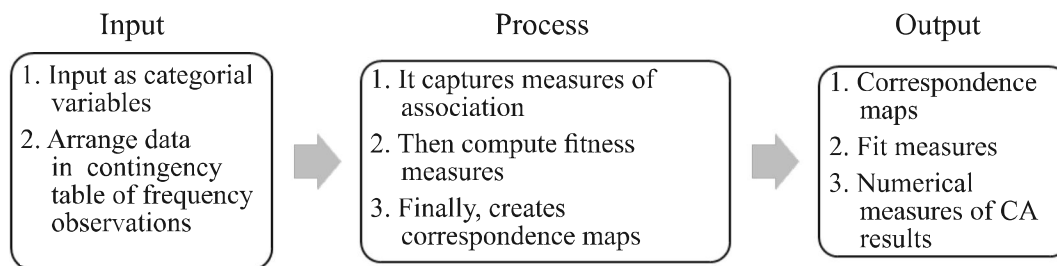


Fig 1: Workflow of the correspondence analysis

Analytical procedure for correspondence analysis

Suppose there are two variables X and Y in nominal scale having levels p and q respectively, which can be expressed in two-way contingency or cross table. The table is bivariate table having i^{th} row and j^{th} column has the cell frequency of $n_{ij} > 0$. If ordinal data is used it is considered as nominal data and the continuous variable like age, plant height, yield, are first converted into nominal data before carrying Correspondence analysis.

This technique examines the correspondence between the observed frequencies and the variables considered either independently or jointly. The contingency table can be expressed in matrix form where n_{ij} are the elements of i^{th} row and j^{th} column of the matrix.

First of all the independence of categories (row by column) is checked by chi-square test statistic and visually the contingency table can be inspected by using graphical matrix and mosaic/ association plots for variables to be interdependent. If the tests are significant, then the further steps are followed for correspondence analysis.

Contingency table

	Y_1	Y_2	...	Y_q	$sum(Y_i)$
X_1	n_{11}	n_{12}	...	n_{1q}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2q}	$n_{2.}$
\vdots	\vdots	\vdots	...	\vdots	\vdots
X_p	n_{p1}	n_{p2}	...	n_{pq}	$n_{p.}$
$sum(X_i)$	$n_{.1}$	$n_{.2}$...	$n_{.q}$	$n_{..}$

ILLUSTRATION

For the illustration purpose, “Housetasks” data has been taken from factextra package of R software. In the Housetasks data rows indicates different household tasks and columns indicate the tasks done by different groups of households i.e. wife, husband, jointly and alternatively. The cell values are the frequencies of the tasks done by different groups of a household.

Housetasks	Wife	Alternatively	Husband	Jointly	Total
Laundry	156	14	2	4	176
Main meal	124	20	5	4	153
Dinner	77	11	7	13	108
Breakfast	82	36	15	7	140
Tidying	53	11	1	57	122
Dishes	32	24	4	53	113
Shopping	33	23	9	55	120
Official	12	46	23	15	96
Driving	10	51	75	3	139
Finances	13	13	21	66	113
Insurance	8	1	53	77	139
Repairs	0	3	160	2	165
Holidays	0	1	6	153	160
Total	600	254	381	509	1744

Our aim in correspondence analysis is to find out the following points:

1. What is the relationship among household tasks with respect to working group of households?
2. What is the relationship among working group of households with respect to household tasks?
3. What is the relationship of household tasks with working group of households?
4. Can these relationships be shown graphically?

To answer the above four questions in correspondence analysis, the following four solutions can be made in correspondence analysis.

Ans 1. Relationship among household tasks = By Row profile (R)

Ans 2. Relationship among working group of households = By Column profile (C)

Ans 3. Relationship between tasks and working group of households = By weighted Chi square distance

Ans 4. Dimension reduction (Singular value decomposition, SVD) and correspondence maps.

Steps for Analytical Procedure

Step 1: First test the independence of categorical variables with chi square test statistics and if test is significant, then we go for next step.

Step 2: Develop correspondence matrix $m_{ij} = \left(\frac{n_{ij}}{n_{..}} \right)$

Step 3: Develop row profile i.e. $(m_{ij} / \text{row mass})$

Step 4: Develop column profile i.e. $(m_{ij} / \text{column mass})$

Step 5: Analyze weighted χ^2 distance = $D = D_r^{-1/2} (M - rc^T) D_c^{-1/2}$

Step 6: Carry out singular value decomposition (SVD) or dimension reduction technique

Step 7: Calculate overall fit measures and plot correspondence maps

Step 1: Pearson's Chi-square test for Housetasks data

First of all the chi-square test is applied to the housetasks contingency table to test whether the categorical variables are independent for further analysis.

$$\chi^2 = 1944.456, df = 36, p\text{-value} = 0$$

The p value is almost zero, which infers that the household tasks and different groups of a households are interdependent. Therefore the correspondence analysis should be applied for inter-relationship of tasks and working group of households.

Step 2: Correspondence matrix M

$$M_{pq} = m_{ij} = \left(\frac{n_{ij}}{n_{..}} \right) =$$

	Y_1	Y_2	...	Y_q	row mass
X_1	m_{11}	m_{12}	...	m_{1q}	$m_{1.}$
X_2	m_{21}	m_{22}	...	m_{2q}	$m_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
X_p	m_{p1}	m_{p2}	...	m_{pq}	$m_{p.}$
column mass	$m_{.1}$	$m_{.2}$...	$m_{.q}$	1

Correspondence matrix M of Housetasks data:

Task	Wife	Alternatively	Husband	Jointly	Row mass
Laundry	0.089	0.008	0.001	0.002	0.025
Main meal	0.071	0.011	0.003	0.002	0.022
Dinner	0.044	0.006	0.004	0.007	0.015
Breakfast	0.047	0.021	0.009	0.004	0.020
Tidying	0.030	0.006	0.001	0.033	0.017
Dishes	0.018	0.014	0.002	0.030	0.016
Shopping	0.019	0.013	0.005	0.032	0.017
Official	0.007	0.026	0.013	0.009	0.014
Driving	0.006	0.029	0.043	0.002	0.020
Finances	0.007	0.007	0.012	0.038	0.016
Insurance	0.005	0.001	0.030	0.044	0.020
Repairs	0.000	0.002	0.092	0.001	0.024
Holidays	0.000	0.001	0.003	0.088	0.023
Column mass	0.026	0.011	0.017	0.022	1.000

The vector of row sums of M_{pq} will be row mass vector i.e.

$$\text{Row sum vector} = r = (m_{1.} \quad m_{2.} \quad \dots \quad m_{p.})' =$$

$$(0.025 \quad 0.022 \quad 0.015 \quad 0.020 \quad 0.017 \quad 0.016 \quad 0.017 \quad 0.014 \quad 0.020 \quad 0.016 \quad 0.020 \quad 0.024 \quad 0.023)'$$

The vector of column sums of M_{pq} will be containing column mass in the following manner.

$$\text{Column sum vector} = c = (m_{.1} \quad m_{.2} \quad \dots \quad m_{.q})' = (0.026 \quad 0.011 \quad 0.017 \quad 0.022)'$$

Therefore, transpose of c will be

$$c^T = (m_{.1} \quad m_{.2} \quad \dots \quad m_{.q}) = (0.026 \quad 0.011 \quad 0.017 \quad 0.022)$$

The D_r and D_c are the diagonal matrix of row and column profiles respectively i.e.

$$D_r = \text{diag}(r) = \begin{bmatrix} m_{1.} & 0 & \dots & 0 \\ 0 & m_{2.} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{p.} \end{bmatrix} = \begin{bmatrix} 0.025 & 0 & \dots & 0 \\ 0 & 0.022 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0.023 \end{bmatrix}$$

and

$$D_c = \text{diag}(c) = \begin{bmatrix} m_{.1} & 0 & \dots & 0 \\ 0 & m_{.2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{.q} \end{bmatrix} = \begin{bmatrix} 0.026 & 0 & 0 & 0 \\ 0 & 0.011 & 0 & 0 \\ 0 & 0 & 0.017 & 0 \\ 0 & 0 & 0 & 0.022 \end{bmatrix}$$

Step 3: Row profiles

When the rows of the correspondence matrix are divided by row mass, we get row profiles.

Interpretation

The row profiles are used to compare the relationship and assess the proportion of row by column. It can be evaluated which column in each row account for more or less percentage of counts. SVD can also be applied to row profiles to visually compare the row categories i.e tasks in case of household tasks data.

Step 4: Column profiles

When the columns of the correspondence matrix are divided by column mass, we get column profiles.

Interpretation

The column profiles are used to compare the relationship and assess the proportion of column with respect to row. It can be evaluated which rows in each column account for more or less percentage of counts. SVD can also be applied to column profiles to visually compare the column categories i.e working groups in case of household tasks data.

Step 5: The weighted χ^2 distance

$$D = D_r^{-1/2} (M - rc^T) D_c^{-1/2}$$

Table 1: Weighted χ^2 distance of Housetasks data

Task	Wife	Alternating	Husband	Jointly
Laundry	3.435899	0.460669	0.035104	0.072572
Main meal	2.927131	0.715917	0.130137	0.081171
Dinner	2.161006	0.465753	0.232713	0.381186
Breakfast	2.017169	1.361655	0.449981	0.167867

Task	Wife	Alternating	Husband	Jointly
Tidying	1.391096	0.436609	0.016304	1.629630
Dishes	0.865506	1.008072	0.122516	1.574530
Shopping	0.865512	0.936114	0.286520	1.585126
Official	0.341471	2.11184	0.852017	0.471749
Driving	0.226737	1.942315	2.331838	0.060181
Finances	0.339318	0.539865	0.713327	1.965414
Insurance	0.176797	0.023437	1.642464	2.066348
Repairs	-0.025020	0.089395	4.581751	0.026722
Holidays	-0.024640	0.019741	0.155606	3.843418

Step 6: Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is required to reduce the dimensions of weighted Chi-square distance so that the categorical variables can have visual representation. Therefore, SVD is applied to partition D matrix into three matrices i.e. U, V and S. Here U is $p \times k$ matrix, V is $q \times k$ matrix and S is $k \times k$ diagonal matrix whose diagonal elements are $s_1 \geq s_2 \geq s_3 \dots \geq s_k$ or $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \sqrt{\lambda_3} \dots \geq \sqrt{\lambda_k}$, here k is the reduced dimension of D matrix. The s_i and λ_i are the singular and eigen values of i^{th} component (PCi). Here $k = \min\{p-1, q-1\}$, therefore in “housetasks” data case $k=q-1=3$. The SVD of D matrix is given by

$$D_{pq} = U_{p \times k} S_{k \times k} V_{q \times k}^T$$

Due to positive definite matrix D, the strict positivity of λ_i is guaranteed. The eigen value λ_k represents the weighted variance explained by k^{th} component. Here PC1 is the component corresponding to λ_1 , PC2 is the component corresponding to λ_2 and so on. The first component PC1 has the maximum variance, the second largest variance is observed for PC2 and so on because of diagonal values arranged in $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \sqrt{\lambda_3} \dots \geq \sqrt{\lambda_k}$ order. The percentage of variation by i^{th} component is given as

$$\% \text{ variance} = \left[\frac{\lambda_i}{\sum_{i=1}^p \lambda_i} \right] * 100$$

Step 7: Overall fit measures and correspondence maps

Overall fit Measures

In overall fit measures, the inertia is defined as percent of variance explained to the total variance by each of the categories. Mass is nothing but row mass and column mass defined earlier in correspondence matrix. Contribution in percent of dimensions indicates what is the contribution of individual categories in building the components and coordinate values are the coordinate values for the principal axis. The correlation in dimensions category indicates about how good the components are good to explain the each of the categories (i.e. row and column elements). Table 2 indicates the eigen

values, percent of variance and cumulative per cent of variance of new components after SVD.

In Table 3, in case of rows, the maximum inertia or per cent of variance is explained by repairs, laundry, driving and holidays tasks while in working group of households the maximum variance is explained by husband group. In case of dimension 1, repairs task contributes maximum and shopping contributes minimum and it can be seen in correspondence plot that along X axis (Dimension 1) the repairs task is farthest while shopping is nearest to origin. The correlation column of dimension 1 is indicating that dimension 1 is good to explain laundry, main meal, dinner, repairs tasks and wife working group of households. Similarly, in case of dimension 2, holidays task contributes maximum and official contributes minimum and it can be seen in correspondence plot that along Y axis (Dimension 2) the holidays task is farthest while official is nearest to origin. The correlation column of dimension 2 is indicating that dimension 2 is good to explain shopping, finances and holidays tasks and jointly working group of households.

Table 2: Eigen value and per cent of variance explained by new components of SVD

Dimension	Eigen value λ	% variance explained by eigen value	Cumulative % variance
PC1	0.54	48.69	48.69
PC2	0.45	39.91	88.60
PC3	0.13	11.40	100.00

Table 3: Overall fit measures of each point in first two dimensions of SVD

Character			Dimension 1			Dimension 2		
Row	Mass	Inertia	Coordinate	Contribution in %	Correlation	Coordinate	Contribution in %	Correlation
Laundry	0.025	0.134	-0.99	18.29	0.74	0.50	5.56	0.18
Main meal	0.022	0.091	-0.88	12.39	0.74	0.49	4.74	0.23
Dinner	0.015	0.038	-0.69	5.47	0.78	0.31	1.32	0.15
Breakfast	0.020	0.041	-0.51	3.82	0.50	0.45	3.70	0.40
Tidying	0.017	0.025	-0.39	2.00	0.44	-0.43	2.97	0.54
Dishes	0.016	0.020	-0.19	0.43	0.12	-0.44	2.84	0.65
Shopping	0.017	0.015	-0.12	0.18	0.06	-0.40	2.52	0.75
Official	0.014	0.053	0.23	0.52	0.05	0.25	0.80	0.07
Driving	0.020	0.102	0.74	8.08	0.43	0.65	7.65	0.34
Finances	0.016	0.030	0.27	0.88	0.16	-0.62	5.56	0.84
Insurance	0.020	0.058	0.65	6.15	0.58	-0.47	4.02	0.31
Repairs	0.024	0.313	1.53	40.73	0.71	0.86	15.88	0.23
Holidays	0.023	0.196	0.25	1.08	0.03	-1.44	42.45	0.96
Column								
Wife	0.026	0.301	-0.84	44.46	0.80	0.37	10.31	0.15
Alternating	0.011	0.118	-0.06	0.10	0.00	0.29	2.78	0.11
Husband	0.017	0.381	1.16	54.23	0.77	0.60	17.79	0.21
Jointly	0.022	0.315	0.15	1.20	0.02	-1.03	69.12	0.98

Correspondence Plot

First two PCs are retained to develop the bi-plot (2d-graph) and three PCs for the tri-plot (3D-plot) as they account the maximum variation among the k components. For example, in case of household's data the bi-plot is explaining 88.60 per cent of variation of data and it can be visualized from the plot that the spread of points in horizontal direction is relatively more than vertical direction, which justifies 48.69 and 39.91 per cent of variation of PC1 and PC2 respectively (Fig 2).

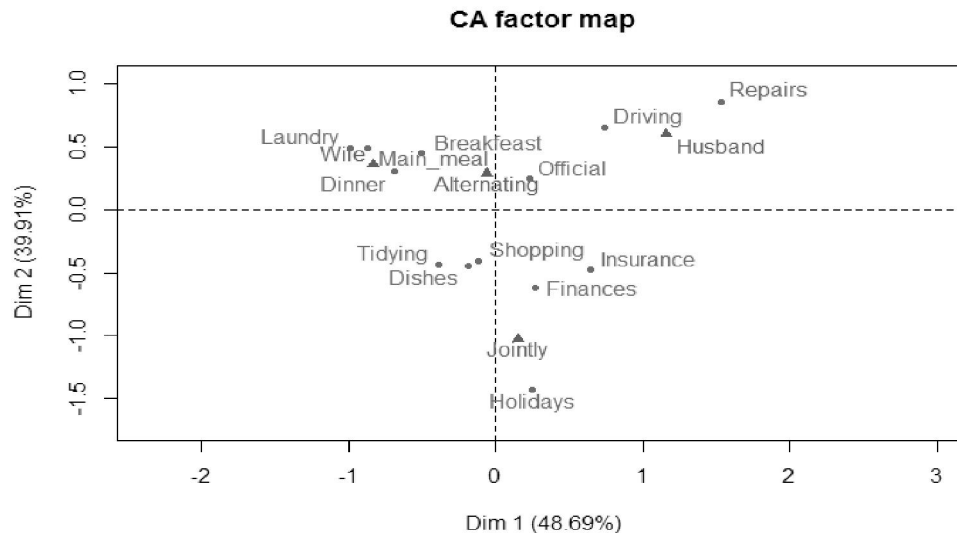


Fig 2: Correspondence plot of correspondence analysis

Inference of Correspondence plot

- Wife and alternating groups together contribute towards laundry, breakfast, main meal and dinner; while Husband and alternating groups together contribute for official, driving and repair tasks. Alternating group is a little biased towards wife group with respect to husband group in doing tasks as it is in the upper left quadrant with wife and highly biased towards wife with respect to jointly group.
- Jointly group is far from wife, alternating and husband group.
- The contribution of alternating is almost negligible compared to wife, husband and jointly group as it is nearest to origin.
- Wife most often do laundry, dinner, breakfast and main meal.
- Husband does official, repairs and driving more often out of which most repairs are done by husband only.
- Jointly they do holidays, shopping, insurance, dishes, tidying and finance tasks more often out of which holidays are most enjoyed jointly.

Finally, it is advised to carefully design and analyze to reach the required objectives to solve any real world problem and the workflow to reach correspondence analysis results is mentioned in the following workflow.

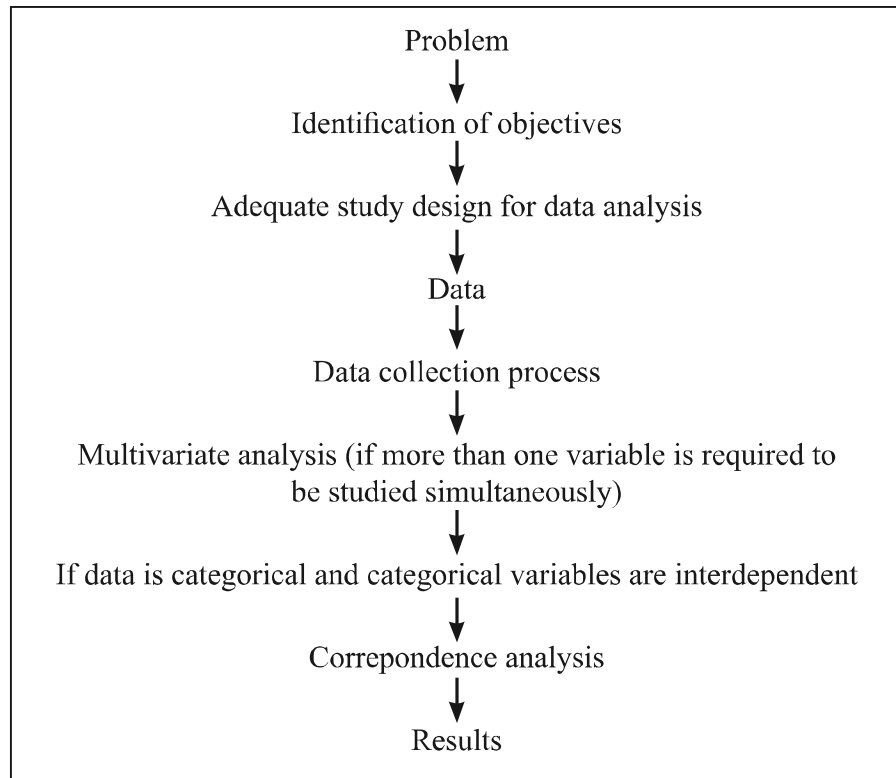


Fig 3: Workflow for correspondence analysis

REFERENCES

Benzécri, J. P. (1973), *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. Paris, France: Dunod

Other Suggested Readings

Greenacre, Michael (2007), *Correspondence Analysis in Practice, Second Edition*. London: Chapman & Hall/CRC.

Hirschfeld, H.O. (1935), A connection between correlation and contingency. *Proc. Cambridge Philosophical Society*, 31, 520–524.

Härdle W. and Simar L. (2007), *Applied Multivariate Statistical Analysis, Second edition*. Berlin: Springer-Verlag.

Hair, J., W. Black, R. Anderson and B. Babib (2018), *Multivariate Data Analysis, eighth edition*. London: Cengage Learning EMEA

Härdle, Wolfgang Karl & Hlávka, Zdenek (2007), *Multivariate statistics: Exercises and solutions*. New York: Springer-Verlag. doi: 10.1007/978-0-387-73508-5.

<http://www.sthda.com/english/rpkgs/factoextra>

PART II

REGRESSION ANALYSIS

Chapter 6

LINEAR AND NON-LINEAR REGRESSION ANALYSIS

Ranjit Kumar Paul and L. M. Bhar

INTRODUCTION

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

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

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

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

$$Y = f(X)$$

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

A regression model that has more than one regressor variables is called a multiple regression model (www.iasri.res.in). In other words, it is a linear relationship between a dependent variable and a group of independent variables. Multiple regression fits a model to predict a dependent (Y) variable from two or more independent (X) variables. Multiple linear regression models are often used as approximating functions. That is, true functional relationship between y and, \dots , is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate estimation to the true unknown function. If the model fits the data well, the overall R^2 value will be high, and the corresponding P value will be low (P value is the observed significance level

at which the null hypothesis is rejected). In addition to the overall P value, multiple regressions also report an individual P value for each independent variable. A low P value here means that this particular independent variable significantly improves the fit of the model. It is computed by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted. If the fit is much worse when that variable is removed from the model, the P value will be low, indicating that the variable has a significant impact on the model.

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

LINEAR REGRESSION MODELS

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

$$Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_i + \hat{\alpha}_1 \dots \quad (1)$$

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

A regression model (1) is defined to be simple and linear in terms of parameters, and linear in the predictor variable. It is “simple” which means there is only one predictor variable, “linear in the parameters” as there are no parameters that is presented in the form of an exponent or its multiplied or divided by another parameter, and “linear in predictor variable” since its variable appears only in the order of first power. A model that is linear in the parameters and in the predictor variable is also called first order model (Belsley *et al.*, 2004).

Regression Parameters

The parameters β_0 and β_1 in regression model (1) called as a regression coefficients, β_1 is identified as the slope of the regression line. It simply implies the change in the mean of Y per unit increase in X . The parameter β_0 is the intercept of the regression line. When the scope of the model gives $X = 0$, β_0 gives the mean of the probability

distribution of Y at $X = 0$. When the scope of the model does not cover this $X = 0$, β_0 does not show any special meaning as a separate term in the regression model.

Method of Ordinary Least Squares

To evaluate the regression parameters β_0 and β_1 , we make use of the method of least squares for the estimation. Using every observation (X_i, Y_i) for the each case, we calculate the deviation of Y from its expected value using the method of least squares, $Y_i - \hat{a}_0 - \hat{a}_1 X_i$. To be specific, the main requirement is to take into consideration sum of the n squared deviations. This criterion can be denoted by Q :

$$Q = \sum_{i=1}^n (Y_i - \hat{a}_0 - \hat{a}_1 X_i)^2 \quad \dots \quad (2)$$

Taking into consideration the definition of method of least squares, β_0 and β_1 are the estimators whose values are b_0 and b_1 , respectively, that can minimize the criterion Q for the observations provided.

Taking into consideration, the analytical approach, it can be shown for regression model (1) that b_0 and b_1 are the values that minimizes Q for any particular set of data are given by the following equations:

$$\begin{aligned} \sum_{i=1}^n Y_i &= nb_0 + b_1 \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i Y_i &= b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 \quad \dots \end{aligned} \quad (3)$$

The two equations that are written above are known as normal equations and can be solved for b_0 and b_1 as:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \dots \\ b_0 &= \frac{1}{n} \left(\sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i \right) = \bar{Y} - b_1 \bar{X} \end{aligned} \quad (4)$$

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

Inferences in Linear Models

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

$$H_0 = \beta_1 = 0$$

$$H_1 = \beta_1 \neq 0$$

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

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

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

where $s(b_1)$ is standard error of b_1 and can be calculated as

$$s(b_1) = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}} \quad \dots \quad (5)$$

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

$$\text{if } |t| \leq t(1 - \alpha/2; n - p), \quad \text{conclude } H_0,$$

$$\text{if } |t| > t(1 - \alpha/2; n - p), \quad \text{conclude } H_1.$$

Similarly testing for other parameters can be carried out.

Measure of Fitting, R^2

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

Represented by $SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$, total sum of squares which computes the variation

in the observation Y_i , or the unpredictability in predicting the value of Y , when no account of the predictor variable X is taken in consideration. Thus, SSTO is a way of measuring of unpredictability in predicting Y when X is not considered. Similarly, SSE (Error Sum Square) gives the variation in the Y_i when a regression model utilizing the predictor variable X is employed. A natural measure of the effect of X in reducing the variation in Y , i.e., in reducing the uncertainty in predicting Y , is to express the reduction in variation ($SSTO - SSE = SSR$ (Regression Error Sum Square)) of as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad \dots \quad (6)$$

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

Variable Selection Techniques

Forward selection procedure

Forward Selection procedure makes use of a subset of predictor variables for the final model. To do forward stepwise in context of linear regression whether n is less than p or n is greater than p . Steps to be followed are:

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

To summarize, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backward elimination procedure

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

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

Stepwise selection procedure

The step wise regression procedure starts of by choosing an equation containing the single best X variable and then attempts to build up with subsequent additions of X 's

one at a time as long as these additions are worthwhile. The order of addition is decided by using the partial F-test values to select which variable should enter next. The highest partial F-value is compared to a (selected or default) F-to-enter value. When variable has been added, the equation is examined to see if any variable should be deleted.

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

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

Diagnostics and Alternative Measures

When a regression model is selected to be applied to an application, we can't assert in advance that the model is appropriate for that particular application, any one, or several other, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to determine the aptness of the model for the dataset before inferences based on that particular model is undertaken. We should take into consideration following six important types of departures from linear regression model with normal errors:

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

NON-LINEAR REGRESSION MODELS

It is not always possible to make use of linear regression model. For example, the engineer or scientist might assume the form of the relationship between the response variable and the regressors from his direct knowledge, perhaps from the theory

underlying the phenomena. In actuality, relationship between the response and the regressors may be a differential equation, or the solution to a differential equation. Often, this will lead to a model of nonlinear form.

Any model that is not linear in the unknown parameters is a non-linear regression model. For example, the model

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

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

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

Where θ is a $p \times 1$ vector of unknown parameters, and ε is an uncorrelated random error term with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. It is assumed that the errors are normally distributed, as in linear regression.

The function $f(x, \theta)$ is called the expectation function for the nonlinear regression model. This is very similar to the linear regression case, except that now the expectation function is a non-linear function of the parameters.

Inside the non-linear regression model, for at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. In linear regression, these derivatives do not behave as functions of the unknown parameters. To illustrate these points, consider a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

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

Now

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

Where $x_0 \equiv 1$. Consider that in the linear case the derivatives are not functions of the β 's.

Now consider this non-linear model

$$\begin{aligned} y &= f(x, \theta) + \varepsilon \\ &= \theta_1 e^{\theta_2 x} + \varepsilon \end{aligned}$$

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

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

Since the derivatives are functions of the unknown parameters θ_1 and θ_2 , the model is non-linear in nature.

Fitting of Nonlinear Models

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

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

The details of these methods are given in Draper and Smith (1998), Chatterjee and Price (1977), Kleinbaum and Kupper (1978), Montgomery *et al.* (2003). The results of linear least square theory in a succession of stages uses in the linearization method uses. Although, neither this method nor the Steepest descent method, is ideal. The latter method is able to converge on true parameter values even though initial trial values are far from the true parameter values, but this convergence tends to be very slow at the later stages of the iterative process. On the other side of it, the linearization method will converge very swiftly provided the vicinity of the true parameter values has been reached, but if initial trial values are too far then convergence may not occur at all.

The most frequently used method of computing nonlinear least squares estimators is the Levenberg-Marquardt's method. This method illustrates a compromise between the other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not 'slow down' at the latter part of the iterative process. The procedure is available in standard statistical software packages.

ILLUSTRATION

Practical on Regression Analysis using R

Example: An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria* (Mol) Standl) Cv. Pusa Hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination and hand pollination under field conditions. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data were collected on number of fruit set (NFS) for the period of 45 days, fruit weight (FW) (kg), seed yield per plant (SY)(g) and seedling length (SL) (cm). The data obtained is as given below: {Here 1 denotes natural pollination

and 2 denotes the hand pollination}

Group	NFS	FW	SY	SL
1	7.0	1.85	147.70	16.86
1	7.0	1.86	136.86	16.77
1	6.0	1.83	149.97	16.35
1	7.0	1.89	172.33	18.26
1	7.0	1.80	144.46	17.90
1	6.0	1.88	138.30	16.95
1	7.0	1.89	150.58	18.15
1	7.0	1.79	140.99	18.86
1	6.0	1.85	140.57	18.39
1	7.0	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.50	18.07
2	7.3	2.58	230.34	19.07
2	8.0	2.62	217.05	19.00
2	8.0	2.68	233.84	18.00
2	8.0	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7.0	2.45	199.87	19.31
2	7.3	2.44	214.30	19.36

#importing the csv file

```
FruitData <-read.csv(file.choose(),header=TRUE)
```

```
attach(FruitData)
```

#finding correlation coefficient matrix from the variables in a dataset

```
cor(FruitData)
```

#estimating and testing correlation coefficient between any two variables say between SY and SL

```
cor.test(SY, SL)
```

#Finding Spearman's Ran correlation coefficient between any variables

```
cor.test(X, Y, method="s")
```

#partial correlating coefficient

```
library(ppcor)
```

```

pcor(FruitData, method = c("pearson"))
#regression model of SY on SL
reg<-lm(SY ~SL)
summary(reg)
#regression model of SY on SL, FW and NFS
reg1<-lm(SY ~SL+FW+NFS)
summary(reg1)
#regression model of SY on SL without intercept
reg<-lm(SY ~-1+SL)
#prediction using regression model
lm.predict<-predict(reg ,interval="confidence")
#Residual diagnostics of fitted model
par(mfrow = c(2, 2))
plot(reg)
#For plotting cook's distance
plot(reg, 4)
#distribution of fitting in r
library(MASS)
fitdistr(SY,"normal")
library(vcd)## loading vcd package
gf<-goodfit(SY,type= "normal",method= "MinChisq")
summary(gf)
library(olsrr)
# Fit the full model
full.model <- lm(SY ~., data = FruitData)
# Fit the model with stepwise selection procedure
ols_step_both_p(full.model)
# Fitting of robust regression model
library(MASS)
reg1<-rlm(SY~ SL+FW+NFS)
summary(reg1)
library(car)

```

```
vif(full.model)
#Plotting Residual QQ Plot
ols_plot_resid_qq(full.model)
#Plotting Residual Normality Test
ols_test_normality(full.model)
# Plotting Residual vs Fitted Values Plot
ols_plot_resid_fit(full.model)
#Plotting Residual Histogram
ols_plot_resid_hist(full.model)
```

R code for nonlinear least square

Let us assume that the relationship between y and x is given as:

$$y = a * \exp(b * x) + \varepsilon$$

To fit the model in r, the code is:

```
nonlin_mod<-nls(y~a*exp(b*x),start=list(a=13,b=0.1)) # the starting values a and b
are assumed to be 13 and 0.1 respectively
```

REFERENCES

- Draper, N. R. and H. Smith (1998), *Applied Regression analysis*, New York: Wiley Eastern Ltd.
- Barnett, V. and T. Lewis (1984), *Outliers in Statistical Data*, New York: Wiley Ltd.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (2004). *Regression diagnostics – Identifying influential data and sources of collinearity*, New York.: Wiley.
- Bhar, L. (2015), *Regression Analysis: Diagnostics and Remedial Measures*, pp 245-270. In: Rajni Jain and S S Raju Edition. *Decision support system in agriculture using quantitative analysis*, Agrotech Publishing Academy, ISBN: 978-81-8321-395-0.
- Chatterjee, S. and B. Price (1977), *Regression analysis by example*, New York: John Wiley & sons.
- Khashei (2011), *A New Hybrid Methodology for Nonlinear Time Series Forecasting*
- Kleinbaum, D. G. and L. L. Kupper (1978), *Applied Regression analysis and other multivariate methods*, Massachusetts: Duxbury Press.
- Montgomery, D. C., E. Peck and G. Vining (2003), *Introduction to linear regression analysis*, 3rd Edition, New York: John Wiley and Sons Inc. www.iasri.res.in

Chapter 7

QUALITATIVE REGRESSION MODEL (LOGIT, PROBIT, TOBIT)

Shivaswamy G. P., K. N. Singh and Anuja A. R.

INTRODUCTION

Regression analysis is a statistical method of studying functional relationship between a dependent or response variable and one or more independent or explanatory variables. In classical linear regression models (CLRM), response variable is implicitly assumed as quantitative and explanatory variables can be quantitative or qualitative. Parameter estimation in the CLRM is based on important assumptions such as linearity of model in parameters, though response and explanatory variables may or may not be linearly related; explanatory variables are independent of the error terms; independent and identically distributed error terms with zero mean and constant variance; and equal reliability of observations. However, when the response variable is qualitative, these basic assumptions of CLRM may not hold. For instance, one wants to study the adoption of high yielding varieties of a crop which is a binary response variable. It takes only two values: 1 if the variety is adopted and 0 if it is not. There are several instances, where the response variable is binary. Suppose one wants to study the determinants of access to institutional credit such as age, gender, education, social status, land holding etc. Whether a farmer has access to institutional credit is a binary variable taking values 0 or 1, 0 meaning no access and 1 meaning access to institutional credit. Similarly, other examples can be crop diversification status as a function of various quantitative or qualitative independent variables. In such cases, error terms are not normality distributed and the variance is not constant thereby violating the homoscedasticity assumption. The statistical models preferred for the analysis of such a binary response are logit and probit models as these models do not make assumptions on the distribution of explanatory variables.

LOGISTIC REGRESSION

Logistic regression analysis is used when the dependent variable is qualitative and normality assumption is not satisfied. Cox (1958) developed this model. Logistic regression is appropriate when the dependent and independent variables are non-linearly related. Logit is transformation of logistic regression to make it linear.

In case of logit transformation, binary variable (Koutsoyiannis, 2001) of adoption of a rice variety, the decision to adopt or not to adopt by i^{th} individual depends on the latent

variable Y_i^* which in turn depends on explanatory variables such as age, education, farm size, access to institutional credit, irrigation facility and training.

$$Y_i^* = BX + u_i \dots \quad (1)$$

where, B vector of parameters, X, vector of explanatory variables; and u_i error term of i^{th} individual

It is assumed that $Y_i = \begin{cases} 0 & \text{if } Y_i^* \leq 0 \\ 1 & \text{if } Y_i^* > 0 \end{cases}$

That is if individual's utility index exceeds Y_i^* farmer will adopt a variety, but if it is less than Y_i^* then variety is not adopted.

In case of probability of adoption of variety ($Y=1$), the logistic regression function can be expressed as

$$P_i = \Pr(Y_i = 1 | X = x) = \frac{1}{(1 + e^{-Z_i})} \dots \quad (2)$$

where, $Z_i = BX + u_i$

The probability that variety is not adopted ($Y=0$) is given by

$$(1 - P_i) = \frac{1}{(1 + e^{Z_i})}$$

where, as z_i ranges from $-\infty$ to $+\infty$, P_i ranges between 0 and 1. And, the model is nonlinear both in response variables X and parameters Bs.

Further, to make the logistic regression function linear in the parameters, we take the ratio of probability that farmer adopts a variety to probability that he is not;

$$\begin{aligned} \frac{P_i}{1 - P_i} &= \frac{\frac{1}{(1 + e^{-Z_i})}}{\frac{1}{(1 + e^{Z_i})}} \\ \frac{P_i}{1 - P_i} &= e^{Z_i} \dots \end{aligned} \quad (3)$$

where, $P_i/(1 - P_i)$ is known as the odds ratio in favor of adoption of a variety i.e. the ratio of probability that a farmer adopts a variety to probability that he does not adopt.

Equation can be transformed by taking natural logarithm as follows

$$\ln\left(\frac{P_i}{1 - P_i}\right) = Z_i \dots \quad (4)$$

Log of the odds ratio is known as the logit which is nothing but a linear transformation of the logistic regression model.

Estimation of logit model

Usual OLS method cannot be used to estimate the logit model despite its linearity properties due to problem of undefined expressions. Rather, Maximum likelihood estimation method is used for estimation.

Variables used for the logit analysis of determinants of variety adoption example are as follows

Adoption=1 for adopters and 0 for non-adopters

Age in years

Education=1 if educated; 0 otherwise Credit=1 if there is access to institutional credit; 0 other wise

Irrigation=1 if there is access to irrigation; 0 for non-access

Training=1 if undergone training; 0 otherwise

Table 1: Sample data set for logit and probit analysis

Observations	Adoption	Age	Education	Farm size	Credit	Irrigation	Training
1	0	70	1	9.01	0	0	0
2	0	30	0	2.136	1	1	0
3	1	40	1	1.12	1	1	0
4	1	60	0	1.003	0	1	0
5	0	30	1	2.61	1	0	1
6	0	60	1	1.23	0	1	1
7	1	45	1	2.434	1	0	0
.							
.							
.							
1763	1	52	1	1.705	1	0	0

Table 2 provides the results of logit model for the adoption of variety example, which are obtained by *STATA* using the command *logit*.

Table 2: Logit estimates of adoption of a rice variety

Particulars	Coefficient	Standard error	Z statistic	Prob>Z
Age	0.002	0.003	0.06	0.951
Education	-0.070	0.102	-0.07	0.945
Farm size	-0.014	0.015	-0.92	0.36

Qualitative Regression Model (Logit, Probit, Tobit)

Particulars	Coefficient	Standard error	Z statistic	Prob>Z
Institutional credit	0.489	0.098	4.98	0
Irrigation access	0.299	0.097	3.07	0.002
Training	0.096	0.285	0.34	0.735
Constant	-0.351	0.213	-1.64	0.1
Number of observations	1763			
McFadden R ²	0.014			

The results of logit model show that access to institutional credit and irrigation are statistically significant at 1 percent level of significance. It is interpreted as access to institutional credit increases the average logit value by 0.489. Access to irrigation is also interpreted similarly. Other variables such as age, education and training are statistically insignificant meaning they do not have visible impact on adoption of a variety.

In case of CLRM, R² indicates the goodness of fit showing the proportion of variation in the dependent variable explained by the independent variables in the model. But, in case of binary regression models, R² is not meaningful for which McFadden R² or pseudo R² is discussed in the literature. The value of McFadden R² ranges between 0 and 1. In our example its value is 0.014. It should be noted that in qualitative regression models, the expected sign of the regression coefficients and their statistical significance are more important than the goodness of fit measures.

We can express the logit coefficients in terms of odds ratio (Table 3) by using following *STATA* command *logit adoption age education farmsize credit irrigation access training*.

Table 3: Odds ratio for adoption versus non-adoption

	Odds ratio	Standard error	Z statistic	Prob>Z
Age	1.000	0.004	0.060	0.951
Education	0.993	0.102	-0.070	0.945
Farmsize	0.986	0.015	-0.920	0.360
Institutional credit	1.632	0.161	4.980	0.000
Irrigation	1.349	0.131	3.070	0.002
Training	1.101	0.315	0.340	0.735
Constant	0.704	0.151	-1.640	0.100
Number of observations	1763			
McFadden R ²	0.014			

The odds ratios are obtained by taking the exponential of logit coefficients given in Table 2. The interpretation of the odds ratio depends on whether its value is greater than 1 or less than 1. Odds ratios of greater than 1 indicate the increased chance of adoption as compared to non-adoption. On the other hand, odds ratio of less than 1 indicates the decreased chance of adoption. Odds ratio of 1 suggests that chances of adopting and not adopting are even. In our example, two variables institutional credit and irrigation have the odds ratios of greater than 1 meaning increased chance of adopting a variety as against non-adoption.

Estimation of marginal effects

Marginal effects depict the marginal impact of one unit change in the explanatory variable on the probability of adoption of a variety. It is a way depicting the model estimates in terms of probabilities which helps in interpreting in terms of magnitude. Note that instead of computing the marginal effect for each independent variable on the probability of adoption, it is computed for the average values of variables. It is to be noted that for quantitative response variables marginal effect is the derivative ($\frac{dy}{dx}$) of dependent variable (y) with respect of independent variable (x) that is rate of change of y with respect to x. However for qualitative independent variable which takes the discrete values 0 and 1 as in our example, marginal effect is estimated for the discrete change in the qualitative variable from 0 to 1.

Table 4: Marginal effects of logit model

Particulars	Marginal effects	Standard error	Z statistic	Prob>Z	Mean value
Age	0	0.001	0.060	0.951	50.558
Education	-0.001	0.026	-0.070	0.945	0.647
Farmsize	-0.003	0.004	-0.920	0.360	2.771
Institutional credit	0.121	0.024	5.030	0.000	0.450
Irrigation	0.074	0.024	3.080	0.002	0.573
Training	0.024	0.071	0.340	0.735	0.029
Number of observations	1763				
McFadden R ²	0.014				

The interpretation of marginal effects in our example is that unit change in age and farm size does not have statistically significant impact on the rate of change in the probability of adoption. For qualitative variable, an access to institutional credit has significant positive impact in the probability of adoption of variety by about 0.121. Similarly access to irrigation increases the probability of adoption by about 0.074.

PROBIT MODEL

Like logit, probit model is used when the response variable is qualitative. Error terms in the probit model follow normal distribution.

For arriving at the probit model, equation 1 can be translated into,

$$\begin{aligned}
 Pr(Y_i = 1|X = x) &= Y_i^* > 0 \\
 &= Pr(u_i > -B'X) \\
 &= Pr\left(\frac{u_i}{\sigma} > \frac{-B'X}{\sigma}\right) \quad \dots \quad (5) \\
 Pr(Y_i = 1|X = x) &= \Phi\left(\frac{-B'X}{\sigma}\right)
 \end{aligned}$$

Estimation of probit model

Probit model is estimated based on the maximum likelihood function which finds coefficients that maximize the probability of $Y_i = 1$.

Using *STATA* command *probit*, ML estimates of the probit model for adoption of variety are given in Table 5 (Spermann, 2008).

Table 5: Probit estimates of adoption of variety

	Coefficient	Standard error	Z statistic	Prob>Z
Age	0.000	0.002	0.070	0.943
Education	-0.004	0.064	-0.060	0.951
Farm size	-0.009	0.010	-0.930	0.354
Institutional credit	0.306	0.061	4.990	0.000
Irrigation	0.187	0.061	3.070	0.002
Training	0.061	0.178	0.340	0.730
Constant	-0.221	0.134	-1.650	0.098
Number of observations	1763			
McFadden R ²	0.014			

Although coefficients of logit and probit models are different, the interpretation of coefficients is similar. Institutional credit and irrigation are statistically significant at 1 percent level of significance. It should be noted that only sign of the logit and probit models are interpreted but not the magnitude. The coefficients of logit and probit models are different and can be comparable after multiplying probit coefficients by about 1.81. However, marginal effects of probit and logit models are similar (Table

6). Logit and probit functions are almost similar with both s shaped curves. The main difference between the logit and probit models is that the logistic distribution has slightly flatter tails. Therefore, there is no compelling reason for choosing one model over another (Halloran, 2018).

Table 6: Marginal effects of probit model

	Marginal effects	Standard error	Z statistic	Prob>Z	Mean value
Age	0	0.001	0.060	0.951	50.558
Education	-0.001	0.026	-0.070	0.945	0.647
Farmsize	-0.003	0.004	-0.920	0.360	2.771
Institutional credit	0.121	0.024	5.030	0.000	0.450
Irrigation	0.074	0.024	3.080	0.002	0.573
Training	0.024	0.071	0.340	0.735	0.029
Number of observations	1763				
McFadden R ²	0.014				

TOBIT MODEL

Tobit model, also called as censored regression model or limited dependent variable regression model, was proposed by Tobin in 1958. A censored sample is a sample in which information on dependent variable is available for only some observations in a sample. If we use OLS on censored data set, estimates obtained will be inconsistent meaning coefficients will not necessarily approach the true population parameters as sample size increases (Gujarati, 2003). In such cases, Tobit model is used for analyzing censored sample.

The model can be expressed as

$$Y = X\beta + u \quad \text{If } \beta'X + u > 0;$$

$$= 0 \quad \text{Otherwise}$$

Such that the residual, $u \sim N(0, \sigma^2)$

where Y , (nx1) vector of dependent variable; β (kX1) vector of unknown parameters; and X vector of exogenous variables.

The model can be estimated using maximum likelihood method or Heckman two step procedure.

The application of the model can be explained with the help of labour economics example. Using the data set which contains information on both working and non-

working married women, suppose we want to estimate the extent of participation of working women in labour force. Here the dependent variable (extent of participation in hours per year) is continuous and lower limit of dependent variable is zero which means non-participation in labour market. Tobit model can be applied here as it employs all information collected for both working women (where dependent variable is more than zero) and non-working women (dependent variable -zero).

In another example, suppose we want to estimate the amount of money spent by an individual on meat in relation to socio economic variables. Now there are two groups of consumers. One set m_1 about whom we have information on the independent variables (age, education, income etc.) as well as the dependent variable (the amount of money spent on meat items) and another set m_2 about whom we have information only on the independent variables but not on dependent variable. Here we cannot neglect observations on second group as the OLS estimates of parameters using only first group of observations will be biased and inconsistent. In this case we can use Tobit model for a better estimation of parameters.

REFERENCES

- Cox, D. R. (1958), The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society*, B, 20: 215-242
- Gujarati, D. N. (2003), Basic Econometrics. 4th Edition, New York: Mc Graw Hill Publications
- Halloran, S. (2018), Logit and probit models, accessed in September 2019 http://www.columbia.edu/~so33/SusDev/Lecture_9.pdf
- Koutsoyiannis, A. (2001), Theory of Econometrics, (2nd ed.), New York: Palgrave Macmillan Limited
- Spermann, A. (2009), The probit model, accessed in September 2019 https://www.empiwifo.uni-freiburg.de/lehre-teaching-1/summer-term-09/materials-microeconometrics/probit_7-5_09.pdf.

Chapter 8

INTRODUCTION TO PANEL DATA REGRESSION MODELS

Ravindra Singh Shekhawat, K. N. Singh, Achal Lama and Bishal Gurung

INTRODUCTION

Different types of data are generally available for empirical analysis, namely, time series, cross section, and panel.

Cross section data: Cross-sectional data in statistics and econometrics is a type of data collected by observing many subjects (such as individuals, firms, countries, states or regions) at the one point or period of time. In cross-section data, values of one or more variables are collected for several sample units, or entities, at the same point of time. For examples, number of tractor per thousand hectare of net sown area in 2012 across the major states of India. The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among selected subjects.

Time series data: A time series is a series of data points that are indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. A data set containing observations on a single phenomenon observed over multiple time periods is called time series (e.g., GDP for several quarters or years).

In time series data, both the values and the ordering of the data points have meaning. Examples of time series are price of wheat from 1990 to 2019, agricultural production from 1966 to 2018 and number of tractor per thousand hectare of net sown area from 1982 to 2012.

PANEL DATA

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data. In panel data the same cross-sectional units (say a family or a firm or a state) is surveyed over time. In short, panel data have space as well as time dimensions. Panel data is combination of time series and cross sectional data. It is also called longitudinal data which means a study over time of a variable or group of subjects (Frees, 2004).

Panel data gives more variability, more information, more efficiency and more degrees of freedom compared to the time series data or cross-section data. The regression models based on such panel data are known as panel data regression models.

Let us consider a data set on number of tractor across the major states of India from 1982 to 2012. For any given year, the data on number of tractor in various states represent a cross-sectional sample. For any given state, there are thirty-time series observations on number of tractor. Thus, we have in all (panel) observations on number of tractor

Table 1: Observation of cross sectional sample

Year	State	Tractor density ('000 ha of net sown area)	Average land holding (ha)	CI (%)	Real wage rate (Rs per day)
2000	Punjab	98.0	4.03	186.85	121.40
2001	Punjab	102.0	4.02	186.67	128.74
2002	Punjab	105.3	4.00	185.03	132.35
2003	Punjab	106.3	3.99	186.49	131.80
2000	MP	13.5	1.66	121.86	73.56
2001	MP	13.5	1.62	128.16	71.36
2002	MP	16.0	1.59	124.52	73.42
2003	MP	17.0	1.54	132.41	72.24

Some other examples:

- Data on yield of rice in 42 villages from 2013 to 2017, for 210 observations total.
- Data on crime rate in 17 Indian states, each state is observed in 6 years, for a total of 102 observations.
- Data on income of 1000 individuals, in four different months, for 4000 observations total.

Importance of panel data

1. It can take heterogeneity explicitly in analysis by allowing individual specific variables.
2. It provides more variability, more information, more degree of freedom, more efficiency and less collinearity among variables.
3. It is more suited to analyses dynamics of changes like mobility of labour forces, unemployment.
4. It helps to study complicated behavioral models like agricultural technology change and economies of scale.
5. It can better detect and measure impact of particular technology changes. For example, impact of tractorization on cropping intensity and farm productivity in household data over the year.
6. More accurate inference of model parameters. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional

data which may be viewed as a panel with $T = 1$ (T is the number of time series), or time series data which is a panel with $N = 1$ (N is the number of cross section), hence improving the efficiency of econometric estimates.

7. By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.

Panel model

A normal panel model can be represented as follow-

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_n X_{nit} + u_{it}$$

where,

$i = i^{\text{th}}$ cross-sectional unit

$t = t^{\text{th}}$ time period

Y_{it} = dependent variable

X 's= explanatory variables

$$E(u_{it}) \sim N(0, \sigma^2)$$

ESTIMATION OF PANEL DATA REGRESSION MODELS

There are two major approach for estimation of a panel model which are

1. Fixed effect panel regression model
2. Random effect panel regression model

Fixed effect panel regression model

For estimation of panel regression model, here we assume that intercepts and slope coefficients are different for all cross section units. It means to say that all the cross sectional unit have different regression equation or function. Therefore, we used the dummy variables to account for individual effect, we can allow for time effect in the sense that the panel regression function shifts over time. For such a situation we introduce time dummies, one for each year. As, it is known that interactive, or differential, slope dummies can account for differences in slope coefficients of panel model. Therefore, we multiply each of the cross sectional dummy by each of the X variables. Fixed effect panel regression model is-

$$Y_{it} = \alpha_0 + \alpha_1 D_{1i} + \cdots \alpha_m D_{mi} + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_n X_{nit} + \gamma_1 (D_{1i} X_{1it}) + \gamma_2 (D_{1i} X_{2it}) + \cdots + u_{it}$$

where,

$i = i^{\text{th}}$ cross-sectional unit

$t = t^{\text{th}}$ time period

n = number of explanatory variable

m = number of cross sectional unit

Y_{it} = dependent variable

X 's = explanatory variables

$E(u_{it}) \sim N(0, \sigma^2)$

D_m = cross sectional or state dummies variables

α_0 = intercept of a particular cross sectional unit

α_m = differential intercept coefficients, tell by how much the intercepts of other cross section differ from the intercept of α_0 .

γ 's = time dummies

Cautions in use of Fixed Effect Panel model (FEM) or Least Square Dummy Variable (LSDV) model

1. If we introduce too many dummy variables we will run up against the degrees of freedom problem (higher the number of dummy variables the lesser the degree of freedom).
2. With so many variables in the model, there is always the possibility of multicollinearity, which might make precise estimation of one or more parameters difficult.
3. Suppose, in the FEM if variables such as gender, colour, and ethnicity, which are time invariant are also included, the LSDV approach may not be able to identify the impact of such time-invariant variables.

Random effect approach

Although fixed effects or LSDV model can be expensive in terms of degrees of freedom if we have several cross-sectional units.

If the dummy variables do in fact represent a lack of knowledge about the (true) model, why not express this ignorance through the disturbance term? This is precisely the approach suggested by the proponents of the so called error components model (ECM) or random effects model (REM).

Let's start from basic model

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_n X_{nit} + u_{it} \dots \quad (1)$$

Instead of treating β_{0i} as fixed, we assume that it is a random variable with a mean value of β_0 . Therefore, intercept value for an individual state can be expressed as-

$$\beta_{0i} = \beta_0 + \varepsilon_i \dots \quad (2)$$

where,

$$i = 1, 2 \dots, N$$

ε_i = random error term with a mean value of zero and σ_ε variance

Here, we have a common mean value for the intercept ($= \beta_0$) and the individual differences in the intercept values of each cross sectional unit are reflected in the random error term ε_i

From equation 1 and 2

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_n X_{nit} + u_{it} + \varepsilon_i$$

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_n X_{nit} + \omega_{it} \dots \quad (3)$$

where,

Composite error term ($(\omega_{it}) = u_{it} + \varepsilon_i$

The composite error term consists of two components, the cross-section or individual-specific error component and the combined time series and cross-section error component (Gujarati and Sangeetha, 2007).

The usual assumptions made by ECM are that

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$u_{it} \sim N(0, \sigma_u^2)$$

$$E(\varepsilon_i u_{it}) = 0 \quad E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)$$

$$E(u_{it} u_{is}) = E(u_{it} u_{jt}) = E(u_{it} u_{js}) = 0 \quad (i \neq j; t \neq s)$$

i.e. the individual error components are not correlated with each other and are not autocorrelated across both cross-section and time series units.

In fixed effect model each cross sectional unit has its own intercept value, while ECM has only single intercept which represents the mean value of all cross sectional intercepts and random error (ε_i) component represents random deviation of the individual intercept from the mean value but random error (ε_i) is not directly observable as it is an unobservable or latent variable.

Composite error follows

$$E(\omega_{it}) = 0$$

$$Var(\omega_{it}) = \sigma_\varepsilon^2 + \sigma_u^2$$

the composite error term ω_{it} is homoscedastic. However, it can be shown that ω_{it} and ω_{is} are correlated i.e. the error terms of a given cross-sectional unit at two different points in time are correlated. The correlation coefficient is as follows:

$$\text{corr}(\omega_{it}, \omega_{is}) = \frac{\sigma_{\varepsilon}^2}{\sigma_{\varepsilon}^2 + \sigma_u^2}$$

For any given cross-sectional unit, the above correlation coefficient between two different times remains same no matter how far apart the two time periods are and always remain same for all cross sectional unit i.e. identical for all individuals. If we do not take this correlation structure into account, and estimate (eq. 3) by OLS, the resulting estimators will be inefficient.

The most appropriate method is GLS or transformed GLS instead of OLS.

Fixed Effects (LSDV) versus Random Effects Model

- Which model is better and give efficient output, FEM or ECM? The answer depends on the assumption made about the likely correlation between the individual, or cross-section, error components and the X's independent variables.
- If it is assumed that the random error component and the X's are uncorrelated, ECM may be appropriate, whereas if they are correlated, FEM may be appropriate.
- If T (the number of time series data) is large and N (the number of cross-sectional units) is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.
- When N is large and T is small, the estimates obtained by the two methods can differ significantly.
- If the individual error component and one or more independent variable are correlated, then the ECM estimators are biased, whereas those obtained from FEM are unbiased.
- If N is large and T is small, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.

Hausman test: A formal test given by Hausman, 1978 that will help us to choose between FEM and ECM.

H_0 : Random effect model (ECM) is appropriate or the FEM and ECM estimators do not differ substantially

H_1 : Fixed effect model (FEM) is appropriate

If the null hypothesis is rejected, the conclusion is that ECM is not appropriate and that we may be better off using FEM.

Note: This is a reminder that panel data regression models may not be appropriate in each situation despite the availability of both time series and cross sectional data.

ILLUSTRATION

We are having a data set related to number of tractor in thousand hectare of net sown area, (tractor density) cropping intensity and average land holding of 14 major state of India from 2009 to 2012. We want to know the factors affecting tractor density. There may be other variables also which affect number of tractor but here we are taking two variables only for understanding the panel data regression model.

Table 2: Illustration

State	Year	Tractor density	CI (%)	Average land holding (ha)
AP	2009	24.25	125.71	1.11
AP	2010	23.38	129.73	1.08
AP	2011	26.2	123.28	1.06
AP	2012	30.8	122.78	1.05
Bihar	2009	26.31	136.83	0.4
Bihar	2010	30.04	135.94	0.39
Bihar	2011	33.2	141.72	0.39
Bihar	2012	37.21	143.98	0.38
Gujarat	2009	37.56	108.11	2.07
Gujarat	2010	39.85	118.88	2.03
Gujarat	2011	42.98	127.09	2.01
Gujarat	2012	48.06	122.31	1.99
Haryana	2009	133.36	178.9	2.24
Haryana	2010	139.52	184.91	2.25
Haryana	2011	138.72	184.71	2.25
Haryana	2012	147.07	181.5	2.25
Karnataka	2009	14.7	123.73	1.57
Karnataka	2010	30.3	124.13	1.55
Karnataka	2011	34.36	121.31	1.54
Karnataka	2012	37.17	119.96	1.53
Kerala	2009	5.12	128.38	0.22
Kerala	2010	5.15	127.75	0.22
Kerala	2011	5.49	130.49	0.22
Kerala	2012	5.67	126.56	0.22
MP	2009	25.48	143.01	1.44
MP	2010	27.03	145.82	1.44
MP	2011	29.39	147.77	1.44
MP	2012	32.09	150.66	1.43

State	Year	Tractor density	CI (%)	Average land holding (ha)
MH	2009	17.37	129.95	1.83
MH	2010	19.06	133.14	1.78
MH	2011	21.34	126.6	1.75
MH	2012	24.17	125.96	1.73
Odisha	2009	10.29	163.38	1.06
Odisha	2010	13.74	115.95	1.04
Odisha	2011	16.94	112.97	1.03
Odisha	2012	18.94	115.57	1.02
Punjab	2009	117.86	188.89	3.81
Punjab	2010	119.66	189.59	3.77
Punjab	2011	125.22	191.22	3.75
Punjab	2012	124.76	189.64	3.73
Rajasthan	2009	33.57	128.11	3.14
Rajasthan	2010	33	141.71	3.07
Rajasthan	2011	35.73	135.88	3.04
Rajasthan	2012	40.04	137.04	3
Tamil Nadu	2009	28.17	113.9	0.81
Tamil Nadu	2010	30.37	116.13	0.8
Tamil Nadu	2011	33.51	118.13	0.8
Tamil Nadu	2012	41.08	113.12	0.79
Uttar Pradesh	2009	53.37	153.35	0.77
Uttar Pradesh	2010	57.45	152.88	0.76
Uttar Pradesh	2011	58.95	156.04	0.76
Uttar Pradesh	2012	63.85	155.89	0.75
West Bengal	2009	5.19	181.32	0.77
West Bengal	2010	5.14	177.31	0.77
West Bengal	2011	5.53	179.93	0.77
West Bengal	2012	6.82	185.94	0.76

Panel data regression analysis is to be done in STATA by using the following command:

- `egen new_id = group (State)`
- `xtset new_id Year, yearly`
panel variable: `new_id` (strongly balanced)
time variable: `Year`, 2009 to 2012
delta: 1 year

➤ *xtreg tr_no_NSA CI Avg_hol_ha, fe* (Fixed effect panel model or FEM)

```
. xtreg tr_no_NSA CI Avg_hol_ha, fe
```

```
Fixed-effects (within) regression      Number of obs      =          56
Group variable: new_id                Number of groups   =          14

R-sq:                                Obs per group:
    within = 0.3348                      min =           4
    between = 0.4141                     avg =          4.0
    overall = 0.4079                     max =           4

corr(u_i, Xb)  = -0.9726                F(2, 40)           =         10.07
                                           Prob > F           =         0.0003
```

tr_no_NSA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
CI	-.0543543	.0793262	-0.69	0.497	-.2146785	.1059699
Avg_hol_ha	-101.3762	22.87148	-4.43	0.000	-147.6011	-55.15117
_cons	201.4951	35.95237	5.60	0.000	128.8327	274.1576
sigma_u	132.54345					
sigma_e	3.8151352					
rho	.99917217	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(13, 40) = 186.03                Prob > F = 0.0000
```

➤ *estimate store fe*➤ *xtreg tr_no_NSA CI Avg_hol_ha, re* (random effect panel mode ECM)

```
. xtreg tr_no_NSA CI Avg_hol_ha, re
```

```
Random-effects GLS regression      Number of obs      =          56
Group variable: new_id            Number of groups   =          14

R-sq:                                Obs per group:
    within = 0.3298                      min =           4
    between = 0.4087                     avg =          4.0
    overall = 0.4025                     max =           4

corr(u_i, X)    = 0 (assumed)          Wald chi2(2)       =         2.01
                                           Prob > chi2        =         0.3661
```

tr_no_NSA	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
CI	.001365	.0973174	0.01	0.989	-.1893736	.1921035
Avg_hol_ha	12.80875	9.083362	1.41	0.159	-4.994309	30.61182
_cons	23.02672	20.68639	1.11	0.266	-17.51786	63.5713
sigma_u	27.476374					
sigma_e	3.8151352					
rho	.98108494	(fraction of variance due to u_i)				

➤ *estimate store re*

➤ *hausman fe re*

```
. hausman fe re
```

	Coefficients		(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
	(b) fe	(B) re		
CI	-.0543543	.001365	-.0557193	.
Avg_hol_ha	-101.3762	12.80875	-114.1849	20.9904

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

```
chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
        =      29.19
Prob>chi2 =      0.0000
(V_b-V_B is not positive definite)
```

From Hausman test, we rejected null hypothesis therefore, we will use fixed effect panel regression model.

REFERENCES

- Frees, E. (2004), Longitudinal and Panel Data, Cambridge University Press.
Gujarati, D. N. and Sangeetha (2007), Basic Econometrics, 4th Edition, Tata McGraw-Hill Edition.
Hausman, J. A. (1978), Specification Tests in Econometrics, *Econometrica*, 46: 1251-71.

SUGESTED READING

- Ahn, S. C. and P. Schmidt (1995), “Efficient Estimation of Models for Dynamic Panel Data”, *Journal of Econometrics*, 68, 5-27.
Arellano, M. (2003), Panel Data Econometrics, Oxford: Oxford University Press.
Balestra, P. and M. Nerlove (1966), “Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas”, *Econometrica*, 34, 585-612.
Baltagi, B. H. (2001), Econometric Analysis of Panel Data, Second edition, New York: Wiley.
Chamberlain, G. (1984), “Panel Data”, in Handbook of Econometrics Vol II, ed. by Z. Griliches and M. Intriligator, pp. 1247-1318. Amsterdam: North Holland.
Frees, E. (2004). Longitudinal and Panel Data, Cambridge University Press.
Gujarati, D. N. and Sangeetha (2007). Basic Econometrics, fourth edition, Tata McGraw-Hill Edition.
Hausman, J. A. (1978), “Specification Tests in Econometrics”, *Econometrica*, 46. 1251-71.
Hsiao, C. (1986), “Analysis of Panel Data, Econometric Society monographs No. 11, New York: Cambridge University Press.

Chapter 9

AUTOREGRESSIVE AND DISTRIBUTED-LAG MODELS

Rajesh T., Harish Kumar H. V., Anuja A. R. and Shivaswamy G. P.

INTRODUCTION

The regression analysis involving time series data, which includes not only the current but also the lagged (past) values of the explanatory variables (the X's) is called a distributed-lag model. Similarly, regression analysis involving time series data, which includes not only the current but also the lagged (past) values of the dependent variables (the Y's) is called an autoregressive model.

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_t$$

represents a distributed-lag model, whereas

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t$$

represents an autoregressive model.

Autoregressive models are also known as dynamic models since they portray the time path of the dependent variable in relation to its past value(s). Autoregressive and distributed-lag models are used extensively in econometric analysis, and in the present chapter we will look at such models with a view to study about the role of lags in economics, reasons for the lags and theoretical justification for the commonly used lagged models in empirical econometrics.

Role of “TIME,” or “LAG,” in economics

The dependence of a variable Y (the dependent variable) on another variable(s) X (the explanatory variable) is rarely instantaneous in economics. Very often, Y responds to X with a lapse of time and such a lapse of time is called a lag. We consider an example to illustrate the nature of the lag.

Example

The consumption function

Assume that a person receives a salary increase of Rs. 1,000 in annual pay, and suppose that this is a permanent increase. Then what will be the effect of this increase in income

on the annual consumption expenditure of that person? In general, people do not rush to spend all the increase immediately after such a gain in income. Thus, that person may decide to increase consumption expenditure by Rs. 400 in the first year following the income increase, by another Rs. 300 in the next year, and by another Rs. 200 in the following year, saving the remainder. By the end of the third year, the person's annual consumption expenditure will be increased by Rs. 900. We can thus write the consumption function as

$$Y_t = \text{constant} + 0.4X_t + 0.3X_{t-1} + 0.2X_{t-2} + u_t \dots \quad (1)$$

where, Y is consumption expenditure and X is income.

The above equation shows that the effect of an increase in income of Rs. 1000 is distributed over a period of 3 years. Such models are therefore called as distributed-lag models because the effect of a given cause (income) is spread over a number of time periods. In general we may write

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + \beta_k X_{t-k} + u_t \dots \quad (2)$$

which is a distributed-lag model with a finite lag of k time periods. The coefficient β_0 is known as the short-run, or impact, multiplier because it gives the change in the mean value of Y following a unit change in X in the same time period. If the change in X is maintained at the same level thereafter, then, $(\beta_0 + \beta_1)$ gives the change in (the mean value of) Y in the next period, $(\beta_0 + \beta_1 + \beta_2)$ in the following period, and so on. These partial sums are called interim, or intermediate, multipliers. Finally, after k periods we obtain

$$\sum_{i=0}^k \beta_i = \beta_0 + \beta_1 + \beta_2 + \dots + \beta_k = \beta \dots \quad (3)$$

which is known as the long-run, or total, distributed-lag multiplier, provided the sum β exists.

If we define

$$\beta_i^* = \frac{\beta_i}{\sum \beta_i} = \frac{\beta_i}{\beta} \dots \quad (4)$$

we obtain “standardized” β_i . Partial sums of the standardized β_i then give the proportion of the long-run, or total, impact felt by a certain time period.

Going back to the consumption regression (1), we can see that the short-run multiplier, which is nothing but the short-run marginal propensity to consume (MPC), is 0.4, whereas the long-run multiplier, which is the long-run MPC, is $0.4 + 0.3 + 0.2 = 0.9$. That is, following a Rs. 1 increase in income, the consumer will increase his or her consumption level by about 40 paise in the year of increase, by another 30 paise in the next year, and by yet another 20 paise in the following year. The long-run impact of an increase of Rs. 1 in income is thus 90 paise. If we divide each β_i by 0.9, we obtain, respectively, 0.44, 0.33, and 0.23, which indicate that 44 percent of the total impact of

a unit change in X on Y is felt immediately, 77 percent after one year, and 100 percent by the end of the second year.

Estimation of distributed-lag models

Consider the following distributed-lag model in one explanatory variable:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + u_t \dots \quad (5)$$

where we have not defined the lag length, that is, how far back into the past we want to go. These type of model is called as infinite (lag) model, whereas a model of the type as shown in equation 2 is called as finite (lag) distributed-lag model, where the lag length k is specified.

We can adopt two approaches to estimate the α and β 's of equation 5.

- (1) ad hoc estimation and
- (2) a priori restrictions on the β 's (Assumption: β 's follows some systematic pattern).

We will discuss ad hoc estimation in this section

Ad hoc estimation of Distributed-Lag models

As the explanatory variable X_t is assumed to be non-stochastic, X_{t-1} , X_{t-2} , and so on, are non-stochastic, too. Therefore, the ordinary least squares (OLS) can be applied to (5). This is the approach taken by Alt and Tinbergen. They suggest that one may proceed sequentially to estimate (5); that is, first regress Y_t on X_t , then regress Y_t on X_t and X_{t-1} , then regress Y_t on X_t , X_{t-1} , and X_{t-2} , and so on. We need to stop this sequential procedure when the regression coefficients of the lagged variables start becoming statistically insignificant and/or the coefficient of at least one of the variables changes signs from positive to negative or vice versa. Based on this principle, Alt (1942) regressed fuel oil consumption Y on new orders X on the quarterly data for the period 1930–1939, and the results were as follows:

$$\hat{Y}_t = 8.37 + 0.171X_t$$

$$\hat{Y}_t = 8.27 + 0.111X_t + 0.064X_{t-1}$$

$$\hat{Y}_t = 8.27 + 0.109X_t + 0.071X_{t-1} - 0.055X_{t-2}$$

$$\hat{Y}_t = 8.32 + 0.108X_t + 0.063X_{t-1} + 0.022X_{t-2} - 0.020X_{t-3}$$

Alt chose the second regression as the “best” one because in the last two equations the sign of X_{t-2} was not stable and in the last equation the sign of X_{t-3} was negative, which may be difficult to interpret economically.

The koyck approach to distributed-lag models

Koyck has proposed a new method of estimating distributed-lag models. If we assume that the β 's is all of the same sign in the infinite lag distributed-lag model (5), Koyck assumes that they decline geometrically as follows.

$$\beta_k = \beta_0 \lambda^k \quad k = 0, 1, \dots \quad (6)$$

where λ , such that $0 < \lambda < 1$, is known as the rate of decline, or decay, of the distributed lag and $1 - \lambda$ is known as the speed of adjustment.

What (6) postulates is that each successive β coefficient is numerically less than each preceding β (since $\lambda < 1$), implying that as one goes back into the distant past, the effect of that lag on Y_t becomes progressively smaller. After all, current and recent past incomes are expected to affect current consumption expenditure more heavily than income in the distant past. Geometrically, the Koyck scheme is depicted in Fig 1.

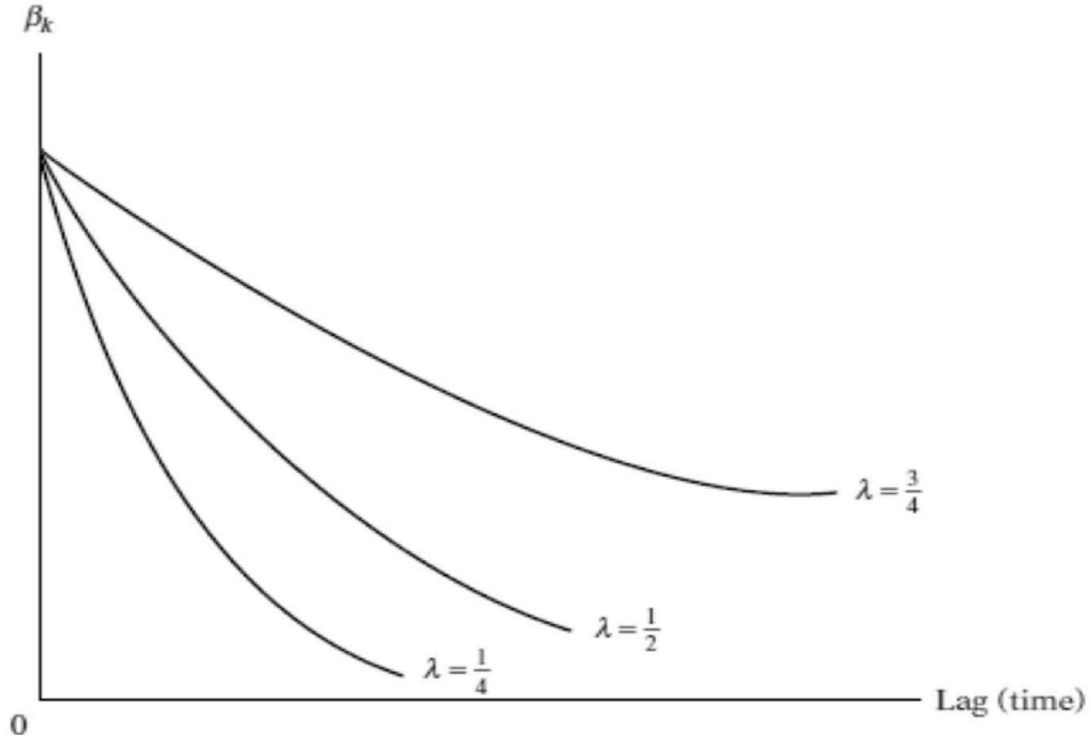


Fig 1: Koyck scheme (declining geometric distribution).

As we can see from the figure that value of the lag coefficient β_k depends on both the common β_0 and the value of λ . The closer the value of λ to 1, the slower the rate of decline in β_k , whereas the closer it is to zero, the more rapid the decline in β_k . Distant past values of X in the former case will exert sizable impact on Y_t , whereas their influence on Y_t in the latter case will diminish quickly. This pattern can be seen from the following illustration:

λ	β_0	β_1	β_2	β_3	β_4	β_5	...	β_{10}
0.75	β_0	$0.75\beta_0$	$0.56\beta_0$	$0.42\beta_0$	$0.32\beta_0$	$0.24\beta_0$...	$0.06\beta_0$
0.25	β_0	$0.25\beta_0$	$0.06\beta_0$	$0.02\beta_0$	$0.004\beta_0$	$0.001\beta_0$...	0.0

Note these features of the Koyck scheme: (1) By assuming non-negative values for λ , Koyck rules out the β 's from changing sign; (2) by assuming $\lambda < 1$, he gives lesser weight to the distant β 's than the current ones; and (3) he ensures that the sum of the β 's, which gives the long-run multiplier, is finite, namely,

$$\sum_{k=0}^{\infty} \beta_k = \beta_0 + \left(\frac{1}{1-\lambda}\right)$$

As a result of (6), the infinite lag model (5) may be written as

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \cdots + u_t \dots \quad (7)$$

The model is still not amenable to easy estimation since a large (literally infinite) number of parameters remain to be estimated and the parameter λ enters in a highly non-linear form: Strictly speaking, the method of linear regression analysis cannot be applied to such a model. But now Koyck suggests an ingenious way out. According to his model, we need to lag (7) by one period to obtain

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \beta_0 \lambda X_{t-2} + \beta_0 \lambda^2 X_{t-3} + \cdots + u_{t-1} \dots \quad (8)$$

Then multiply (8) by λ to obtain

$$\lambda Y_{t-1} = \lambda \alpha + \lambda \beta_0 X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \beta_0 \lambda^3 X_{t-3} + \cdots + \lambda u_{t-1} \dots \quad (9)$$

By subtracting (9) from (7), we will get

$$Y_t - \lambda Y_{t-1} = \alpha(1 - \lambda) + \beta_0 X_t + (u_t - \lambda u_{t-1}) \dots \quad (10)$$

or, rearranging,

$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + v_t \dots \quad (11)$$

Where $v_t = (u_t - \lambda u_{t-1})$, a moving average of u_t and u_{t-1} .

The procedure described above is known as the Koyck transformation. By comparing (11) with (5), we can see the tremendous simplification accomplished by Koyck. Whereas before we had to estimate α and an infinite number of β 's, but now we have to estimate only three unknowns: α , β_0 , and λ . Now there is no reason to expect multicollinearity. In a sense multicollinearity is resolved by replacing X_{t-1} , X_{t-2} , \dots , by a single variable, namely, Y_{t-1} .

The partial sums of the standardized β_i tell us the proportion of the long-run, or total, impact felt by a certain time period. In general, the mean or median lag is often used to characterize the nature of the lag structure of a distributed lag model.

The Median Lag

The median lag is the time required for the first half, or 50 percent, of the total change in Y following a unit sustained change in X. For the Koyck model, the median lag is as follows

$$\text{Koyck model: Median lag} = -\frac{\log 2}{\log \lambda}$$

Thus, the median lag is 0.4306 if $\lambda = 0.2$, but the median lag is 3.1067 if $\lambda = 0.8$. In the former case 50 percent of the total change in Y is accomplished in less than half a period, whereas in the latter case it takes more than 3 periods to accomplish the 50 percent change.

The Mean Lag

Provided all β_k are positive, the mean, or average, lag is defined as

$$\text{Koyck model: Mean lag} = -\frac{\lambda}{1 - \lambda}$$

Thus, if $\lambda = 0.5$, the mean lag is 1. The median and mean lags serve as a summary measure of the speed with which Y responds to X.

ILLUSTRATION**Per Capita Personal Consumption**

This example studies per capita personal consumption expenditure (PPCE) in relation to per capita disposable income (PPDI) in India for the period 1972–2018. As an illustration of the Koyck model, consider the data given in Table 1 (Gujarati *et al.*, 2012).

Table 1 : Per capita personal consumption

Year	PPCE	PPDI	PPCE(-1)	Year	PPCE	PPDI	PPCE(-1)
1972	7639	8501	7542	1996	15109	17777	14422
1973	7639	8667	7639	1997	15806	18242	15109
1974	7936	8993	7639	1998	16336	18672	15806
1975	8178	9221	7936	1999	16760	18838	16336
1976	8605	9827	8178	2000	17320	19506	16760
1977	9097	10360	8605	2001	17664	19886	17320
1978	9559	10831	9097	2002	17833	20047	17664
1979	9760	11223	9559	2003	17614	19875	17833
1980	10276	11658	9760	2004	17974	20314	17614
1981	10586	11929	10276	2005	18359	20259	17974
1982	10721	12329	10586	2006	18848	20578	18359
1983	11022	12767	10721	2007	19148	20919	18848

Year	PPCE	PPDI	PPCE(-1)	Year	PPCE	PPDI	PPCE(-1)
1984	11634	13278	11022	2008	19601	21312	19148
1985	12137	14111	11634	2009	20131	21831	19601
1986	11914	13860	12137	2010	20949	22897	20131
1987	12086	14057	11914	2011	21816	23330	20949
1988	12685	14504	12086	2012	22626	24235	21816
1989	13130	14894	12685	2013	22971	24453	22626
1990	13603	15470	13130	2014	23378	24983	22971
1991	13796	15697	13603	2015	23809	25301	23378
1992	13582	15706	13796	2016	24477	25998	23809
1993	13645	15983	13582	2017	25043	26202	24477
1994	13710	16184	13645	2018	25594	26771	25043
1995	14422	16594	13710				

The result of regression of Per Capita Consumption Expenditure (PPCE) on Per Capita Personal Disposal Income (PPDI) and lagged PPCE is as follows:

Dependent Variable: PPCE

Method: Least Squares

Sample (adjusted): 1972-2018

Regression Statistics	
Multiple R	0.999105
R Square	0.998211
Adjusted R Square	0.998129
Standard Error	225.1053
Observations	47

	Coefficients	Standard Error	t Stat	P-value
Intercept	-237.251	153.9967	-1.54063	0.130569
PPDI	0.2132	0.070481	3.024651	0.004145
PPCE(-1)	0.7978	0.073183	10.9018	4.36E-14

Adjusted R Square is 99.82 percent, which indicates that 99.82 percent of the variation in PPCE is explained by PPDI and PPCE Lag. From the regression analysis, we found that $\beta_0 = 0.2132$ and $\lambda = 0.7978$. β_0 gives the short run (immediate) effect i.e. if PPDI increases by 1 percent then PPCE will increase by 0.2132 percent in the same year. Long run multiplier is given by the following equation;

$$\text{Long run multiplier} = \beta_0 \left(\frac{1}{1-\lambda} \right) \approx 1.0537$$

In words, a sustained increase of 1 rupee in PPDI will eventually lead to about 1.05 rupees increase in PPCE. The long-run consumption function can be written as:

$$PPCE_t = -1247.1351 + 1.0537(PPDI_t)$$

It can be obtained by dividing the short-run consumption function by 0.2029 and dropping the lagged PPDI term. The median lag is given by;

$$\text{Median lag} = -\frac{\log 2}{\log \lambda} = -\frac{\log(2)}{\log(0.7971)} = 3.0589$$

i.e. 50 percent of this total effect of increase in PPDI on PPCE is felt after 3 years.

REFERENCES

- Alt, F. F. (1942), Distributed Lags, *Econometrica*, 10: 113-128
- Gujarati, D. N., Porter, D. C. and S. Gunashekar (2012), Basic Econometrics (Fifth edition). Mc Graw Hill Education (India) Private Limited, New Delhi, 656-713.
- Tinbergen, J. (1949), Long-Term Foreign Trade Elasticities, *Metroeconomica*, 1: 174-185

Chapter 10

CONJOINT ANALYSIS

Sukanta Dash, Krishan Lal and Rajender Parsad

INTRODUCTION

Conjoint Analysis is about the people make choices between products, services or both, so that businesses can design new products or services that better meet customers' underlying needs. Conjoint analysis has been found to be an extremely powerful of way of capturing what really drives customers to buy one product over another and what customers really value. A key benefit of conjoint analysis is to produce dynamic market models that enable companies to test out what steps they would need to take for improvement in market share, or how competitors' behavior will affect their customers.

Conjoint analysis is a statistical technique used in market research to determine how people value different features that make up an individual product or service. The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making. A controlled set of potential products or services is shown to respondents and by analyzing how they make preferences between these products, the implicit valuation of the individual elements making up the product or service can be determined. These implicit valuations (utilities or part-worths) can be used to create market models that estimate market share, revenue and even profitability of new designs.

Conjoint originated in mathematical psychology and was developed by marketing professor, Paul Green at the Wharton School of the University of Pennsylvania and Data Chan. Other prominent conjoint analysis pioneers include Richard Johnson (founder of Sawtooth Software) who developed the Adaptive Conjoint Analysis technique in the 1980s and Jordan Louviere who invented and developed Choice-based approaches to conjoint analysis and related techniques such as MaxDiff.

History of Conjoint Analysis

The earliest forms of conjoint analysis can be traced back to the 1970s having developed from the psychology of decision making and econometric choice theory. Key developers have been Paul Green (Marketing use of decompositional models), Jordan Louviere (Choice-based conjoint) and Rich Johnson (Sawtooth Software and Adaptive conjoint).

The first academic papers describing conjoint came in 1971 (Paul Green) and Harvard Business Review was describing the technique to the wider business audience in 1975.

The advent of computer-based personal interviewing in the 1980s saw major strategy consultants start to use conjoint, while from the economics field choice theory and the ability to estimate parameters using Logit-based MLE lead to the development of choice-based conjoint. For further details about conjoint analysis one can refer to Green and Srinivasan (1978), Green *et al.* (1981), Marder (1999) and Orme (2005).

Early years - full profile and part-worths

Conjoint analysis has its earliest roots in psychology and the testing of the theory that when people take decisions the result is the ‘sum’ of all the bits of value for each part in that decision. So when we buy a computer, there is a difference in value between computers with a small screen compared to a large screen. Hence, the difference in value to the customer can be linked just to this one feature and by adding up different values in a kind of ‘configurator’ style you can come to a conclusion about the overall value of the product or service. This seemed sensible in theory, but to test it required a method for breaking a product down into constituent parts, building profiles from these parts, then gathering preference data, then finally testing untried combinations to see if the customer preference was as expected. This is then the root of conjoint design. A demonstration can be seen here.

The first such tests were called full-profile designs (since the respondent saw profiles with one of each type of part) and were designed on cards using some form of ranking exercise typically. A major hurdle was in reducing the number of profiles down to a manageable number. This drew on work from statistics looking at experimental design - how do you minimize the number of experiments required to test a number of combinations of properties, for instance in drug development? The result was the use of ‘fractional factorial orthogonal designs’ to build the profiles to test. Since the objective was to validate the decision making process, tests were carried out and analyzed and then compared to ‘hold-out’ profiles. These were profiles that were not counted for statistical accuracy but were required to validate the model.

The method of analysis initially was analysis of variance (ANOVA) to produce a statistical model to predict the preference drivers. Early studies evaluated attributes as continuous (*eg.* price) or discrete (*eg.* colour), but it soon became apparent that it was more effective to treat all variables as if they were discrete variables as even ‘linear’ attributes such as price were often non-linear in nature. The result of the analysis was the calculation of ‘part-worths’. That is the model betas which describe how much each variable contributes to the final model. A higher beta is being more important. Since these part-worths have no units and because they are predicting an abstract entity such as preference, it is possible to scale and adjust the parameters without affecting the underlying model outcomes. This meant many of the consultancies that started to use conjoint turned these part-worths into the more user-friendly ‘utilities’.

Developments

Research into conjoint type techniques continues to develop. The advent of online research means that computer-based advanced techniques such as adaptive are much lower cost than when computer interviewing was carried out. There is an ongoing development of computer learning techniques using elements such as genetic algorithms and evolutionary learning to 'search' the decision making space with the consumer and so make judgments about people would make decisions. In addition, the development of product configurations, such as those used by Dell selling computers, means that there are other ways of getting at choice data. There is also controversy at the academic level since there are still some very fundamental assumptions being made about the types of models that people use - we blithely treat everything as linear functions and often disregard non-linear factors such as diminishing returns.

Conjoint Analysis Design

Conjoint analysis and trade-off studies are amongst the most sophisticated forms of market research because they provide estimates of demand based on product or service design. But like any form of research, the quality of the output depends heavily on the quality of the design. For conjoint analysis, in particular this means choosing the right flavour or type of conjoint is to use and ensuring that the design of the attributes and levels meets the research, analysis and business requirements.

The basis of conjoint design splits into three interrelated parts. Firstly, there is the choice of the flavour or type of conjoint analysis that is to be used. Different types of conjoint analysis have different strengths and weaknesses. Some forms can be carried out on paper, some require the use of computer or internet-based interviews, some allow large complex designs, others are more robust in the face of issues like pricing. The second element of design is the attributes and levels that make up the product. The final inter-related element is the sample itself and the contact methods available. In the conjoint interview can be carried out face-to-face, or over the internet, or by post and phone, and how long will the questionnaire be.

Attributes and Levels in Conjoint Analysis

Attributes and levels form the fundamental basis of conjoint analysis. The idea is that a product or service can be broken down into its constituent parts - so for instance a mobile phone has a size, weight, battery life, size of address book, type of ring. Each of these elements making up a generic mobile phone is known as an attribute. When we compare between mobile phones each will have a different specification on each of these attributes. We might have choices in terms of battery life between 12, 24, 36, 48 hours of battery life. Each of these options is known as a level of the battery life attribute. In another example, a car might have an attribute colour which is made up of the levels red, blue, green, white. This breaking down of products and services into attributes and levels is an extremely powerful tool for examining what a business

offers and what it should be offering. For new product development, combining this product breakdown with an understanding of what the customer values most means that the business can focus its efforts on those areas of most importance to customers.

In addition, each level can also be thought of as a performance target. If the customer, on the attribute 'delivery', wants the level "next day delivery" and we are perceived to be offering "48hr delivery" then there is a gap to make up. What is more with conjoint analysis, it is possible to calculate what the value of making up that gap is to the business. Compare this to the cost of doing it and the business can decide whether it is worth it. Note that this is far more useful than simply knowing that we score 6.9 out of 10 on deliveries. This has no meaning. It doesn't say where the customer thinks we are and it doesn't tell us what we need to do to get higher, or whether it is worth trying to improve this score, or what it will cost. For a lot of strategic work attributes and levels are far more powerful tools than scales and scores.

In its attributes and levels have to behave in certain ways so that the conjoint analysis is valid, and in certain other ways to make the conjoint useful. Firstly, each attribute has to be independent, that is it should not overlap with other attributes. So colour and fuel economy are clearly not related, so they can appear together. However, some things like "car shape" and "number of passengers" aren't independent - a 7 seated coupe is not realistic. There are also subtler effects - certain attributes have *halo-effect* on others around them. For instance, if one level were "gold-plated handle", many people would infer that the rest of the product was also of better quality when there is no other information to support this. The main difficulty this causes is that price and brand need to be treated extremely carefully in conjoint studies to produce valid results.

Each level also needs to be capable of being read and understood on its own. Although attributes are used to help break a product down and in analysis, when presented to respondents the entire respondent sees are the levels. Independent and readable levels are important from an analysis point of view, but for the conjoint to be useful it also needs to ensure that the range of attributes cover all the areas that are important to the customer, and that the range of levels cover all possibilities from worst-case to blue-sky. Having said that if we do want to focus on a particular element of choice, it is possible to use an "all other things being equal" scenario, but if we are looking at the type of car-radio this limitation would not tell us about the importance of the car-radio in the overall choice of car.

For many products, particularly in business markets, service can be more important than the actual product. By using both product and service attributes in the same conjoint it is possible to see how customers trade-off service against features. However, care has to be taken to balance the attributes to prevent biasing the outcome one way or another. It is also vital that the levels used describe real and where possible measurable and actionable performance. We could run a conjoint with levels "very good fuel economy" "good fuel economy" "not so good fuel economy". But, as with the scales above, these

do not mean anything. If we come out with “good fuel economy” and the customer wants “very good fuel economy” what do you do? One good test is if we showed it to the shop-floor (or the call centre) would they know how to deliver it. A second test is whether it is a statement you could use on a packet or in a brochure

PROCEDURES OF CONJOINT ANALYSIS

1) Hybrid designs and Adaptive Conjoint Analysis (ACA)

The main problems in conjoint is reducing the number of profiles that need to be evaluated by respondents. Richard Johnson was developing a range of techniques including a ‘pairwise grid’ approach on paper, but then went on to develop Sawtooth Software’s Adaptive Conjoint Analysis (ACA) - a computer based approach to conduct conjoint analysis which relied on initial self-explicated exercises where respondents pre-rank and pre-value items before undertaking the preference task. These techniques are described as ‘hybrid’ containing as they do a combination of trade-off and non-trade-off elements. ACA was taken up as a very practical method for estimating and valuing customer demands by management consultancy firms, although it tended to be disfavoured by academics because of the lack of a firm underlying theoretical model and the somewhat arbitrariness of the design. That’s not to say it didn’t work, just that it was more a practical tool than a theoretical one. Certainly at the end of 1990s ACA was the most common form of conjoint analysis in use.

Adaptive Conjoint Analysis (ACA) is one of two most common methods for carrying out conjoint analysis. The benefits of ACA are that it allows for a large number of attributes (up to 30) and levels (up to 7 per attribute) to be used. However, ACA does require a computer-based interview and the large number of attributes means that it is common for an ACA interview to last 45 minutes or more. In addition, some of the methods it uses to simplify the task of working out utilities mean that some care is needed in choosing and designing the attributes in order to get reliable results.

Technically ACA is known as a hybrid technique as it contains elements of ‘self-explication’ followed by the trade-off tasks themselves. ACA itself is produced by Sawtooth Software and can be conducted face-to-face or on-line. Telephone use of ACA is difficult and paper-base questionnaires are not possible.

2) Choice Based Conjoint Analysis (CBC)

The most common alternative to ACA is Choice-based conjoint (CBC). Although this uses the same over-arching principles as ACA, in design, implementation and calculation it is completely different. Whereas ACA has respondents selecting from products described with two or three attributes, CBC shows full descriptions using all the attributes available. In addition, CBC can show more than just two “products” at the same time, together with a none-of-these option enabling more realistic choice decisions to be evaluated.

The limitation on the amount a respondent can absorb at a time, combined with the rapidly increasing number of “full-profile” combinations that are possible means that choice-based conjoint is typically limited to 5-7 attributes (in contrast to 25-30 for ACA). An additional twist is that utilities and importance in CBC are calculated across a sample as a whole, whereas for ACA we get utilities and importance for each individual in the sample. Combined with the lower number of attributes, this means that choice-based studies require far shorter questionnaires (15-20 minutes) and can be designed to be purely paper-based.

The advantages choice-based conjoint gives us are greater robustness of results - particularly for pricing work (although there are ways of getting around ACA's pricing limitations), combined with shorter and therefore less costly fieldwork. It is also favoured for its rigour academically. The disadvantages are the lower number of attributes that are possible, and the lack of individual level utilities - which means that post-hoc cluster analysis for things like needs-based segmentation are not possible using choice based conjoint although techniques such as hierarchical bayesian analysis seeks to remedy this by post-hoc simulation of individual level values.

3) Discrete Choice Analysis

A more advanced form of choice-based conjoint is Discrete Choice Analysis (also known as “stated preference research”). DCA studies are particularly popular for transportation studies looking at modal choice - the preference between a train, car and airline for instance. The main difference from CBC is the inclusion of continuous variables such as price and time. This allows to examine the varying costs of the ticket with varying times taken to travel and so to establish the value of time for the journey. This enables transport economists to make statements like “2cm extra leg room is worth 10 minutes longer journey time or £40 extra fare” or “an extra train every 15 minutes would encourage x per cent of car drivers to switch to the train”.

4) Full Profile Conjoint Analysis

An additional option that dates back a long time but that is still used is full profile conjoint analysis. Full-profile is the original form of conjoint and is still in use, though predominantly in the US it would appear. Like choice-based conjoint this uses a more limited number of attributes to describe the product or service, but sufficient cards or treatments are shown to one respondent to enable individual level utilities to be calculated. A fractional factorial design is used to specify a fixed set of profiles that need to be shown for analysis. The difficulty is that this does limit the number of attributes quite severely. However, these old school studies are still popular for simple, non-computer-based conjoint projects and are most common for students learning about conjoint for the first time.

5) Choice of Conjoint Analysis

The earliest forms of conjoint analysis were what are known as Full Profile studies, in which a small set of attributes (typically 4 to 5) are used to create profiles that are shown

to respondents, often on individual cards. Respondents then rank or rate these profiles. Using relatively simple dummy variable regression analysis the implicit utilities for the levels can be calculated. Two drawbacks were seen in these early designs. Firstly, the number of attributes in use was heavily restricted. With large numbers of attributes, the consideration task for respondents becomes too large and even with fractional factorial designs the number of profiles for evaluation can increase rapidly. In order to use more attributes (up to 30), hybrid conjoint techniques were developed. The main alternative was to do some form of self-explication before the conjoint tasks and some form of adaptive computer-aided choice over the profiles to be shown.

The second drawback was that the task itself was unrealistic and did not link directly to behavioural theory. In real-life situations, the task would be some form of actual choice between alternatives rather than the more artificial ranking and rating originally used. Jordan Louviere pioneered an approach that used only a choice task which became the basic of choice-based conjoint and discrete choice analysis. This stated preference research is linked to econometric modeling and can be linked revealed preference where choice models are calibrated on the basis of real rather than survey data. Originally choice-based conjoint analysis was unable to provide individual level utilities as it aggregated choices across a market. This made it unsuitable for market segmentation studies. With newer hierarchical Bayes analysis techniques, individual level utilities can be imputed back to provide individual level data.

Information Collection

Data for conjoint analysis is most commonly gathered through a market research survey, although conjoint analysis can also be applied to a carefully designed configurator or data from an appropriately design test market experiment. Market research rules of thumb apply with regard to statistical sample size and accuracy when designing conjoint analysis interviews.

A typical Adaptive Conjoint questionnaire with 20-25 attributes may take more than 30 minutes to complete. Choice based conjoint, by using a smaller profile set distributed across the sample as a whole may be completed in less than 15 minutes. Choice exercises may be displayed as a store front type layout or in some other simulated shopping environment.

Conjoint Analysis Models

Most conjoint analysts fit what is known as part-worth model to respondents' evaluating judgments, whether obtained by trade-off tables or full profile, self explicated, or hybrid approaches. Let $p = 1, 2, \dots$ and t denote the set of t attributes that are used in the study design. Let y_{ip} denote the level of p^{th} attribute for the j^{th} stimulus. First we assume that y_{ip} is inherently continuous. The vector model assumes that the preferences s_j for the j^{th} stimulus is given by

$$s_j = \sum_{p=1}^r w_p y_{jp}$$

where w_p denotes a respondent's weight for each of the t attributes.

The ideal point model shows that preferences s_j is negatively related to the weighted squared distance d_j^2 of the location y_{ip} of the j^{th} stimulus from the individual's ideal point x_p of the location y_{ip} of the j^{th} stimulus from the individual's ideal point x_p where d_j^2 is defined as

$$d_j^2 = \sum_{p=1}^r w_p (y_{jp} - x_p)^2$$

The part-worth model assumes that

$$s_j = \sum_{p=1}^t f_p(y_{jp})$$

where f_p is a function denoting the part-worth of difference levels of y_{ip} for the p^{th} attribute. In practice $f_p(y_{jp})$ is estimated for a selected set of discrete levels of y_{ip} .

Any number of algorithms may be used to estimate utility functions. These utility functions indicate the perceived value of the feature and how sensitive consumer perceptions and preferences are to changes in product features. The actual mode of analysis will depend on the design of the task and profiles for respondents. For full profile tasks, linear regression may be appropriate, for choice based tasks, maximum likelihood estimation, usually with logistic regression are typically used. The original methods were monotonic analysis of variance or linear programming techniques, but these are largely obsolete in contemporary marketing research practice. Conjoint analysis makes extensive use of *Orthogonal Arrays* (Addelman, 1962) to reduce the number of stimulus descriptions to a small fraction of the total number of descriptions.

Conjoint Analysis Applications

Calculation of the part worth utilities becomes just the starting point for many interesting applications of conjoint analysis. The important ones among them are described next:

Optimum product design: Since all possible product concepts can be compared after adding their respective attribute levels part worth utilities, it is possible to determine the demand for different products out of any given set of available products in the market place. The demand levels can be converted into profit figures as cost of producing and marketing can also be calculated. These cost calculations are possible as the volume of operations and the features of the products are now known. Thus, the optimum product can be chosen from the profits point of view (or any of the other given management's objective). Customers differential rates of purchase of products are also duly considered at this stage.

Quite often, a manager may like to know the effects of slight change in any of the attribute by his own company or the competitor's. Conjoint Analysis allows this kind of "What if" analysis very easily with the data base of part worth utilities. In fact, different kinds of scenarios can be simulated and the manager can optimize not only the product but other aspects of his marketing strategy. Similarly, whenever there is any change in competitor's actions or in the environment a fresh scenario can be drawn for simulation. Of course, the simulation shall be limited to the attributes considered in the analysis. This feature does also help in increasing the shelf life of the conjoint analysis output.

Market segmentation: Since the Conjoint Analysis is done at the individual customer level, the individual customer's identity can be retained throughout the analysis. Thus, customers can be segmented according to their sensitivities to different product attributes.

It is also possible to identify the customers segments, which would be attracted most for the proposed product position. This helps in having a focused matching between the chosen product position and the target customer segment. It can also help in identifying that part of competitor's market which needs to be poached for snatching market share from them. Similarly, the same type of analysis can be done to identify the most vulnerable section of one's own market segment.

Sometimes, an additional product offer appears to be quite attractive. But, this may be at the cost of cannibalisation. Conjoint Analysis can help in estimating the effects of cannibalisation as well. Thus, it helps in maximizing net profits of the organization.

SWOT analysis: First of all, the part worth utility of the brand itself can tell about the relative brand strength. Similarly, by looking at the other features of one's own and competitor's offers Conjoint Analysis enables the marketers to conduct his detailed SWOT Analysis.

Estimating customer level brand equity: Conjoint Analysis is a good bridge between the consumer level perceptions and the financial worth of the offers. This can be used for estimating the important parameter of brand equity at the consumers level. There is scope of differentiating the "Loyal", "Acceptors" and "Switchers" for more accurate calculations of brand equity.

Advantages and Disadvantages

Advantages

- It estimates psychological tradeoffs that consumers make when evaluating several attributes together.
- It measures preferences at the individual level.
- It uncovers real or hidden drivers which may not be apparent to the respondent themselves.

- It is realistic choice or shopping task.
- It is able to use physical objects.
- If it is appropriately designed, the ability to model interactions between attributes can be used to develop needs based segmentation.

Disadvantages

Designing the conjoint studies can be complex with too many options, respondents resort to simplification strategies. It is difficult to use for product positioning research because there is no procedure for converting perceptions about actual features to perceptions about a reduced set of underlying features. The respondents are unable to articulate attitudes toward new categories, or may feel forced to think about issues they would otherwise not give much thought to poorly designed studies may over-value emotional/preference variables and undervalue concrete variables. It does not take into account the number items per purchase so it can give a poor reading of market share.

ILLUSTRATION

Use of Fractional Factorial Plans in Conjoint Analysis: An Example

In Conjoint Analysis the profile of different products is presented to the consumers for their responses. These profiles are generated by varying the levels of its attribute. For example, suppose we are conducting a Conjoint Analysis based study of Laptop product. Let us assume that the most important attributes considered by its customers are OS, Price, Display size, Processor and RAM. Let us further assume that the following levels of attributes are considered relevant and interesting by the marketer for the study (Table 1).

Table 1: Levels of attributes

S. No.	Attribute	Levels
1.	OS	Windows Linux MAC
2.	Price	<Rs. 20,000 Rs. 20,000/- to Rs. 40,000/ Rs. 40,000/- to Rs. 60,000/ >Rs. 60,000/
3	Display Size	13” 14” 15.6”
4.	Processor	I3 I5 I7
5.	RAM	8 GB 16 GB

Since the five attributes can take 3, 4, 3, 3 and 2 levels, the total number of possible product concepts that can be generated by configuring these attributes is $3 \times 4 \times 3 \times 3 \times 2 = 216$. In

order to determine the part worth utilities of each of the levels of all these attributes, we shall have to take 216 different product concepts for getting his/her responses. This number is certainly too large for any consumer. Therefore, we resort to the method of Fractional Factorial Design of Experiment to make it manageable.

The statistical technique of Fractional Factorial Design of Experiment finds out the minimum number of product designs which are necessary to use in the study and yet provide us all the information that we originally sought. These designs are also mutually independent (orthogonal) to avoid any redundancy in the data and allow the representation of each of the attributes and their respective levels in an unbiased manner.

In the example of Laptop considered here, this technique has given us only 16 designs out of the 216 possible Laptops. However, it should be noted that such reduction in number of product designs is possible only after making certain assumptions. For example, we had assumed that none of the attributes interact among themselves. Or in other words, the attributes are considered to be independent of each other. Only under this assumption we got the number of product concepts as 16. At the other end, if we would have allowed all the attributes to interact with each other the required number of product concepts would have remained as 216. With different types of assumptions, the number of concepts required would be in between these extremes.

The 16 product concepts found through this method are not unique (Table 2). Many other sets of 16 cards would have also been equally good. However, all of these sets would have to be independent and represent all the attributes and their respective levels in an unbiased manner. We are illustrating below one such set of 16 cards representing the product concepts of Laptops using Fractional Factorial Design of Experiment. (In SPSS from the menus choose Data > Orthogonal Design > Generate...)

Table 2: Description of product concept cards

Card#	OS	Price	Display Size	Processor	RAM
Card 1	Windows	<Rs. 20,000	15.6"	I3	8 GB
Card 2	Linux	Rs. 40,000/- to Rs. 60,000/	14"	I3	16 GB
Card 3	MAC	Rs. 20,000/- to Rs. 40,000/	14"	I3	8 GB
Card 4	Windows	>Rs. 60,000/	13"	I7	16 GB
Card 5	Linux	<Rs. 20,000	14"	I5	8 GB
Card 6	MAC	Rs. 20,000/- to Rs. 40,000/	13"	I5	8 GB
Card 7	Windows	Rs. 40,000/- to Rs. 60,000/	15.6"	I5	16 GB
Card 8	Linux	>Rs. 60,000/	14"	I5	16 GB
Card 9	MAC	<Rs. 20,000	13"	I7	8 GB

Conjoint Analysis

Card#	OS	Price	Display Size	Processor	RAM
Card 10	Windows	Rs. 15,000	14"	I5	8 GB
Card 11	Linux	Rs. 20,000/- to Rs. 40,000/	14"	I7	8 GB
Card 12	MAC	Rs. 20,000	15.6"	I3	8 GB
Card 13	Windows	Rs. 40,000/- to Rs. 60,000/	14"	I7	16 GB
Card 14	Linux	Rs. 20,000/- to Rs. 40,000/	15.6"	I5	8 GB
Card 15	MAC	>Rs. 60,000/	13"	I7	16 GB
Card 16	Windows	Rs. 40,000/- to Rs. 60,000/	14"	I7	16 GB

a) Physical design of stimuli

After selecting the product concepts required for Conjoint Analysis study, they need to be exposed to the consumers as stimuli. This may be done in a variety of ways mainly depending on the demands of the situation and the convenience of the researcher. Of course, it would be most desirable to present real life prototypes of the products according to the product concepts specified. These, may be given to the consumers for their usage or trials. But, such extreme ways of presenting the products may not always be possible or even necessary. In such cases, product models, diagrams or even verbal descriptions may be adopted.

b) Data collection

Ease of data collection is a key feature of Conjoint Analysis. The consumers are asked only to assign rating scores to each of the product stimuli or even rank the different concepts presented to them. This is quite a realistic task and is close to the shopping experiences where the customer merely makes choices. He does not have to respond to each of the attributes separately.

This feature of conjoint analysis is possible due to the use of Fractional Factorial Design of Experiment before collection of data and the use of Conjoint Analysis after collecting the data. In other words, the use of the technique eases the burden of the respondents.

c) Determination of part worth utilities

The rating or ranking data obtained from the consumers are analyzed next. Two methods are more popular for this purpose. In one method, the part worth utilize for each of the levels of each attributes are arbitrarily assigned. Based on these assumed values, consumers overall rating or ranking (as the case may be) are estimated. These estimated responses may understandably, be quite different from the actual data. After a few iterations convergence is achieved so that the part worth utilities found approximate the estimate responses to the actual data best.

In the alternative method, the part worth utilities are derived in one step. Here, an error function describing the difference between the estimated and actual data is defined. This function is then minimized.

After using any of the available method, the output is obtained for each of the respondent separately. This is quite significant as the disaggregate data can be combined in any of the desired way. But, if the output was only at the aggregate level then disaggregation might not have been possible.

In our example of Laptop, the part worth utilities may be found for each of the attributes and their levels (Table 3). It is evident that price plays most important role (56.2%) in the minds of customers. This is followed by processor (32.6%), RAM (6.8%), OS (3.4%) and Display Size (1.0%).

Table 3: Utilities for each of attributes and their levels

S. No.	Attribute	Levels	Part Worth Utility
1.	OS (3.4%)	Windows Linux MAC	1.6 1.2 -0.3
2.	Price (56.2%)	<Rs. 20,000 Rs. 20,000/- to Rs. 40,000/ Rs. 40,000/- to Rs. 60,000/ >Rs. 60,000/	12.8 -3.6 11.6 8.5
3	Display Size (1%)	13” 14” 15.6”	0.9 2.8 0.3
4.	Processor (32.6%)	I3 I5 I7	-5.5 7.2 10.9
5.	RAM (6.8%)	8 GB 16 GB	1.3 1.4

Customers respond quite predictably towards different price levels. They prefer lesser price to the higher ones. Among Processor, we find that I7 is the most liked and I3 in the least liked one. However, they prefer the 16 GB RAM most. They prefer 14” Display size most. In terms of the OS, they prefer Windows operating system the most.

The part worth utilities can now be added to determine the total utility for each of the possible product concepts. This allows us to scan the consumer preference pattern for all of the 216 product concepts although he has been exposed to only 16 of them. We can now also rank all of these 216 product concepts

REFERENCES

- Addelman, S. (1962), Symmetrical and asymmetrical fractional factorial plans. *Technometrics*, 4: 47-58.
- Green, P. and V. Srinivasan (1978), Conjoint analysis in consumer research: Issues and outlook, *Journal of Consumer Research*, 5: 103-123.
- Green, P., J. Carroll and S. Goldberg (1981), A general approach to product design optimization via conjoint analysis, *Journal of Marketing*, 43: 17-35.
- Marder, E. (1999), The assumptions of choice modelling: Conjoint analysis and SUMM. (Single unit marketing model). *Canadian Journal of Market Research*, 18 (1): 1 - 12.
- Orme, B. (2005), *Getting Started with Conjoint Analysis* Madison, WI: Research Publishers LLC. ISBN 0-9727297-4-7.

ANNEXURE I

Syntax of conjoint analysis using SPSS:

```
CONJOINT PLAN='file specification'  
/DATA='file specification'  
/SEQUENCE=PREF1 TO PREF22  
/SUBJECT=ID  
/FACTORS=OS (DISCRETE) DISPLAY SIZE (DISCRETE)  
PRICE (LINEAR LESS)  
PROCESSOR (LINEAR MORE) RAM (LINEAR MORE)  
/PRINT=SUMMARYONLY.
```

Chapter 11

TWO STAGE LEAST SQUARE SIMULTANEOUS EQUATION MODEL

Shivendra Kumar Srivastava and Jaspal Singh

INTRODUCTION

Most of the empirical research works in economics are undertaken using single-equation type economic relationships. In such models, one variable (the dependent variable Y or explained) is expressed as a linear function of one or more other variables (the independent/explanatory/ repressor variables, the X's). Here, the implicit assumption is that the cause-and-effect relationship, if any, between Y and the X's is unidirectional. The explanatory variables are the cause and the dependent variable is the effect. But in many situations, such a one-way or unidirectional cause-and-effect relationship is not meaningful. This occurs if Y is determined by the X's, and some of the X's are, in turn, determined by Y. In short, there is a two way, or simultaneous, relationship between Y and (some of) the X's, which makes the distinction between dependent and explanatory variables of dubious value (Madhani, 2005 and Gujarati, 2005). Under such circumstances, more than one regression equations, one for each independent variable, can be taken to understand the multi-flow of influence among the variables.

Illustration: Demand for any particular commodity depends on its own price (P), on other prices (P_0), and on income (I), so that;

$$Q_d = \alpha_0 + \alpha_1 P + \alpha_2 P_0 + \alpha_3 I + U_1 \dots \quad (1)$$

If we apply Ordinary Least Square (OLS) to obtain estimates on $\alpha_0, \alpha_1, \alpha_2$ and α_3 , it will violate one of the crucial assumption of the OLS procedure i.e. the explanatory variables are either nonstochastic, or if stochastic, are distributed independently of the stochastic disturbances. Since the price of the commodity is also affected by the quantity demanded of that commodity, the above single equation model cannot be treated as a complete model. In other words, there is two-way causation as below;

$$Q_d = f(P), \text{ and also}$$

$$P = f(Q_d)$$

That is, there needs to be at least one more equation (to describe relation between P and Q_d) to estimate the given demand function.

Assume that the required relation is described as under:

$$P = f(Q_d) = \beta_0 + \beta_1 Q_d + \beta_2 W + U_2 \dots \quad (2)$$

(W depicts the weather conditions)

To obtain estimates of α_1 , α_2 and α_3 of the demand function through OLS procedure, it is to be shown that explanatory variables P_0 and I in equation (1) are distributed independently of U_1 and the explanatory variable Q_d and W in equation (2) are distributed independently of U_2 .

Combining equation (1) and equation (2):

$$P = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 P + \alpha_2 P_0 + \alpha_3 I + U_1) + \beta_2 W + U_2$$

It is clear that P is not independent of U_1 in the above relation; i.e., $E(PU_1) \neq 0$. The estimates of α_1 , α_2 and α_3 will, therefore, turn out to be biased if OLS is applied to equation (1).

In such situations, simultaneous equation models are applied to describe joint dependence of variables. In contrast to single-equation models, in simultaneous-equation models more than one dependent or endogenous, variable is involved necessitating as many equations as the number of endogenous variables. Thus, the number of equations in such models is equal to the number of jointly dependent or endogenous variables involved in the phenomenon under analysis. A unique feature of simultaneous-equation models is that the endogenous variable (i.e. regressand) in one equation may appear as an explanatory variable (i.e. regressor) in another equation of the system. Also unlike the single-equation models, in simultaneous-equation model it is not possible (possible only under specific assumptions) to estimate a single equation of the model without taking into account information provided by other equations of the system. If one applies OLS to estimate the parameters of each equation disregarding other equations of the model the estimates so obtained are not only biased but also inconsistent i.e. as the sample size increases indefinitely, the estimators do not converge to their true population values. The bias arising from application of such procedure of estimation which treats each equation of the simultaneous-equation model as though it were a single-equation model is known as Simultaneity Bias or Simultaneous-Equation Bias.

SIMULTANEOUS-EQUATION METHODS

There are several methods of estimating simultaneous equation models with varying statistical properties. These methods may be divided into two categories; namely single equation methods also known as limited information methods) and system methods (also known as full information methods).

1. Single-equation methods
 - a. The Indirect Least Squares (ILS)

- b. The Method of Instrumental Variables (IV)
- c. Two Stage Least Squares (2 SLS)
- 2. System methods
 - a. Limited Information Maximum Likelihood (LIML)
 - b. Three Stage Least Squares (3 SLS)
 - c. Full Information Maximum Likelihood (FIML)

In single-equation methods, each equation in the system (of simultaneous equations) is estimated individually, taking into account any restriction placed on that equation without worrying about the restrictions on the other equations of the system. In other words, each equation of simultaneous-equations model is estimated individually disregarding the restrictions on the other equations in the model. On the other hand, in the system methods, all equations of the model are considered together and estimated simultaneously taking into account all restrictions placed on the equations; hence the name full information methods. In practice, the system methods are not commonly used for their complex nature, enormous computational works and sensitiveness to specification errors.

Two Stage Least Square Method

The two-stage least squares (2SLS), developed by Henri Theil and Robert Basman, involves two successive applications of OLS and seeks to remove the defect of existence of the correlation between the disturbance terms and the independent variable(s) so that when we apply OLS technique to each structural equation separately, the simultaneity bias gets eliminated.

Let us consider following demand-supply model

$$D_t = \alpha_0 + \alpha_1 P_t + \alpha_2 Y_t + U_1$$

$$S_t = \beta_0 + \beta_1 P_t + \beta_2 W_t + U_2$$

$$D_t = S_t$$

Where $D_t = S_t$ = quantity demanded and supplied, P_t = price of the commodity in question, Y_t = income, and W_t = weather index. In this simultaneous model, the variables P and Q are endogenous variables because their values are determined within the system. The variables Y_t and W_t are exogenous variable because its value is determined outside the system. The application of OLS to solve these equations will produce bias and inconsistent estimates because P_t is correlated with the disturbance term. By applying 2SLS method, we remove the defect of existence of correlation between the disturbance term and the independent variable(s). In this method, we purge the explanatory variable (P_t) which is correlated with the disturbance term with its own estimated value. This is done in two stages;

First obtain reduced form equations which express endogenous variables (P_t and D_t/S_t) as a function of exogenous variables (Y_t and W_t). To solve for our endogenous variables P_t and D_t , we set the supply and demand function equal to each other to get;

$$D_t = \pi_{10} + \pi_{11}Y_t + \pi_{12}W_t + V_1$$

$$P_t = \pi_{20} + \pi_{21}Y_t + \pi_{22}W_t + V_2$$

where the parameters π 's and V 's are reduced-form parameters and the error reduced-form errors, respectively. π 's are estimated by applying OLS to these reduced-form equations. This is the first stage of estimation.

Having estimated the π 's, we obtain values of \hat{P}_t or different values of Y_t and W_t . We now replace P_t in the structural model by \hat{P}_t obtained in the first stage as follows;

$$D_t = \alpha_0 + \alpha_1 \hat{P}_t + \alpha_2 Y_t + U_1$$

$$S_t = \beta_0 + \beta_1 \hat{P}_t + \beta_2 W_t + U_2$$

This is now a transformed model. Since \hat{P}_t is based on the estimates from the reduced-form equations, it acts as an instrument variable for the original data on P_t .

All the structural parameters are estimated by applying OLS to these transformed equations. This is the second stage of estimation.

Since in 2SLS, the endogenous variable(s) is purged by its own estimated value by taking into consideration all the exogenous variables, this method assumes complete knowledge of exogenous variables in the model. If the specification of these variables is not correct, the estimates of the structural parameters will not possess the desired properties. 2SLS estimates are thus sensitive to specification errors. They are also asymptotically unbiased and consistent i.e. their distribution collapses on the true parameter (value) as sample size tends to infinity. As such, this method requires a rather large number of observations, specially if the model includes many exogenous variables, which will be used in the first stage to estimate \hat{P}_t (endogenous) variable.

To summarize the procedure:

1. Least squares estimation of the reduced-form equation for endogenous variable (P) and the calculation of its predicted value (\hat{P}_t).
2. Least squares estimation of the structural equation in which the right-hand side of the endogenous variable (P) is replaced by its estimator (\hat{P}_t).

Empirical 2SLS model in agriculture

The level of agriculture performance has direct implications for overall economic development, particularly in rural areas where agriculture is a predominant sector. Agricultural productivity is a significant factor in determining level of household

income and poverty in rural India. Besides, agriculture income of a household was hypothesized to be a function of both productivity as well as number of persons dependent on same size of the land or conversely land-man ratio. Accordingly, poverty was hypothesized to be affected by agricultural productivity and number of workers per hectare of net sown area (labour to land ratio). Further, agricultural productivity itself depends on various production inputs, infrastructure, topographic and climatic factors. These relationships can be expressed through following equations (Chand and Srivastava, 2016)

$$RURALPOOR = \alpha_0 + \beta_1.AGRILPRODTY + \beta_2.WORKERPERLAND + \varepsilon$$

$$AGRILPRODTY = \delta_0 + \gamma_1.CROPINTENSITY + \gamma_2.IRRICOV + \gamma_3.FERTUSE + \gamma_4.RAINFALL + \gamma_5.PROBLEMISOIL + \gamma_5.GWDEV + \theta$$

where,

RURALPOOR = rural poverty (%)

AGRILPRODTY = agricultural productivity (Rs/ ha of net sown area)

WORKERPERLAND = agricultural workers per ha of net sown area

CROPINTENSITY = cropping intensity (%)

IRRICOV = irrigation coverage (share of gross irrigated area in gross sown area)

FERTUSE = fertilizer use (kg/ha)

RAINFALL = annual rainfall (mm)

PROBLEMISOIL = share of problem soil in total area (%)

GWDEV = groundwater development (%)

Given the endogeneity of the independent variable ‘agricultural productivity’, 2SLS was used to solve above equations. The study used district level data on above variables obtained from various sources. The district level data set used in the analysis includes 487 districts of the country which covers about 94 per cent of the net sown area of the country. District-wise rural poverty rate was estimated using unit-level consumption expenditure survey data of National Sample Survey Office for the year 2011-12. The average monthly per capita consumption expenditure (MPCE) was compared with state specific official poverty line to estimate district-level poverty estimates. Agricultural productivity was computed by taking sum of output of selected agricultural commodities multiplied by state level implicit prices of respective agricultural commodities, divided by net sown area. The output prices data was generated by dividing the state level value of output of each crop estimated by Central Statistical Organization (CSO), by output of the crop for the year 2010-11.

The value of output for the crops considered in the study was multiplied by ratio of GCAt/GCAc, where GCAt is the reported gross cropped area and GCAc is the

sum of area under crops considered in the study to arrive at estimate of value of crop output for GCAt. This figure was then divided by net sown area to arrive at per hectare productivity. The advantage of taking productivity per hectare of net sown area instead of gross cropped area is that it provides estimate of productivity based on the output of the whole year.

The worker per unit land was estimated as a ratio of cultivators and agricultural labours to net sown area which indicates the pressure of work force on agricultural land. Cropping intensity is the share of gross cropped area in net sown area. Similarly, irrigation coverage was estimated as the share of gross irrigated area in gross cropped area. The district wise data for estimating these variables along with fertilizer use and annual rainfall was obtained from the data set with International Crop Research Institute for Semi-Arid Tropics (ICRISAT), Hyderabad for the Year 2010-11 and 2011-12.

For the problematic soil, we relied upon the district-wise degraded and waste land statistics of National Bureau of Soil Survey and Land Use Planning (NBSS&LUP), Nagpur for the year 2009-10. NBSS&LUP classifies problematic soils into 14 categories and estimates area under each. The data on district-wise level of groundwater development was collected from central groundwater board for the year 2011.

Model estimates

Model was estimated using SAS software (SAS Code given in Appendix 1). The summary of estimated variables is given in table 1 and other model parameters are given in Appendix 2. In the first stage of the model, determinants of agricultural productivity came out to be significant and were as per the expectations. The effect of change in cropping intensity on agricultural productivity (Rs/ha) was strongest among other factors. Thus, agricultural productivity can be improved by bringing fallow land under cultivation in a year. Similarly, one per cent increase/decrease in fertilizer use, groundwater use, irrigation coverage, and rainfall would result in 0.20 per cent, 0.19 per cent, 0.18 per cent and 0.13 per cent increase/decrease in agricultural productivity, respectively. It is to be noted that irrigation development has a stronger effect on agricultural productivity as compared to rainfall. This implies that adverse effects of rainfall variation on agricultural productivity can be mitigated by improving irrigation infrastructure in the country. Thus, access to irrigation would reduce the dependency of crop production on monsoon. However, the pattern of irrigation development has remained uneven across the geographical regions and unsustainable water resource development in north-western part coexists with its under-utilization in eastern region of the country. This accentuates the regional disparity in agricultural performance and therefore emphasizes location specific strategies for equitable development in the country. The occurrence of problem soils adversely affects agricultural productivity as indicated by negative elasticity coefficient.

Table 1: Estimated parameters of two stage least square regression analysis

Stage 2		Stage 1	
Parameter	Coefficient	Parameter	Coefficient
Dependent variable		Dependent variable	
Rural poverty (%)		Agril. productivity (000, Rs/ha)	
Independent variables		Independent variables	
Intercept	38.600*** (2.182)	Intercept	-29.920*** (5.116)
Agril productivity (Rs. 000/ha)	-0.332*** (0.0391)	Cropping intensity (%)	0.310*** (0.030)
Worker per ha	1.399*** (0.297)	Irrigation coverage (%)	0.0232*** (0.039)
		Fertilizer use (kg/ha)	0.075*** (0.010)
		Rainfall (mm)	0.006*** (0.002)
		Problem soil (%)	-0.148*** (0.034)
		Groundwater development (%)	0.155*** (0.022)
R ²	0.1646	R ²	0.5511
F-value	47.68***	F-value	98.22***
No. of observations	486	No. of observations	486

Figures within parentheses are standard error of estimated coefficients, *** significant at 1 % level of significance

In the second stage, as expected, estimated coefficient of agricultural productivity was negative and significant indicating an inverse association between improvement in agricultural productivity and rural poverty. Further, elasticity estimates (Table 2) show that one per cent increase/decrease in land productivity would result in 0.80 per cent decrease/increase in rural poverty. On the other hand, a decline of one per cent in pressure of work force on agricultural land results in 0.17 per cent decrease in the rural poverty. These results indicate that improvement in agricultural productivity through technological and policy interventions, and employment diversification away from agriculture sector towards non-farm sectors would contribute positively in reducing poverty among rural households.

Table 2: Estimated elasticity per hectare of agricultural productivity and rural poverty with respect to various factors

Elasticity of rural poverty		Elasticity of agricultural productivity	
Variable	Coefficient	Variable	Coefficient
Per ha productivity	-0.80	Cropping intensity	1.08
Agril. worker/ha	0.17	Irrigation coverage	0.18
		Fertilizer use	0.20
		Rainfall	0.13
		Extent of problem soil	-0.09
		Groundwater development	0.19

REFERENCES

Chand, R. and S. K. Srivastava (2016), Disadvantaged agricultural regions: Is there a way forward, *Indian Journal of Agricultural Economics*, 71(1): 36-48.

Gujarati, D. N. (2005), Basic Econometrics, Fourth Edition, Tata McGraw-Hill Publishing Company Limited, New Delhi

Madnani, G. M. K. (2005), Introduction to Econometrics: Principles and Applications, Seventh Edition, Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi

APPENDIX I

SAS code for 2SLS model

data a;

```
input Agridependence   agriworkers      agricovergae      cropintensity
irricoverage           Fertuse         Ann Rainfall      Ruralpoor
Problesoil             Agrilprdyperha  agrkprdyperworker workerperha
GWDEV;
```

datalines;

43	64	36	111	18	193	1143	11	49	32554	22644	1.4	28
21	36	71	150	67	213	288	23	84	73546	41522	1.8	114
43	70	64	133	25	140	663	23	62	46096	32561	1.4	82
28	42	54	156	17	29	660	6	36	26559	27444	1.0	144
46	66	80	146	5	98	735	29	33	42831	36519	1.2	45
22	46	82	178	83	145	430	20	64	66085	37326	1.8	82
49	89	45	118	13	46	632	44	49	15179	7982	1.9	29
.
.
.
.
41	71	61	126	42	109	688	25	43	27097	21799	1.2	28
18	34	73	170	48	322	1200	25	19	125444	118398	1.1	138
49	79	63	117	6	122	959	36	12	32559	25917	1.3	19
19	39	38	147	36	117	1538	6	32	53785	16830	3.2	56

;

run;

data a ; set a;

prdtperha= Agrilprdyperha/1000;

workerperhal = workerperha*100

run;

proc syslin data = a 2sls;

endogenous prdtperha;

instruments cropintensity irricoverage Fertuse AnnRainfall Problesoil GWDEV;

poor: model Ruralpoor = prdtperha workerperhal ;

prod: model prdtperha = cropintensity irricoverage Fertuse AnnRainfall Problesoil GWDEV;

run;

APPENDIX II*The SAS System**The SYSLIN Procedure**Two-Stage Least Squares Estimation*

Model	POOR
Dependent Variable	Rural poor

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	29230.45	14615.23	47.68	<.0001
Error	484	148351.8	306.5120		
Corrected Total	486	168949.0			

Root MSE	17.50748	R-Square	0.16460
Dependent Mean	26.42505	Adj R-Sq	0.16115
Coeff Var	66.25336		

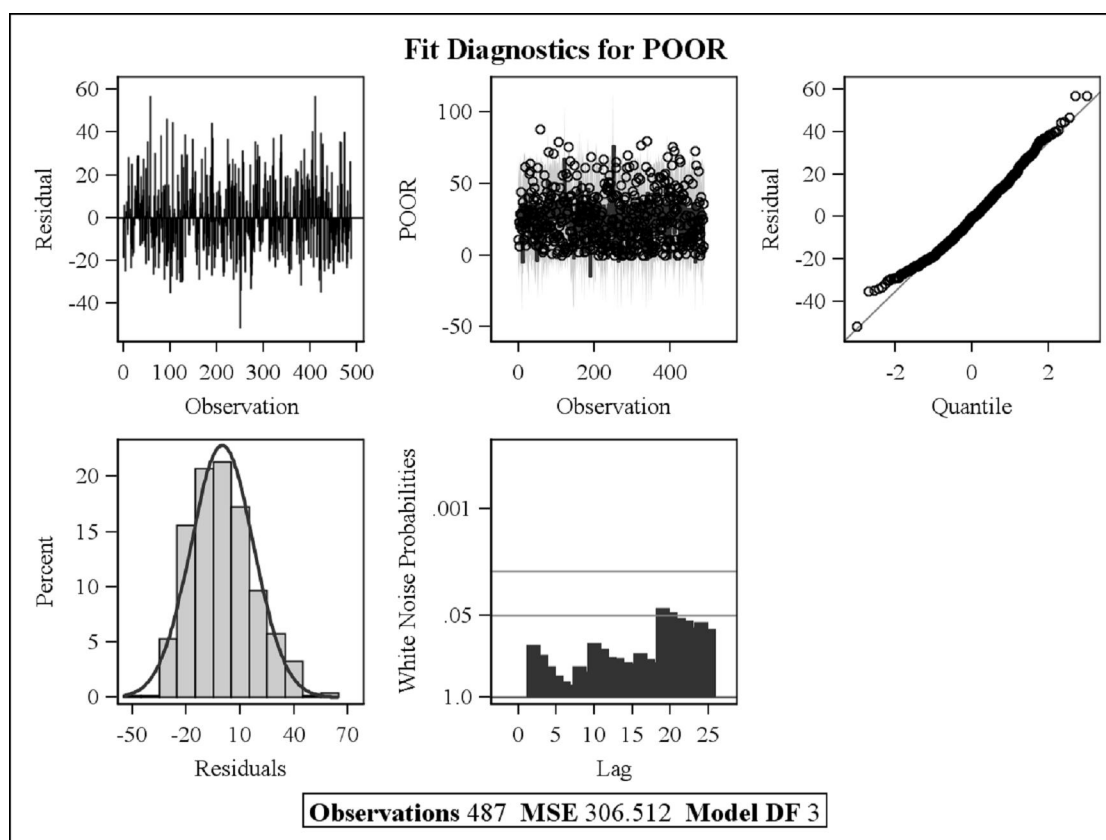
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	38.60046	2.182882	17.68	<.0001
prdtperha	1	-0.33174	0.039265	-8.45	<.0001
workerperha	1	1.398885	0.296670	4.72	<.0001

Model	PROD
Dependent Variable	prdtperha

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	198897.0	33149.51	98.22	<.0001
Error	480	162002.7	337.5055		
Corrected Total	486	360899.7			

Root MSE	18.37132	R-Square	0.55111
Dependent Mean	47.57329	Adj R-Sq	0.54550
Coeff Var	38.61689		

Root MSE		18.37132	R-Square	0.55111	
Dependent Mean		47.57329	Adj R-Sq	0.54550	
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-29.9202	5.116145	-5.85	<.0001
cropintensity	1	0.310336	0.030393	10.21	<.0001
irricoverage	1	0.232395	0.038675	6.01	<.0001
Fertuse	1	0.074875	0.009616	7.79	<.0001
AnnRainfall	1	0.005792	0.001582	3.66	0.0003
Problesoil	1	-0.14832	0.034129	-4.35	<.0001
GWDEV	1	0.155167	0.021485	7.22	<.0001



Chapter 12

DISCRIMINANT FUNCTION ANALYSIS

Achal Lama, K. N. Singh, R. S. Shekhawat, Kanchan Sinha and Bishal Gurung

INTRODUCTION

Discriminant function analysis (DFA) is primarily used for predicting the probability of an observation belonging to a given class or category based on one or multiple predictor variables (continuous and/or categorical). According to Friedman (1989), linear and quadratic discriminant analysis for small sample high dimensional setting account for a considerable gain in classification accuracy. It differs from logistic regression on the grounds that, the DFA can be used for predicting the category of an observation in the situation where the outcome variable contains more than two classes i.e., for multi-class classification problems. It is noteworthy that, both logistic regression and DFA can be used for binary classification tasks. It can be stated that the main aim of a DFA is to predict group membership based on a linear combination of the interval variables. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known. The second purpose of DFA is to better understand the data set. It can be done by careful examination of the resultant prediction model and the variables used to predict group membership. The manner of interactions among the interval variables allows a greater insight and simplification of the multivariate data set. DFA, based on matrix theory is inherent with the advantage of a clearly defined decision-making process.

DFA is assumed to be a multivariate technique for describing a mathematical function that will distinguish among predefined groups of samples. As a technique which makes use of the eigen values, its similarity can be established with multiple regression and principal components analysis. In addition, DFA is the counterpart to ANOVA and MANOVA. In DFA, continuous variables (measurements) are used to predict a categorical variable (group membership), whereas ANOVA and MANOVA use a categorical variable to explain variation in (predict) one or more continuous variables. Let us consider two examples to understand the usefulness of DFA.

Firstly, let us consider a situation where we have a series of morphological measurements on several species and we want to know how well those measurements allow these species to be classified uniquely. One approach would be to perform a series of t-tests or ANOVAs to test for differences among the species, but this would be tedious,

especially if there are many variables. Another approach can be a principal components analysis to see how the groups plot in multidimensional space and this is often a good exploratory approach. DFA takes a similar approach to PCA, but DFA seeks a linear function that will maximize the differences among the groups. The function will show how well the species can be distinguished, as well as where the classification is robust and where it may fail.

For the second example let us, consider Fisher's (1936) classic example of DFA involving three varieties of iris and four predictor variables (petal width, petal length, sepal width, and sepal length). Fisher not only wanted to determine if the varieties differed significantly on the four continuous variables, but he was also interested in predicting variety classification for unknown individual plants. DFA can be used to find a function that uses these measurements of predictor variables to separate the observation into the varieties to which they belong. That function can then be applied to the predictor variables to predict which variety is the source of each predictor variable.

Assumptions of DFA

DFA is a parametric method and it holds several assumptions. The most important assumptions are:

- The groups must be mutually exclusive.
- The number of cases for each group must not be greatly different.
- The cases must be independent.
- DFA performs better as sample size increases. A good guideline is that there should be at least four times as many samples as there are independent variables
- Discriminant function analysis is highly sensitive to outliers. Each group should have the same variance for any independent variable (that is, be homoscedastic), although the variances can differ among the independent variables. For many types of data, a log transformation will make the data more homoscedastic (that is, have equal variances).
- The predictor variables must follow multivariate normal distribution.

Types of DFA

There are various types of DFA in use, but in this chapter we will focus mainly on Linear Discriminant Analysis (LDA). The other types DFA such as Quadratic Discriminant Analysis (QDA), Mixture Discriminant Analysis (MDA) and Regularized Discriminant Analysis (RDA) will be dealt in briefly. To begin with the LDA algorithm starts by finding directions that maximize the separation between classes, and then use these directions to predict the class of individuals. These directions, called linear discriminants, are linear combinations of predictor variables. Discriminant scores are

calculated for each observation for each class based on these linear combinations. The scores are calculated using the below equation:

$$\Delta_j(Y) = \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \mu_j^T \Sigma^{-1} Y + \ln(\pi_j)$$

where Δ_j is the discriminant score for class j

Y is matrix of independent variables

μ_j is the vector containing the means of each variable for class j

Σ is the covariance matrix of the variables (assumed to be same for all classes)

π_j is the prior probability that an observation belongs to class j

LDA assumes that predictors are normally distributed (Gaussian distribution) and that the different classes have class-specific means and equal variance/covariance.

Before performing LDA, we need to make sure that the univariate distribution of each variable is normally distributed. If not, we can transform them using log and root for exponential distributions and Box-Cox for skewed distributions. Removing of outliers from the data and standardizing the variables to make their scale comparable is a must.

QDA is considered to be bit more flexible than LDA, as it allows the covariance matrix to be different for each class. LDA tends to be a better than QDA for a small training set. In contrary, QDA is to be used if the training set is very large, so that the variance of the classifier is not a major issue, or if the assumption of a common covariance matrix for the K classes is clearly untenable (James *et al.*, 2014).

FDA is a flexible extension of LDA that uses non-linear combinations of predictors such as splines. FDA is useful to model multivariate non-normality or non-linear relationships among variables within each group, allowing for a more accurate classification.

RDA builds a classification rule by regularizing the group covariance matrices (Friedman, 1989) allowing a more robust model against multicollinearity in the data. This might be very useful for a large multivariate data set containing highly correlated predictors. Regularized discriminant analysis is a kind of a trade-off between LDA and QDA. QDA assumes different covariance matrices for all the classes. Regularized discriminant analysis is an intermediate between LDA and QDA. RDA shrinks the separate covariance of QDA toward a common covariance as in LDA. This improves the estimate of the covariance matrices in situations where the number of predictors

is larger than the number of samples in the training data, potentially leading to an improvement of the model accuracy.

ILLUSTRATION

For the illustration purpose we will analyse the famous Fisher's (1936) classic dataset (<https://archive.ics.uci.edu/ml/datasets/iris>) involving three varieties of iris and four predictor variables (petal width, petal length, sepal width, and sepal length). The same dataset was taken up earlier to highlight the usefulness of DFA. The analysis is done using R software. Following are the codes and outputs obtained

Table 1: Predictor varieties for three species of iris

S. No.	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
.
.
.
.
51	7	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
.
.
101	6.3	3.3	6	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
.
.
148	6.5	3	5.2	2	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3	5.1	1.8	virginica

R Code:

```
install.packages("MASS")
library(MASS)
data(iris)
write.csv(iris,"iris data.csv")
# set the seed to make your partition reproducible
set.seed(123)
smp_size<- floor(0.80 * nrow(iris))
```

```

train_ind<- sample(seq_len(nrow(iris)), size = smp_size)
train.data<- iris[train_ind, ]
test.data<- iris[-train_ind, ]
# Fit the LDA model
model<- lda(Species~., data = train.data)
model
# Make predictions
predictions<- predict(model,test.data)
predictions
write.csv(predictions,"predictions.csv")
# Model accuracy
mean(predictions$class==test.data$Species)
#Fit the QDA model
Model1<- qda(Species~., data = train.data)
Model1
# Make predictions
predictions<- predict(Model1,test.data)
# Model accuracy
mean(predictions$class == test.data$Species)

```

Output:

```
lda(Species ~ ., data = train.data)
```

Prior probabilities of groups:

setosa	versicolor	virginica
0.35	0.34	0.31

Group means:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Setosa	5.01	3.47	1.45	0.25
versicolor	5.94	2.75	4.29	1.33
virginica	6.58	2.97	5.52	1.98

Coefficients of linear discriminants:

	LD1	LD2
Sepal.Length	0.60	-0.05
Sepal.Width	1.71	-2.29

Petal.Length -2.04 0.66

Petal.Width -2.80 -2.39

Proportion of trace:

LD1 LD2

0.99 0.01

QDA model

```
qda(Species ~ ., data = train.data)
```

Prior probabilities of groups:

setosa	versicolor	virginica
0.35	0.34	0.31

Group means:

	Sepal. Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.01	3.47	1.45	0.25
versicolor	5.94	2.75	4.29	1.33
virginica	6.58	2.97	5.52	1.98

The results thus obtained clearly indicate that the LD1 is explaining 99 per cent of the variability present in the dataset. Hence we can proceed with using just the LD1. In addition to it we can also obtain the variables which contributing positively and negatively for classifying the observations to various classes (3 in this case). Sepal width and sepal length are characters which have positive impact on classification and the remaining two have negative effect. Further, when the model was tested on the test dataset (20%), the probability of classification was found to be 1. This indicates that the model built upon training data set (80%) was appropriate.

Thus, various extensions of the DFA have been discussed along with the implementation to the famous Fisher's (1936) classic dataset. This chapter also highlights the various assumptions required for implementing the DFA. It can be said that this DFA technique is very useful method not only for classification, but also for prediction the class or group of a new observation.

REFERENCES

- Fisher, R. A. (1936), The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7(2): 179-188.
- Friedman, J. H. (1989), Regularized discriminant analysis. *Journal of the American Statistical Association*, 84 (405): 165–75.
- James, G., D. Witten, T. Hastie and R. Tibshirani (2014), An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated.

PART III

TIME SERIES ANALYSIS

Chapter 13

PRICE FORECASTING USING ARIMA MODEL

Raka Saxena, Ranjit Kumar Paul and Rohit Kumar

INTRODUCTION

In recent years, the issue of high price volatility in agricultural commodities in domestic as well as international market has assumed critical importance. Considering the extreme price situations and volatility, the market intelligence plays a significant role in farmers' decisions regarding production and marketing of agricultural commodities. Agricultural prices determine the farm income and thus have a significant impact on the farmers' well-being (Chand, 2017). Hence, volatility in agricultural commodity prices has been a major concern for policy makers in India as it significantly affects the gains to farmers (Saxena and Chand, 2017). Market Intelligence and the dissemination of market information play an important and significant role in the farmers' decisions regarding production and marketing of agricultural commodities. Thus, the availability of accurate, timely and adequate market-related information enables farmers to take informed decision as to when and where to sell their produce (Acharya, 2003).

Globally, continuous efforts are being made to capture the trends in prices and commodity outlook/projections to gain from the market dynamics in terms of demand-supply interplay. Three organizations, namely, OECD-FAO, Food and Agricultural Policy Research Institute (FAPRI) and the US Department for Agriculture (USDA) carry out the agricultural outlook exercise and provide the annual and medium term projections for selected agricultural commodities globally. In India, 14 Market Intelligence Units (MIU) were established by the Directorate of Economics and Statistics as early as in 1954. The market intelligence activities are now being given strong emphasis by the Government of India and efforts are being made to institutionalize the capacity within the institutions to carry out the efforts in long run. After a prolonged gap, the next initiative to provide the farmers with the agricultural price information was led by the Indian Council of Agricultural Research (ICAR) in 2009. ICAR implemented the sub-project "Establishing and Networking of Agricultural Market Intelligence Centres in India" as part of the National Agricultural Innovation Project (NAIP) with the objective to establish an institutionalized network of Agricultural Market Intelligence Centres in India; enabling and empowering farmers and entrepreneurs. It aimed at providing up-to-date information on prices and other market factors enabling farmers to negotiate with the traders and also facilitating spatial distribution of products among markets. Therefore, the ICAR set up the Domestic and Export Market Intelligence Cell (DEMIC) in order to provide price forecasts for various crops (FICCI, 2017). The DEMICs were

established to help the farmers to realise higher prices, provide improved regional linkages to generate, disseminate, and sharing market information for better decision making and also improve the access and use of market intelligence to all stakeholders in the marketing chain for better production and marketing strategies. Scientists from the state agricultural universities (SAUs) who were involved as collaborating centres, developed the price forecasts whereas the lead team, based in Tamil Nadu Agricultural University (TNAU), monitored the overall implementation of the project activities. The team regularly brought out pre-sowing and pre-harvest price forecasts for 34 crops (including cereals, pulses, oilseeds, cotton, vegetables and spices) with 90 to 100 per cent accuracy, the price forecasts were widely disseminated through print and visual media, mobile applications, radio broadcasts and also through tie-ups with organizations having networks with farmers. Regular feedbacks were received from stakeholders and analysed (Acharya, 2017).

It was further suggested that the market intelligence should be continued at ICAR-National Institute of Agricultural Economics and Policy Research (NIAP). This work was continued at NIAP with ICAR funding with a network of 14 institutions (covering 12 state/central agricultural universities and 2 ICAR institutes) and was named 'Network Project on Market Intelligence'. The project intended to provide the short-term price forecasts of regionally important commodities and also focused on policy studies with relevance to price behaviour, price transmission, market infrastructure along with market linkages. More than 40 agricultural commodities, most of them being high-value commodities (horticulture), were selected to provide reliable and timely price forecasts to farmers in 13 major states across the country. To create longer and larger impact and acceptability in the system, e-solution for market intelligence can be developed by combining various algorithms of suitable techniques and models in single software package, which would be easy to use even by the line departments.

Various Statistical Approaches

Various statistical approaches viz, regression, time series, stochastic and, of late, machine learning approaches are in vogue for statistical modeling. Some of the tools and models which can be used for time series analysis, modeling and forecasting are briefly discussed in the following sections. However, the same cannot be claimed to be complete and exhaustive. Every approach has its own advantages and limitations. Engle (1982) proposed the Autoregressive Conditional Heteroscedastic (ARCH) families of parametric nonlinear time-series models. It captures the volatility in prices through inbuilt modeling mechanisms. More recently, Artificial Neural Networks (ANN) have been studied as an alternative to these non-linear model driven approaches. ANN belongs to the data-driven approach, i.e. the analysis depends on the available data, with little apriori rationalization about relationships between variables and about the models. These models typically utilize a host of empirical data and attempt to forecast market behaviour and estimate future values of key variables by using past values of core economic indicators.

For many agricultural commodities, data are usually collected over time. Forecasts for them can be obtained using different modeling techniques; however, the choice of the method depends on the purpose and importance of the forecasts as well as the costs involved in using the alternative methods. The most widely used technique for analysis of time-series data is; the Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology, as these models are found to be more flexible in handling different patterns of time series data. In this chapter, the 'Univariate' Box-Jenkins models, also referred to as ARIMA models are discussed. Univariate or single series means that the forecasts are based only on past values of a single variable and not on any other data series.

An important characteristic of time series data is that the successive observations are dependent on the past values of the series. Each observation of the observed data series, Y_t , may be considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t, t \in T\}$, where $T = \{0, \pm 1, \pm 2, \dots\}$, and applying standard time-series approach to develop an ideal model will adequately represent the set of realizations and also their statistical relationships in a satisfactory manner.

We denote by Y_t , the observation made at time t ($t = 1, 2, \dots, n$). Thus, a time-series involving n points may be represented as sequence of n observations (Y_1, Y_2, \dots, Y_n). The statistical analysis of time series data differs from the classical regression analysis. Time series data typically violates the assumption that the error terms/successive observations are uncorrelated with each other. This effect, known as autocorrelation, biases the standard error associated with regression slope parameters estimates and makes the relevant t -test invalid. Contrary to the assumption of statistical independence of observations in the classical regression analysis, we assume that the time sequenced observations ($Y_1, Y_2, \dots, Y_{t-1}, Y_t, Y_{t+1}, \dots$) may be statistically related to the past observations in the same series, in the Box-Jenkins method. Our objective is to specify the relationship between the time series observations in the form of a statistical model.

Stationarity of a time series process

A time series 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.

In general, unit root tests are used to test the stationarity of a series. A statistical test for stationarity or test for unit root has been proposed by Dickey and Fuller (1979). This test is also referred to as the Augmented Dickey Fuller (ADF) test. The test is applied for the parameter in the auxiliary regression

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

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

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

Autocorrelation functions

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

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

It ranges from -1 to $+1$. Box and Jenkins have suggested that maximum number of useful r_p is roughly $N/4$ where N is the number of periods upon which information on y_t is available.

Partial autocorrelations are used to measure the degree of association between y_t and y_{t-p} when the y -effects at other time lags $1, 2, 3, \dots, p-1$ are removed.

Description of ARIMA Models

Autoregressive (AR) Model

Autoregressive terms mention the lags of the stationary series. In this model, the current value of the process is expressed as a finite, linear aggregate of lagged values of the process along with the shock ε_t . Let us denote the values of a process at equally spaced time epochs $t, t-1, t-2, \dots$ by $y_t, y_{t-1}, y_{t-2}, \dots$, then can be described by the following expression:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

If we define an autoregressive operator of order p by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p,$$

Where, B is the backshift operator such that $B y_t = y_{t-1}$, the autoregressive model can be written as $\phi(B) y_t = \varepsilon_t$.

Moving Average (MA) Model

Another kind of model of great practical importance in the representation of observed time-series is the finite moving average process. A moving average series essentially

refers to the lags of the forecast error terms. MA (q) model is defined as

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

If we define a moving average operator of order q by

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

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

Autoregressive Moving Average (ARMA) Model

To achieve greater flexibility in fitting of actual time-series data, it is sometimes advantageous to include both autoregressive and moving average processes. This leads to the mixed autoregressive-moving average model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

or

$$\phi(B)y_t = \theta(B)\varepsilon_t.$$

This is written as ARMA (p, q) model. In practice, it is frequently true that adequate representation of actually occurring stationary time-series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than 2 and often less than 2.

Autoregressive Integrated Moving Average (ARIMA) Model

A generalization of ARMA models which incorporates a wide class of non-stationary time-series is obtained by introducing the differencing into the model. The simplest example of a non-stationary process which reduces to a stationary one after differencing is Random Walk. A process $\{y_t\}$ is said to follow an Integrated ARMA model, denoted by ARIMA (p, d, q), if $\nabla^d y_t = (1 - B)^d \varepsilon_t$ is ARMA (p, q). The model is written as

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

Where, $\varepsilon_t \sim WN(0, \sigma^2)$, WN indicating white noise. The integration parameter d is a nonnegative integer. When $d = 0$, ARIMA (p, d, q) \equiv ARMA (p, q).

The ARIMA methodology is carried out in three stages, viz. identification, estimation and diagnostic checking. Parameters of the tentatively selected ARIMA model are estimated at the estimation stage. Adequacy of the selected model is tested at the diagnostic checking stage. If the model is found to be inadequate, all the three stages are repeated till the selected ARIMA model provides satisfactory results for the data series under consideration. An excellent discussion of various aspects of this approach is given in Box *et al.* (2007). Most of the standard

software packages, like SAS, SPSS, R and E Views contain programs for fitting the ARIMA models. Most of the standard software packages, like SAS, SPSS, R and EViews contain programs applicable for the ARIMA models. In this chapter EViews software has been used to forecast the monthly wholesale price of mustard for Bharatpur Market, Rajasthan

A typical ARIMA model has three components p , d and q

Theoretical ACFs and PACFs (Autocorrelations versus lags) are available for the various models chosen (Pankratz, 1983) for various values of orders of autoregressive and moving average components i.e. p and q . Thus, compare the correlograms (plot of sample ACFs versus lags) obtained from the given TS data with these theoretical ACF/PACFs to find a reasonably good match and tentatively select one or more ARIMA models. The general characteristics of theoretical ACFs and PACFs are as follows: (here spike represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

Table 1: Thumb rule for identifying the ARMA or ARIMA model

Model	ACF	PACF
AR	Spikes decay towards zero	Spikes cutoff to zero
MA	Spikes cutoff to zero	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

Model building

Identification

The preliminary step for modelling is to check the stationarity of the series. If the original series is non-stationary, then we have to make the series stationary as the estimation procedures are offered only for stationary series. A non-stationary series may be stationarized by differencing, logging, deflating and other transformation techniques. Further in this process, the initial values for the orders of AR(p) and MA(q) can be obtained by examining the significant autocorrelation and partial autocorrelation coefficients. Say, if first order auto correlation coefficient is significant and data is stationary at level, then an AR(1), or MA(1) or ARMA(1, 0, 1) model could be tried to start with. This is not a hard and fast rule, as sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients. Still, they can be used as initial values, while the final models are achieved after going through the stages repeatedly.

Estimation

At the identification stage, one or more models are tentatively selected that may provide statistically acceptable representation of the data series. After that, we attempt to obtain precise estimates of parameters in the model by following Box and Jenkins

approach. Standard computer packages like SAS, SPSS, EViews etc. are available for estimation of relevant parameters using iterative procedures.

Diagnostics

Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained based on the following diagnostics.

- (a) Low Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC)/ Schwarz-Bayesian Information Criteria (SBC)
- (b) Plot of residual ACF

Once the appropriate ARIMA model has been estimated, the goodness of fit can be investigated by plotting the ACF of residuals of the selected model. If most of the sample autocorrelation coefficients of the residuals are within the limits i.e., in the range of where, N is the number of observations used for model then the residuals $\pm 1.96 / \sqrt{N}$ are white noise signifying that the model is a good fit.

- (c) Non-significance of auto correlations of residuals via Portmanteau tests (Q-tests based on Chisquare statistics)-Box-Pierce or Ljung-Box tests

After tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and, if need be, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity Q known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square.

Validation

It is important to evaluate/validate the forecasts obtained in terms of accuracy of the predicted values. The commonly used measures for validation are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Standard Error. MAE and RMSE are based on the forecast error, which is the difference between the actual price and the forecasted price. MAE is the mean of the absolute value of the error terms, while, RMSE is given by the square root of the mean of the squared error terms. Sometimes, the mean absolute per cent error (MAPE) is also used, as it is a scale independent measure. Standard Error of a forecast is the standard deviation of the error term estimated using the sample mean and it provides the precision of the forecasts.

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i|$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2}$$

Where, e_i are error terms calculated as ($e_i, i= 1,2,\dots, n$)

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

Where, A_t is the actual value and F_t is the forecast value

Generally, it is recommended not to use the entire data series for forecasting and rather to hold a set of time series observations (may be recent 4-5 price observations) and develop the forecasts based on the remaining historical observations. The ‘hold out set’ may be used to evaluate the forecasting model using the forecast error based measures. If one is interested in slightly long forecasting, say 12 months, the short term forecast of 4-5 time periods need to be developed first. These forecasts are plugged in the original dataset and forecasts are developed based on this new dataset. The process is repeated once more to get the complete forecasts of 12 months.

ILLUSTRATION

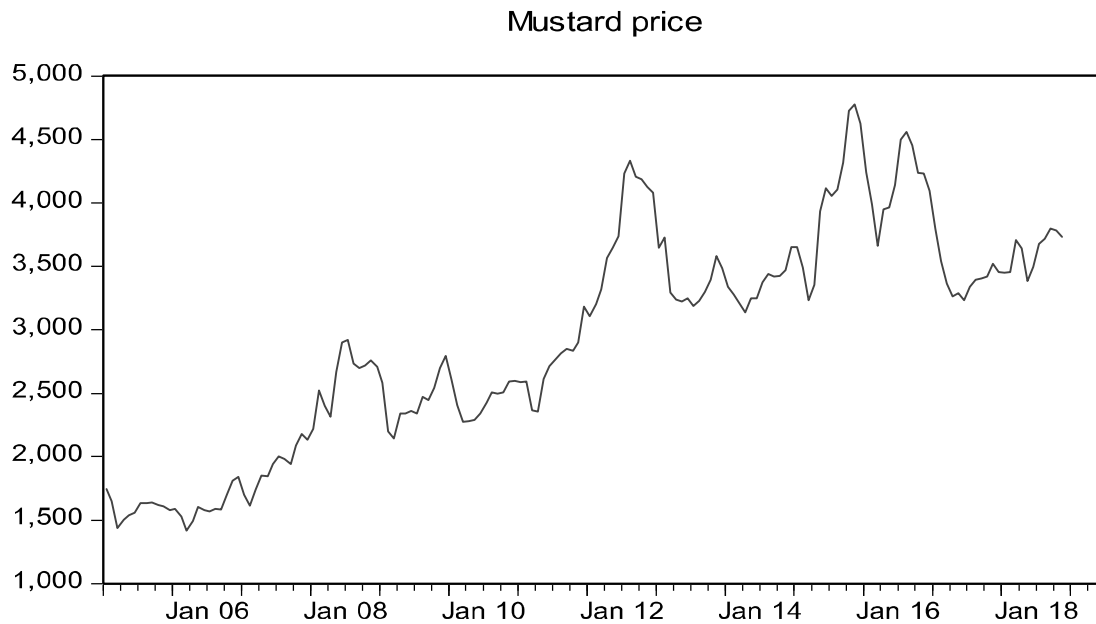
Application of ARIMA Using E-Views Software

Step1: Import data file in Eviews

1. Prepare excel sheet of monthly mustard price data (Mustard price)
2. File – Open – Foreign data as work file- Desktop- file Open (excel sheet)
 - Data : Monthly wholesale price of Mustard for Bharatpur Market, Rajasthan
 - Time Period: January, 2005 to November, 2018 (174 Observations)
 - Source: AGMARKNET Website

Step 2: Sequence Chart of the data

Click mustard price sheet in Eviews – View –Graph - Time axis label (Date) - Ok



Step 3: Checking Stationarity of the series

Stationarity at level

Null Hypothesis: MUSTARD_PRICE has a unit root

Exogenous: Constant, Linear Trend

Lag Length: 1 (Automatic - based on SIC, maxlag=13)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-3.117051	0.1058
Test critical values: 1% level	-4.014635	
5% level	-3.437289	
10% level	-3.142837	

*MacKinnon (1996) one-sided p-values.

Step 4: Stationarity of the series (1st Differencing) and Identification of the Model

Stationarity at 1st difference

Null Hypothesis: D(MUSTARD_PRICE) has a unit root

Exogenous: Constant, Linear Trend

Lag Length: 0 (Automatic - based on SIC, maxlag=13)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-9.441007	0.0000
Test critical values: 1% level	-4.014635	
5% level	-3.437289	
10% level	-3.142837	

*MacKinnon (1996) one-sided p-values.

Autocorrelation and Partial autocorrelation function

Date: 09/25/19 Time: 11:10

Sample: 2005M01 2019M06

Included observations: 166

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. **	. **	1	0.292	0.292	14.374	0.000
. .	* .	2	-0.048	-0.146	14.771	0.001
* .	. .	3	-0.071	-0.014	15.643	0.001
. .	. .	4	0.039	0.067	15.901	0.003
. .	* .	5	-0.035	-0.088	16.111	0.007
* .	* .	6	-0.104	-0.068	18.008	0.006
* .	. .	7	-0.075	-0.022	18.987	0.008
** .	** .	8	-0.210	-0.233	26.794	0.001
* .	. .	9	-0.151	-0.035	30.843	0.000
. .	. *	10	0.059	0.105	31.468	0.000
. *	. *	11	0.178	0.088	37.150	0.000
. *	. .	12	0.105	0.049	39.163	0.000
. .	. .	13	0.070	0.067	40.051	0.000
. *	. *	14	0.129	0.076	43.122	0.000
. .	* .	15	-0.034	-0.144	43.341	0.000
** .	** .	16	-0.229	-0.232	53.072	0.000
* .	* .	17	-0.175	-0.089	58.799	0.000
. .	. .	18	-0.026	0.020	58.927	0.000
. .	. .	19	-0.028	0.031	59.076	0.000
* .	. .	20	-0.149	-0.064	63.313	0.000
* .	. .	21	-0.107	-0.028	65.526	0.000
* .	* .	22	-0.068	-0.080	66.410	0.000
. .	. .	23	0.062	0.022	67.155	0.000
. *	. .	24	0.151	-0.008	71.605	0.000
. *	* .	25	0.075	-0.112	72.723	0.000
. .	. .	26	-0.050	-0.051	73.230	0.000
* .	. .	27	-0.131	-0.060	76.690	0.000
* .	* .	28	-0.156	-0.181	81.589	0.000
. .	. .	29	-0.037	0.024	81.870	0.000
. .	* .	30	-0.062	-0.068	82.647	0.000
* .	* .	31	-0.139	-0.131	86.622	0.000
. .	. .	32	-0.045	0.039	87.049	0.000
. .	* .	33	0.013	-0.069	87.087	0.000
. *	. *	34	0.146	0.114	91.610	0.000
. *	. .	35	0.096	-0.027	93.558	0.000
. *	. .	36	0.153	0.023	98.559	0.000

Price Forecasting Using Arima Model

Step 5: Estimation of the ARIMA Model

Model Fit

ARIMA (1, 1, 0)

Dependent Variable: D(MUSTARD_PRICE)

Method: Least Squares

Date: 09/24/19 Time: 15:27

Sample (adjusted): 2005M03 2018M11

Included observations: 165 after adjustments

Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.69229	16.17097	0.784881	0.4337
AR(1)	0.291941	0.074847	3.900511	0.0001
R-squared	0.085369	Mean dependent var		12.59945
Adjusted R-squared	0.079758	S.D. dependent var		153.3187
S.E. of regression	147.0775	Akaike info criterion		12.83184
Sum squared resid	3525981.	Schwarz criterion		12.86949
Log likelihood	-1056.627	Hannan-Quinn criter.		12.84713
F-statistic	15.21399	Durbin-Watson stat		1.914298
Prob(F-statistic)	0.000140			
Inverted AR Roots	.29			

ARIMA (1, 1, 2)

Dependent Variable: D(MUSTARD_PRICE)

Method: Least Squares

Date: 09/24/19 Time: 15:28

Sample (adjusted): 2005M03 2018M11

Included observations: 165 after adjustments

Convergence achieved after 37 iterations

MA Backcast: 2005M01 2005M02

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	14.41443	4.151038	3.472489	0.0007
AR(1)	0.898679	0.040956	21.94255	0.0000
MA(1)	-0.629653	0.079007	-7.969585	0.0000
MA(2)	-0.354453	0.076540	-4.630936	0.0000
R-squared	0.131765	Mean dependent var		12.59945
Adjusted R-squared	0.115587	S.D. dependent var		153.3187
S.E. of regression	144.1859	Akaike info criterion		12.80403
Sum squared resid	3347119.	Schwarz criterion		12.87932

Log likelihood	-1052.332	Hannan-Quinn criter.	12.83459
F-statistic	8.144585	Durbin-Watson stat	1.953212
Prob(F-statistic)	0.000044		
<hr/>			
Inverted AR Roots	.90		
Inverted MA Roots	.99	-.36	

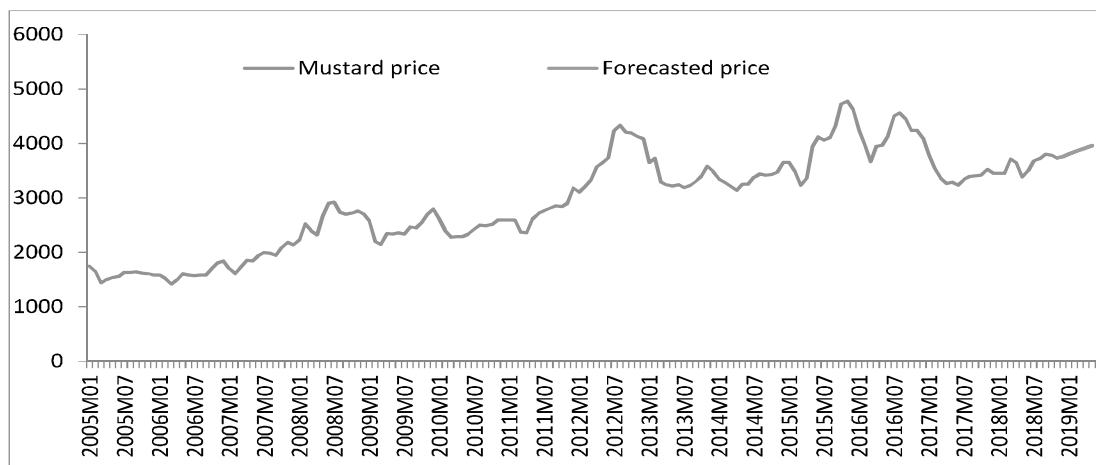
Step 6: Forecasting and model validation of the ARIMA (1, 1, 2)

Validation of model ARIMA (1, 1, 2)

Month	Actual Price (Rs./qtl)	Forecasted Price (Rs./qtl)
Jun-18	3,496	3374
Jul-18	3,677	3450
Aug-18	3,719	3520
Sep-18	3,800	3584
Oct-18	3,785	3643
Nov-18	3,731	3697
MAPE		4.23

Step 7: Forecast Results of ARIMA (1, 1, 2) model

Month	Forecasted Price (Rs./qtl)
Dec-18	3754
Jan-19	3801
Feb-19	3845
Mar-19	3886
Apr-19	3924
May-19	3960



REFERENCES

- Acharya, S. S. (2003), Analytical framework for review of agricultural marketing institutions. In: Institutional Change in Indian Agriculture, (Pal, S., Mruthyunjaya, Joshi, P. K. and Saxena, R., eds.), NCAP, New Delhi.
- Acharya, S. S. (2017), Effective implementation of agricultural price and marketing policy incomes: doable priority actions for doubling farmers. *Agricultural Economics Research Review*, 30 (Conference Issue): 1-12.
- Box, G. E. P., Jenkins, G. M. and G. C. Reinsel (2007), Time-series analysis: forecasting and control. 3rd edition. Pearson education, India.
- Chand, R. (2017), Doubling farmers' income: rationale, strategy, prospects and action plan. NITI Aayog, New Delhi.
- Dickey, D. A. and W. A. Fuller (1979), Distribution of the estimators for the autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74: 427-31.
- Engle, R. F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50: 987-1008.
- FICCI (2017), Agriculture marketing: an overview and way forward. A knowledge paper on agriculture marketing. Federation of Indian Chambers of Commerce and Industry
- Pankratz, A. (1983), Forecasting with univariate Box – Jenkins models: concepts and cases. John Wiley, New York, U.S.A.
- Saxena, R. and R. Chand (2017), Understanding the recurring onion price crisis: revelations from production-trade-price linkages. Policy paper 33, ICAR-National Institute of Agricultural Economics and Policy Research (NIAP), New Delhi.

Chapter 14

VOLATILITY MODELS

Girish Kumar Jha and Achal Lama

INTRODUCTION

In conventional econometric models, the variance of the disturbance term is assumed to be constant (Makridakis, 1998 and Pankratz, 1983). In many practical applications, this assumption may not be realistic. For instance, many economic time series exhibit phases of relative tranquillity followed by periods of high volatility. Much of the econometric research is concerned with extending the Box-Jenkins methodology to analyse these types of time series variables. A model incorporating the possibility of a nonconstant error variance is called a heteroscedasticity model.

Volatility has been one of the most active areas of research in time series econometrics in recent decades. Volatility refers to the variability of the random (unforeseen) component of time series. In economic theory, volatility connotes two principal concepts: variability and uncertainty; the former describing overall movement and the latter referring to movement that is unpredictable.

There are various ways of measuring price volatility. To begin with in the naive approach all the price movements are treated as indication of the instability by calculating standard deviation of the price index. In this approach uncertainty is overstated as it does not account for predictable components like trends in the price evolution process. This approach does not account for predictable components like trends in the price evolution process thereby overstating the uncertainty. An improvement over the naive approach is the ratio method which quantifies the price instability by measuring the standard deviation of $\log(P_t / P_{t-1})$ over a period, where P_t and P_{t-1} are the price in period t and $t-1$ respectively. In the third approach price series is divided between predictable and unpredictable components, with an assumption of price volatility being time invariant. The fourth approach is an improvement over the third one, as it not only divides the series into predictable and unpredictable components but also allows the price volatility to be time varying. Such time varying conditional variances can be estimated by using a Generalised Autoregressive Conditional Heteroscedastic (GARCH) model (Bollerslev, 1986).

The main aim of this chapter is to describe the techniques necessary to model and forecast conditional heteroscedasticity. The simplest model suggests that the conditional heteroscedasticity can be estimated as an autoregressive process (ARCH), Engle (1982). This chapter describes standard GARCH and exponential GARCH

(EGARCH) models. It concludes with an illustration of fitting GARCH and EGARCH models using real data.

GARCH MODEL

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

$$\begin{aligned} \varepsilon_t | \Psi_{t-1} &\sim N(0, h_t) \\ h_t &= a_0 + \sum_{i=1}^q a_i \varepsilon_{t-i}^2 \dots \end{aligned} \quad (1)$$

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 nonnegative and finite unconditional variance of stationary $\{\varepsilon_t\}$ series.

However, ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required. Secondly, conditional variance of ARCH(q) model has the property that unconditional autocorrelation function (Acf) of squared residuals; if it exists, decays very rapidly compared to what is typically observed, unless maximum lag q is large. To overcome the weaknesses of ARCH model, 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

$$\begin{aligned} \varepsilon_t &= \xi_t h_t^{1/2} \\ h_t &= a_0 + \sum_{i=1}^q a_i \varepsilon_{t-i}^2 + \sum_{j=1}^p b_j h_{t-j} \dots \end{aligned} \quad (2)$$

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

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

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 (2) 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 of decay is too rapid as compared to what is generally observed in a financial time-series, unless the maximum lag q is large. As (2) is a more parsimonious

model of the conditional variance than a high-order ARCH model, hence preferred mostly by the users as an alternative to ARCH.

The most popular GARCH model in applications is the GARCH(1,1) model. To express GARCH model in terms of ARMA model, denote $\eta_t = \varepsilon_t^2 - h_t$. Then from equation (2)

$$\varepsilon_t^2 = a_0 + \sum_{i=1}^{Max(p,q)} (a_i + b_i) \varepsilon_{t-i}^2 + \eta_t + \sum_{j=1}^p b_j \eta_{t-j} \dots \quad (3)$$

Thus a GARCH model can be regarded as an extension of the ARMA approach to squared series $\{\varepsilon_t^2\}$. Using the unconditional mean of an ARMA model, we have

$$E(\varepsilon_t^2) = \frac{a_0}{1 - \sum_{i=1}^{Max(p,q)} (a_i + b_i)} \dots \quad (4)$$

provided that the denominator of the prior fraction is positive.

Properties of GARCH model

The properties of GARCH models can easily be studied by focusing on the simplest GARCH (1, 1) model with

$$\begin{aligned} \varepsilon_t &= \xi_t h_t^{1/2} \\ h_t &= a_0 + a_1 \varepsilon_{t-1}^2 + b_1 h_{t-1} \dots \end{aligned} \quad (5)$$

where $\xi_t \sim \text{IID}(0,1)$ and $0 \leq a_1, b_1 \leq 1, (a_1 + b_1) < 1$.

The (1,1) in parenthesis is a standard notation in which the first number refers to number of autoregressive lags, or ARCH terms, appearing in the equation, whereas the second number refers to lags specified for moving average, which here is often called the number of GARCH terms. Occasionally models with more than one lag are needed to find good variance forecasts.

First, a large ε_{t-1}^2 or h_{t-1} gives rise to a large h_t . This means that a large ε_{t-1}^2 tends to be followed by another large ε_t^2 , generating the well known behaviour of volatility clustering in financial time-series.

Second it can be seen that if

$$1 - 2a_1^2 - (a_1 + b_1)^2 > 0,$$

then

$$\frac{E(\varepsilon_t^4)}{[E(\varepsilon_t)]^2} = \frac{3[1 - (a_1 + b_1)^2]}{1 - (a_1 + b_1)^2 - 2a_1^2} > 3 \dots \quad (6)$$

Consequently, similar to ARCH models, the tail distribution of a GARCH (1, 1) process is heavier than that of a normal distribution. Third, the model provides a simple parametric function that can be used to describe the volatility evolution.

Forecasting volatility using GARCH model

Forecasts of a GARCH model can be obtained using methods similar to those of an ARMA model. Although this model is directly set up to forecast for just one period, it turns out that based on the one-period forecast, a two-period forecast can be made. Ultimately, by repeating this step, long-horizon forecasts can be constructed. For the GARCH (1,1), the two-step forecast is a little closer to the long-run average variance than is the one-step forecast, and, ultimately, the distant-horizon forecast is the same for all time periods as long as $(a_1 + b_1) < 1$. This is just the unconditional variance. Thus, the GARCH models are mean reverting and conditionally heteroscedastic, but have a constant unconditional variance.

Consider the GARCH(1,1) model in (5) and assume that the forecast origin is t , the one-step ahead forecast is

$$h_t(1) = a_0 + a_1 \varepsilon_t^2 + b_1 h_t$$

For multi-step ahead forecasts, use $\varepsilon_t^2 = \xi_t^2 h_t$ and rewrite the volatility equation in (5) as

$$h_{t+1} = a_0 + (a_1 + b_1) h_t + a_1 h_t (\varepsilon_t^2 - 1)$$

For two-step ahead forecasts is

$$h_{t+2} = a_0 + (a_1 + b_1) h_{t+1} + a_1 h_{t+1} (\varepsilon_{t+1}^2 - 1)$$

Since

$$E((\varepsilon_{t+1}^2 - 1) / \psi_t) = 0,$$

the two-step ahead volatility forecast at the forecast origin t satisfies the equation

$$h_t(2) = a_0 + (a_1 + b_1) h_t(1)$$

In general we have

$$h_t(l) = a_0 + (a_1 + b_1) h_t(l-1), \quad l > 1 \quad \dots \quad (7)$$

Consequently, the multi-step ahead volatility forecast of a GARCH (1, 1) model converge to the unconditional variance of ε_t as the forecast horizon increases to infinity provided that $\text{Var}(\varepsilon_t)$ exists.

Estimation of parameters

In order to estimate the parameters of GARCH model, three different types of estimators are proposed in literature, namely conditional maximum likelihood estimator, White's

estimator and the least absolute deviation estimator. In this chapter we restrict our discussion to the conditional maximum likelihood estimator.

Conditional maximum likelihood estimator

Similar to the estimation for ARMA models, the most frequently used estimators for ARCH/GARCH models are those derived from a (conditional) Gaussian likelihood function.

The loglikelihood function of a sample of T observations, apart from constant, is

$$L_T(\theta) = T^{-1} \sum_{t=1}^T \left(\log h_t + \varepsilon_t^2 h_t^{-1} \right),$$

where,

$$h_t = a_0 + \sum_{i=1}^q a_i \varepsilon_{t-i}^2 + \sum_{j=1}^p b_j h_{t-j}$$

EXPONENTIAL GARCH (EGARCH) MODEL

An interesting feature of asset price is that negative news seems to have a more pronounced effect on volatility than does positive news. The EGARCH model was developed to allow for asymmetric effects between positive and negative shocks on the conditional variance of future observations. One problem with the standard GARCH model is that it is necessary to ensure that all of the estimated coefficients are positive. Nelson (1991) proposed a specification that does not require nonnegativity constraints on the parameters. In the EGARCH model, the conditional variance, h_t , is an asymmetric function of lagged disturbances. The model is given by

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

$$\ln(h_t) = a_0 + \frac{1 + b_1 B + \dots + b_{q-1} B^{q-1}}{1 - a_1 B + \dots + a_p B^p} g(\varepsilon_{t-1}) \quad \dots \quad (8)$$

where,

$$\begin{aligned} g(\varepsilon_t) &= (\theta + \gamma) \varepsilon_t - \gamma E(|\varepsilon_t|), \quad \text{if } \varepsilon_t \geq 0, \\ &= (\theta - \gamma) \varepsilon_t - \gamma E(|\varepsilon_t|), \quad \text{if } \varepsilon_t < 0, \end{aligned}$$

B is the backshift (or lag) operator such that

$$Bg(\varepsilon_t) = g(\varepsilon_{t-1})$$

The EGARCH model can also be represented in another way by specifying the logarithm of conditional variance as

$$\ln(h_t) = a_0 + \beta \ln(h_{t-1}) + \alpha \left| \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \right| + \gamma \frac{\varepsilon_{t-1}}{\sqrt{h_{t-1}}} \dots \quad (9)$$

This implies that the leverage effect is exponential, rather than quadratic, and that forecasts of the conditional variance are guaranteed to be nonnegative.

ILLUSTRATION

Illustration presented in this chapter is a part of work done by Lama, *et al.* (2015). In this study, the three sets of data are used. The series include domestic and international edible oils price indices as well as international cotton price index, 'Cotlook A' index. The base year is 2004-05 for all the three series. The 'Cotlook A' index data were collected from the commodity price bulletin, published by the United Nations Convention of Trade and Development (UNCTAD). The International edible oils price index data was collected from the World Bank Commodity Prices Indices (Pink Sheet) available from its official website. And, domestic edible oils price index data were collected from Office of the Economic Advisory, Ministry of Commerce, Government of India. Each series contains 360 data points (April, 1982 to March, 2012). Out of which, first 348 data points are used for model building purpose and rest 12 data points are kept for validation, except for the 'Cotlook A' series for which 346 data points are used for modelling and remaining 14 points for forecasting. All the characteristics of the data sets used are presented in Table 1. From, the visual inspection of time plots (Fig 1) of these series it is clear that volatility is present at several time-epochs.

The basic assumption in time series econometrics is that the underlying series is stationary in nature. Thus, the test for stationarity of the three series under consideration was done using Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) test statistics (Dickey and Fuller, 1979; Phillips and Perron, 1988). For all the three series, both the tests were found insignificant at 5% level of significance, thus confirming the non-stationarity of the level series. But, on differencing the series once, both the test was found highly significant at 1% level of significance confirming their stationarity. The detailed results of the tests are given in Table 2.

Various combinations of the ARIMA models were tried after first differencing of all the three series. Among all, the AR (1) model had minimum AIC and BIC values for all series. As, the root mean square error (RMSE) value of series were high, it confirms that the ARIMA cannot model and forecast volatile data efficiently. In addition the square of the residuals of these series had significant autocorrelation. Thus, the need of modelling these series with nonlinear models of the GARCH family was felt. The parameter estimates of the ARIMA model along with the standard errors in bracket are given in Table 3.

The basic assumption of Box-Jenkins approach is that the residuals remain constant over time. Thus, the ARCH – LM test was carried out on the square of the residuals obtained after fitting the ARIMA model on all the three series. The results of the test revealed the presence of ARCH effect for all three series (Table 4).

The GARCH model was fitted to all three series and then forecasting was done. The estimates of the parameters along with their standard errors in brackets for individual series are given in Table 5. Results revealed that domestic and international edible oils price indices exhibit a high persisting volatility as the sum of a and b are close to one. The good fit of the model for both domestic and international edible oils price indices is depicted clearly by Fig 2 and 3 respectively for both series. The modelling of ‘Cotlook A’ index series was not satisfactory, where a sudden rise in the volatility was seen which is evident from Figure 4. This motivated us to model and forecast the series using EGARCH.

To capture the asymmetric nature of volatility in the data, EGARCH model was been employed to Cotlook A index series. The AIC value for fitted EGARCH model is 2279.45, which is less than the corresponding value, 2288.88 for the fitted GARCH model. This clearly showed the superiority of EGARCH model over GARCH model for the data under consideration for modelling purposes. Fitted EGARCH model along with data is exhibited in Fig 5. Evidently, the fitted model is able to capture quite well the volatility present at time-epochs especially towards the end.

One step ahead forecasts for the monthly index of the domestic and international edible oils price for the period April, 2011 to March, 2012 along with its corresponding standard errors in parentheses is given in the Tables 6 and 8 respectively. Table 10 presents the forecast for Cotlook A index series for the period from February, 2011 to March, 2012. The forecasting ability of both the models was judged on the basis of root mean square error (RMSE) and relative mean absolute prediction error (RMAPE) and reported in Tables 7, 9 and 11 respectively. The R code used for the present analysis is appended in the end.

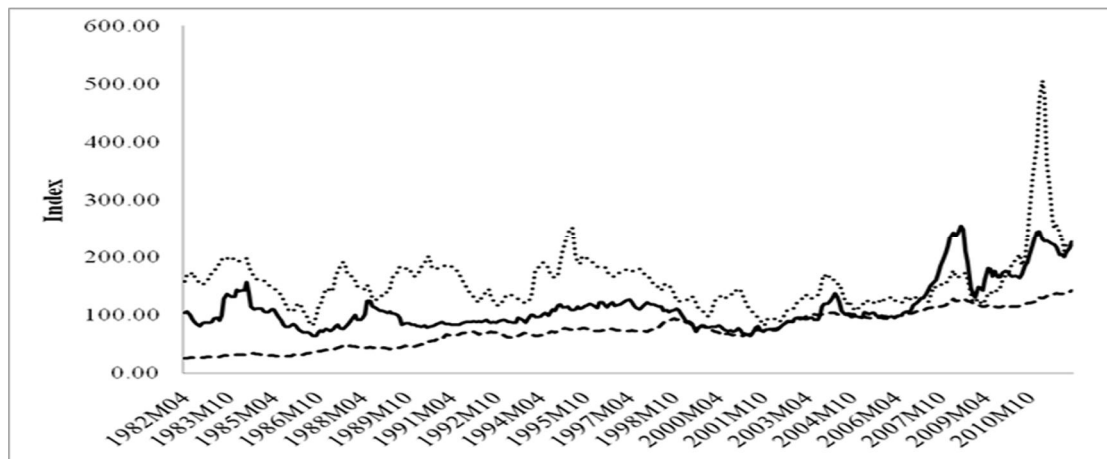


Fig 1: Fitted AR (2) – GARCH (1, 1) model along with its data points (bold) and Cotlook A (Dotted)

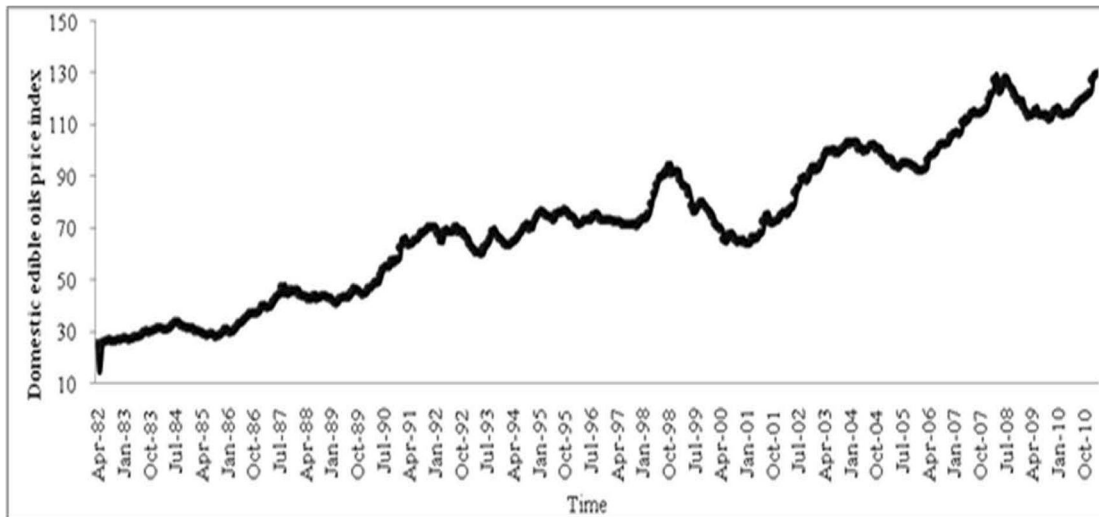


Fig 2: Time plot of the three series Domestic edible oils (dashed), International edible oils (bold) and Cotlook A (dotted)

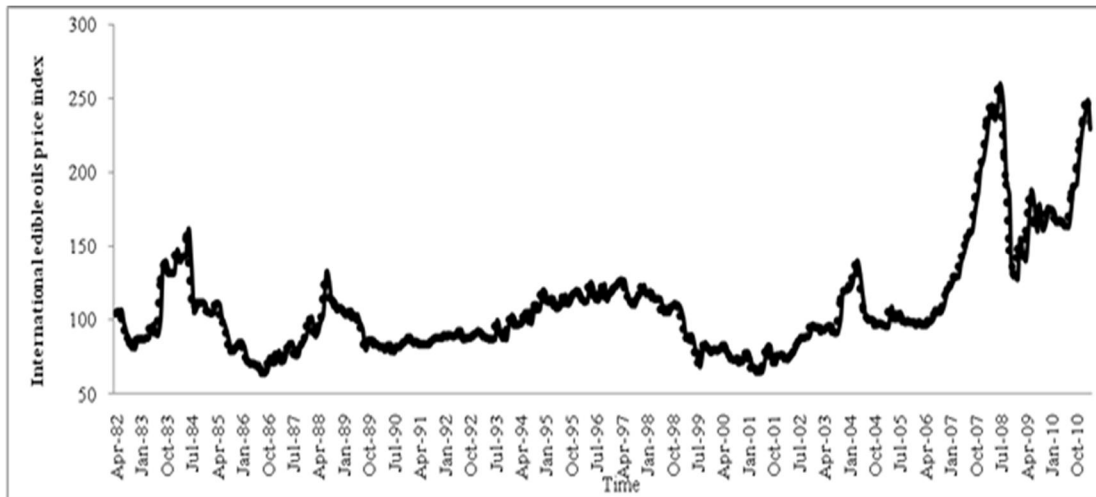


Fig 3: Fitted AR (2) – GARCH (1, 1) model along with its data points look A (dotted)

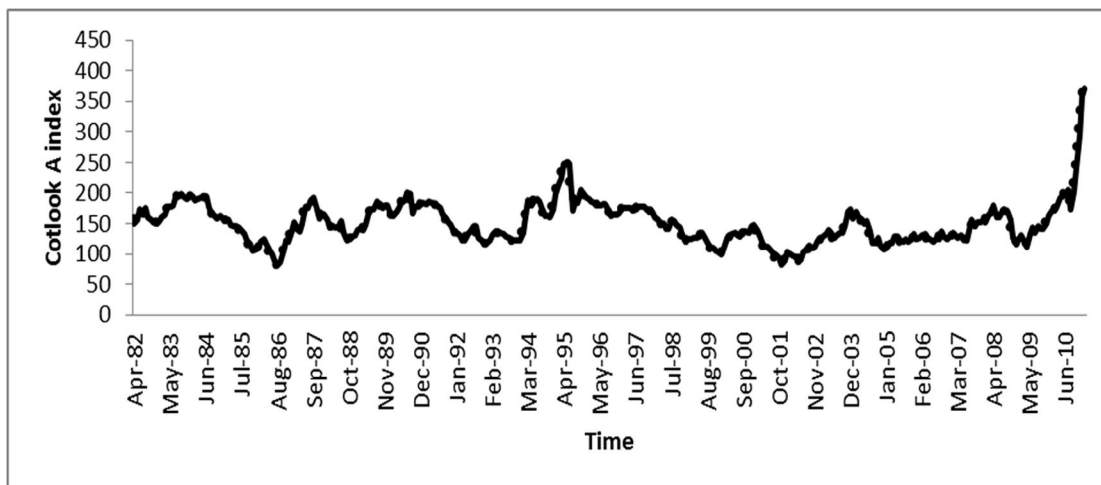


Fig 4: Fitted AR (2) – GARCH (1, 1) model along with its data point

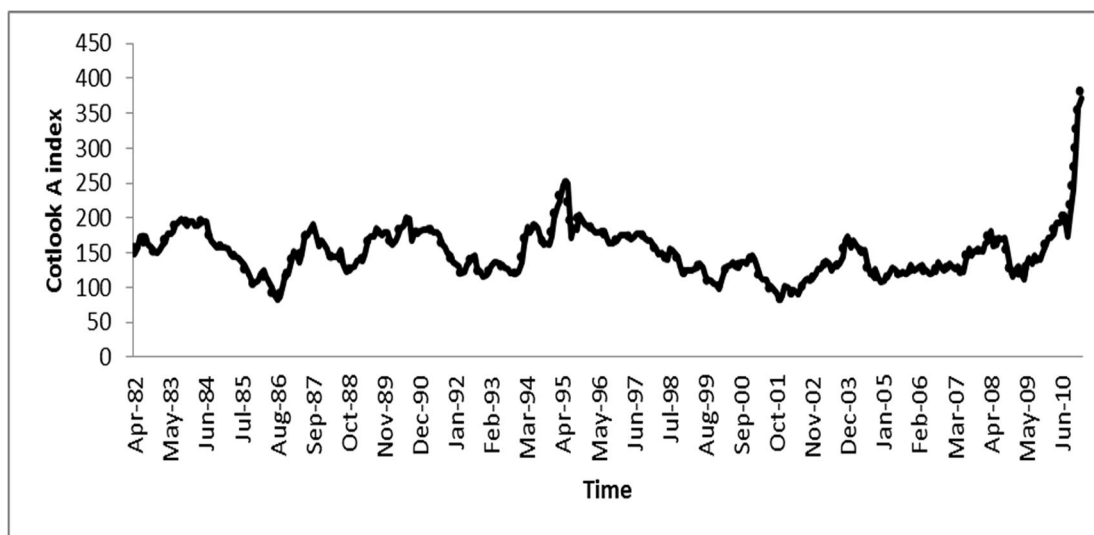


Fig 5: Fitted AR (2) – EGARCH (1, 1) model along with its data point

Table 1: Summary statistics of data

Statistic	Domestic edible oils price index	International edible oils price index	Cotlook A index
Mean	74.47	112.95	157.24
Median	72.57	100.11	149.68
Maximum	141.60	256.22	506.34
Minimum	25.22	64.13	81.93
Standard deviation	30.31	41.369	51.66
Skewness	0.16	1.64	3.030
Kurtosis	2.12	5.158	17.84

Table 2: Stationarity test

Series		ADF Test	P value	PP Test	P value
Domestic edible oils price index	Level	0.2	.97	0.0016	.95
	Differenced	12.69	<.0001	13.25	<.0001
International edible oils price index	Level	1.48	.58	1.25	.65
	Differenced	12.09	<.0001	12.16	<.0001
Cotlook A index	Level	.99	.29	0.84	.35
	Differenced	10.20	<.0001	9.05	<.0001

Table 3: Parameter estimates of ARIMA model

Series	Model	Parametr	Estimate	P value
Domestic edible oils price index	ARIMA(1,1,0)	AR(1)	0.38 (0.05)	<0.0001
International edible oils price index	ARIMA(1,1,0)	AR(1)	0.42 (0.05)	<0.0001
Cotlook A index	ARIMA(1,1,0)	AR(1)	0.56 (0.05)	<0.0001

Table 4: ARCH - LM test for three index series

Lags	Q values			P value
	Domestic price	International price	Cotlook A	
1	336.32	309.32	179.55	<.0001
2	656.57	550.00	253.21	<.0001
3	962.50	724.16	290.18	<.0001
4	1256.64	843.79	303.43	<.0001
5	1538.86	921.25	305.74	<.0001
6	1808.99	970.81	306.06	<.0001

Table 5: Parameter estimates of GARCH model

Series	Model	a	b	AIC value
Domestic edible oils price index	AR(2)- GARCH(1,1)	0.09 (0.03)	0.88 (0.03)	1191.90
International edible oils price index	AR(2)- GARCH(1,1)	0.40 (0.07)	0.54 (0.06)	2091.03
Cotlook A index	AR(2)- GARCH(1,1)	0.20 (0.09)	0.45 (0.05)	2288.88

Table 6: Forecast of the domestic edible oil index series

Month	Actual value	Forecast	
		ARIMA(1,1,0)	AR(2)-GARCH(1,1)
Apr-11	129.70	128.75(1.37)	128.90(1.57)
May-11	132.10	128.92(2.33)	129.27(1.57)
Jun-11	133.40	129.17(3.13)	129.74(1.57)
Jul-11	133.70	129.45(3.80)	130.26(1.56)
Aug-11	135.60	129.74(4.38)	130.79(1.56)

Month	Actual value	Forecast	
		ARIMA(1,1,0)	AR(2)-GARCH(1,1)
Sep-11	136.30	130.04(4.90)	131.33(1.56)
Oct-11	135.40	130.33(5.37)	131.88(1.56)
Nov-11	135.30	130.63(5.80)	132.43(1.56)
Dec-11	137.00	130.92(6.21)	132.98(1.56)
Jan-12	139.20	131.22(6.59)	133.53(1.56)
Feb-12	139.30	131.52(6.95)	134.09(1.56)
Mar-12	141.60	131.81(7.23)	134.65(1.56)

Table 7: Forecast evaluation of the domestic edible oil index series

MODEL	RMSE	RMAPE (%)
ARIMA(1,1,0)	1.71	4.10
AR(2)-GARCH(1,1)	1.25	2.96

Table 8: Forecast of the international edible oil index series

Month	Actual value	Forecast	
		ARIMA(1,1,0)	AR(2)-GARCH(1,1)
Apr-11	227.73	224.058(6.07)	227.72(10.78)
May-11	228.26	221.59(10.53)	226.39(10.64)
Jun-11	225.25	220.75(14.30)	225.08(10.51)
Jul-11	222.90	220.60(17.51)	223.77(10.38)
Aug-11	221.38	220.73(20.31)	222.49(10.26)
Sep-11	216.79	220.98(22.81)	221.21(10.14)
Oct-11	203.13	221.28(25.07)	219.95(10.03)
Nov-11	204.64	221.60(27.15)	218.71(9.92)
Dec-11	199.87	221.92(29.08)	217.48(9.82)
Jan-12	208.08	222.26(30.89)	216.26(9.73)
Feb-12	215.89	222.59(32.06)	215.05(9.63)
Mar-12	226.80	222.92(34.23)	223.84(9.55)

Table 9: Forecast evaluation of the international edible oil index series

Model	RMSE	RMAPE (%)
ARIMA(1,1,0)	3.19	3.90
AR(2)-GARCH(1,1)	2.48	2.78

Table 10: Forecast of the Cotlook A index series

Month	Actual value	Forecast		
		ARIMA(1,1,0)	Forecast AR(2)-GARCH(1,1)	Forecast AR(2)-EGARCH(1,1)
Feb-11	469.98	408.34(8.30)	389.59(26.46)	391.77(22.11)
Mar-11	506.34	416.47(15.56)	371.55(25.74)	376.72(18.12)
Apr-11	477.56	421.40(22.35)	348.54(25.05)	356.74(15.36)
May-11	364.91	424.53(28.55)	324.69(24.39)	335.61(13.39)
Jun-11	317.75	426.66(34.17)	301.98(23.75)	315.13(11.94)
Jul-11	268.96	428.23(39.29)	281.25(23.13)	296.14(10.87)
Aug-11	251.55	429.49(43.97)	262.76(22.54)	278.91(10.04)
Sep-11	257.63	430.57(48.29)	246.50(21.97)	263.48(9.41)
Oct-11	243.85	431.55(52.30)	232.32(21.42)	249.78(8.91)
Nov-11	230.78	432.48(56.05)	220.01(20.90)	237.66(8.52)
Dec-11	210.43	433.37(59.58)	209.35(20.39)	226.96(8.21)
Jan-12	222.91	434.25(54.45)	200.15(19.91)	217.55(7.96)
Feb-12	222.12	435.12(57.13)	192.21(19.44)	209.26(7.76)
Mar-12	219.36	435.99(59.68)	185.37(19.01)	201.97(7.59)

Table 11: 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
AR(2)-EGARCH(1,1)	14.41	3.99

R code for analysing a time series data using ARIMA and AR-GARCH model.

```

library("tseries")
library("forecast")
library("fgarch")
setwd("C:/Users/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

- Bollerslev, T. (1986), Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31: 307-327
- Chatfield, C. and H. Xing (2019), *The Analysis of Time-Series: An Introduction*. Chapman and Hall, Washington, D.C.
- Dickey, D. A. and W. A. Fuller (1979), Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74: 427–431.
- Engle, R. F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of U. K. inflation. *Econometrica*, 50: 987-1008.
- Lama, A., G. K. Jha, R. K. Paul and B. Gurung (2015), Modelling and forecasting of price volatility: An application of GARCH and EGARCH models. *Agricultural Economics Research Review*, 28: 73-82.
- Makridakis, S., S. C. Wheelwright and R. J. Hyndman (1998), *Forecasting, methods and applications*, 3rd edition, John Wiley, New York.
- Nelson, D. (1991), Conditional Heteroskedasticity in asset returns: A new approach. *Econometrica*, 59: 347-370.
- Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, John Wiley, New York.
- Phillips, P. C. B. and P. Perron (1988), Testing for a unit root in time series regression. *Biometrika*, 75: 335–346.
- Taylor, S. J. (1986), *Modeling financial time series*. Wiley, New York.

Chapter 15

ARTIFICIAL NEURAL NETWORK FOR TIME SERIES MODELLING

Mrinmoy Ray, K. N. Singh, Kanchan Sinha and Shivaswamy G. P.

INTRODUCTION

An artificial neural network (ANN), otherwise called neural network (NN), is one of the popular machine learning techniques that is inspired by the structure as well as functional aspects of biological neural networks the human brain especially. A neural network made out of various interconnected simple processing elements called neurons or nodes. Each node receives an input signal which is the aggregate “information” from other nodes or external stimuli, processes it locally through an activation or transfer function and produces a transformed output signal to other nodes or external outputs. This information processing characteristic makes ANNs an effective computational device and able to learn from examples and is then to generalize to examples never seen before.

A Time series (TS) is an ordered sequence of observations of a variable at equally spaced time intervals (monthly price data of a commodity, yearly crop yield and daily temperature data etc.). The motivation behind the use of time series model is to predict future values based on previously observed values. The most popular time series model is Auto Regressive Integrated Moving Average (ARIMA) (Box *et al.*, 2009 and Makridakis *et al.*, 1989). However, the main drawback of ARIMA model is that it can only deal with linear time series data (Ray *et al.*, 2016). Even though there are divergent statistical models to deal with non-linear time series data such as bilinear, Autoregressive Conditional Heteroscedasticity (ARCH) model, Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model and threshold autoregressive (TAR) model etc. But these models have some specific assumption which is not always plausible to identify while dealing with real world time series data (Jha and Sinha, 2014). Therefore, in the gamut of time series modeling Artificial Neural Network (ANN) (Zhang *et al.*, 1998; Remus and O’Connor, 2001 and Mukerjee *et al.*, 2016) has notched up much popularity in modeling non-linear dynamics and subsequently rendering the non-linear forecasting. The main feature of this approach is that it doesn’t require prior assumption of the time series data under consideration, rather it is to a great extent depends upon pattern of the data popularly known as data-driven approach.

ANN Architecture

In general, an ANN can be partitioned into three sections, named layers, which are discussed as

Input layers

These layers are responsible for receiving information (data), signals, features, or measurements from the external environment. These inputs (samples or patterns) are usually normalized within the limit values produced by activation functions. This normalization results in better numerical precision for the mathematical operations performed by the network.

Hidden, intermediate, or invisible layers

These layers are composed of neurons which are responsible for extracting patterns associated with the process or system being analyzed. These layers perform most of the internal processing from a network.

Output layers

These layers are also composed of neurons, and thus are responsible for producing and presenting the final network outputs, which result from the processing performed by the neurons in the previous layers.

ANN approach to time series forecasting

In the domain of time series analysis, the inputs are typically the past observations series and the output is the future value. The ANN performs the following nonlinear function mapping between the input and output

$$y_t = f(y_{t-1} + y_{t-2}, \dots, y_{t-p}, w) + \epsilon_t$$

where, w is a vector of all parameters and f is a function of network structure and connection weights. Therefore, the neural network resembles a nonlinear autoregressive model.

Single hidden layer multilayer feed forward network is the most popular for time series modeling and forecasting. This model is characterized by a network of three layers of simple processing units. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer (Fig 1).

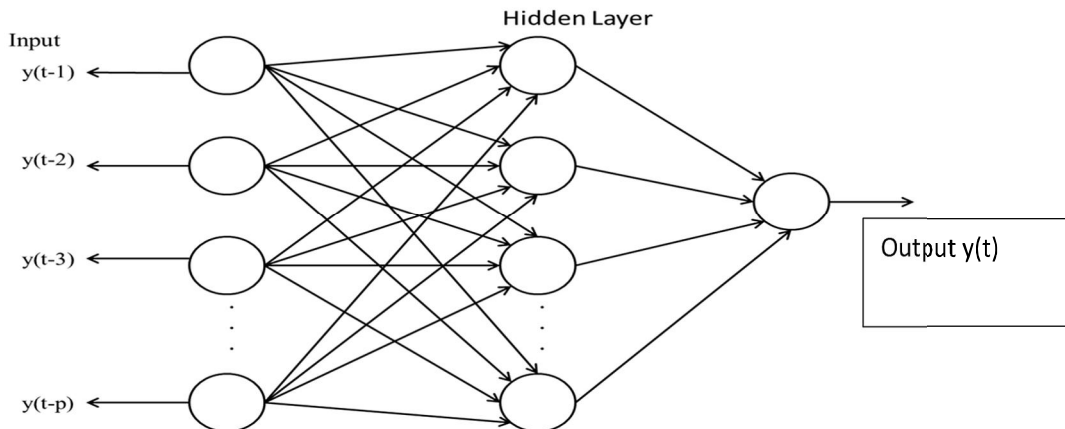


Fig 1: Architecture of ANN for time series forecasting

The relationship between the output (y_t) and the inputs ($y_{t-1}, y_{t-2}, \dots, y_{t-p}$) can be mathematically represented as follows:

$$y_t = f \left(\sum_{j=0}^q \omega_j g \left(\sum_{i=0}^p \omega_{ij} y_{t-i} \right) \right)$$

where, $\omega_j (j = 0, 1, 2, \dots, q)$ and $\omega_{ij} (i = 0, 1, 2, \dots, p, j = 0, 1, 2, \dots, q)$ model parameters often called the connection weights, p number of input nodes q number of hidden nodes, g and f activation function at hidden and output layer respectively. Activation function defines the relationship between inputs and outputs of a network in terms of degree of the non-linearity.

For time series forecasting sigmoid activation function is employed in hidden layer and identity activation function is employed in the output layer which are given in Table 1.

Table 1: Time series forecasting sigmoid activation function

Activation function	Equation
Identity	x
Sigmoid	$\frac{1}{1 + e^{-x}}$

The selection of appropriate number of hidden nodes as well as optimum number of lagged observation p for input vector is important in ANN modeling for determination of the autocorrelation structure present in a time series. Though there are no established theories available for the selection of p and q , hence experiments are often conducted for the determination of the optimal values of p and q . The connection weights of ANNs are determined by Gradient decent back propagation algorithm which is described here.

The objective of training is to minimize the error function that measures the misfit between predicted value and actual value. The error function which is widely used is mean squared error which can be written as:

$$E = \frac{1}{N} \sum_{n=1}^N (e_i)^2 = \frac{1}{N} \sum_{n=1}^N \left\{ y_t - f \left(\sum_{j=0}^q \omega_j g \left(\sum_{i=0}^p \omega_{ij} y_{t-i} \right) \right) \right\}^2$$

where N total number of error terms. The parameters of the neural network are ω_j and ω_{ij} estimated by iteration. Initial connection weights are taken randomly from uniform distribution. In each iteration the connection weights changed by an amount $\Delta \omega_j$

$$\Delta\omega_j(t) = -\eta \frac{\partial E}{\partial \omega_j} + \delta \Delta\omega_j(t-1)$$

where, η , learning rate and $\frac{\partial E}{\partial \omega_j}$, partial derivative of the function E with respect to the weight ω_j ; δ , momentum rate. The $\frac{\partial E}{\partial \omega_j}$ can be represented as follows-

$$\frac{\partial E}{\partial w_j} = -e_j(n) \times f'(x) \times y_j(n)$$

where $e_j(n)$, residual at n^{th} iteration

$f'(x)$ = derivative of the activation function in the output layer. As in time series forecasting the activation function in the output layer is identity function hence $f'(x) = 1$. $y_j(n)$ is the desired output. Now connection weights in from input to hidden nodes changed by an amount $\Delta\omega_{ij}$

$$\Delta\omega_{ij}(t) = -\eta \frac{\partial E}{\partial \omega_{ij}} + \delta \Delta\omega_{ij}(t-1)$$

where

$$\frac{\partial E}{\partial w_{ij}} = g'(x) \times \sum_{j=0}^q e_j(n) * w_j(n)$$

where $g'(x)$ is the activation function in the hidden layer. For sigmoid activation function

$$g'(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}$$

Learning rate is user defined parameter known as tuning parameter of neural network which determine how slow or fast the optimal weight is obtained. The learning rate must be set small enough to avoid divergence. The momentum term prevents the learning process from setting in a local minimum. Though there are no established theories available for the selection of learning rate and momentum, hence experiments are often conducted for the determination of the learning rate and momentum.

Step by step modeling procedure

Division of the data

Data divided into training and test sets. The training sample is used for ANN for model development and the test sample is utilized to evaluate the forecasting performance. Sometimes a third one called the validation sample is also utilized to avoid the overfitting problem or to determine the stopping point of the training process. It is common to use one test set for both validation and testing purposes particularly for small data sets.

The literature suggests little guidance in selecting the training and testing sets. Most commonly used rule are 90% vs. 10%, 80% vs. 20% or 70% vs. 30%, etc.

Data normalization

Nonlinear activation functions such as the sigmoid function typically have the squashing role in restricting the possible output from a node to, typically, (0, 1). Hence, data normalization is done prior to training process begins.

Normalization procedure

Linear transformation to [0,1]: $X_n = (X_0 - X_{\min}) / (X_{\max} - X_{\min})$

Statistical normalization: $X_n = (X_0 - \text{mean}(X)) / \text{var}(X)$

Simple normalization: $X_n = X_0 / X_{\max}$

Selection of appropriate number of hidden nodes as well as optimum number of lagged

There are no established theories available for the selection of p and q , hence experiments are often conducted for the determination of the optimal values of p and q .

Estimation of connection weights

Estimation of connection weights are determined by learning algorithm. For time series forecasting most commonly used learning approach is gradient decent back propagation algorithm.

Evaluating forecasting performance

Forecasting performance can be computed by several approaches. Some of the approaches are given here

$$MAPE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| / y_t \times 100$$

$$MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2$$

where, n , total number of forecast values; y_t , actual value at period t and, corresponding forecast value. The model with less MAPE/MSE is preferred for forecasting purposes.

ILLUSTRATION

Monthly data on wholesale price of rice (January, 2010 to October, 2018) of all-India level data have been utilized for Illustration purpose. The plot of considered time series data is given here (Fig 2).

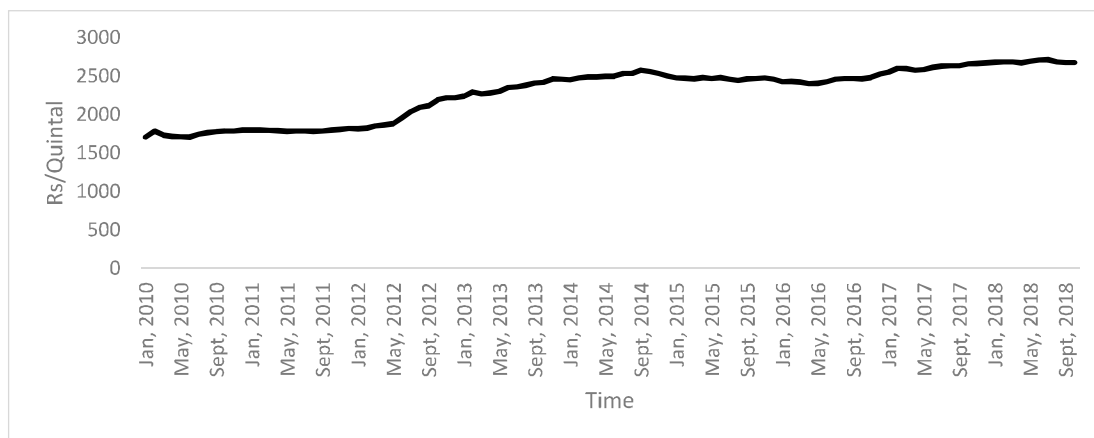


Fig 2: Wholesale price of rice (January, 2010 to October, 2018) of all-India level

Data from January, 2010 to November, 2017 (90%) were used for model construction. The summary of the fitted neural network model is given in Table 2. The ANN was fitted employing “forecast” package of R software. The R code is given in Annexure.

Table 2: Summary of fitted ANN

Parameters	ANN
Number of input (lag)	1
Number of hidden unit	2
Activation function in hidden unit	Sigmoid
Activation function in output unit	linear
Learning algorithm	Gradient decent back propagation

From December, 2018 to October, 2018 (10%) were used to check the forecasting performance. The Actual versus ANN model values plot is given in Fig 3.

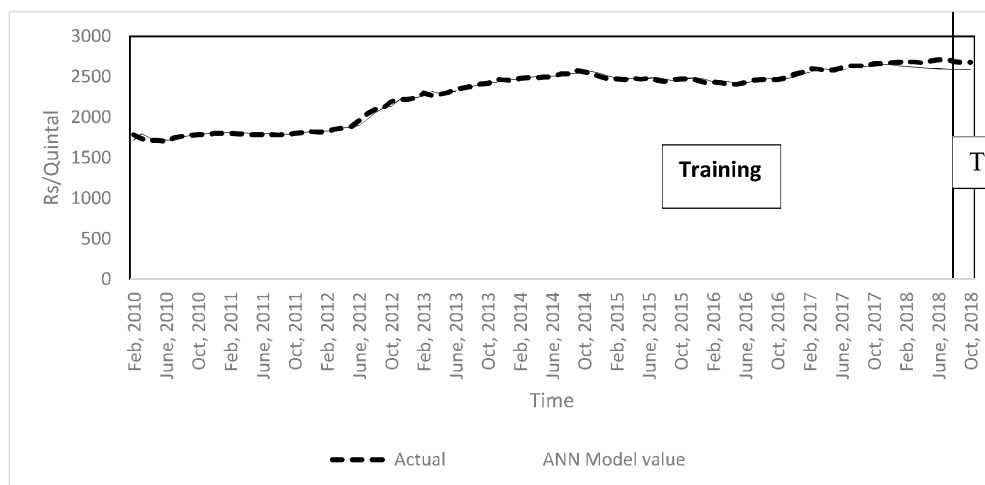


Fig 3: The actual versus ANN model values plot

Fig 3 reveals that ANN model is good fit for the considered time series data. The MAPE and MSE have been computed for training as well as testing set Table 3.

Table 3: MAPE and MSE values of fitted ANN model

Model	MAPE		MSE	
	Training	Testing	Training	Testing
ANN	0.81	2.83	567.20	6584.01

Based on the values of MAPE and MSE the performance of the ANN model can be compared with other time series model like ARIMA, GARCH etc.

REFERENCES

- Box, G. E. P., G. M. Jenkins and G. C. Reinsel (2009), Time Series Analysis: *Forecasting and Control* (3rd ed.), San Francisco: *Holden-Day*.
- Jha, G. K. and K. Sinha (2014), Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*, 24 (3): 563-571.
- Makridakis, S., S. C. Wheelwright and R. J. Hyndman (1998), *Forecasting: Methods and Applications*, 3rd Edition, Chichester: Wiley.
- Mukherjee, A., S. Rakshit, A. Nag, M. Ray, H. L. Kharbikar, S. Kumari, S. Sarkar, S. Paul, S. Roy, A. Maity, V. S. Meena and R. R. Burman (2016), Climate Change Risk Perception, Adaptation and Mitigation Strategy: An Extension Outlook in Mountain Himalaya. In: Jaideep Kumar Bisht, Vijay Singh Meena, Pankaj Kumar Mishra and Arunava Pattanayak Edition. *Conservation Agriculture* (pp. 257-292). Singapore. Springer Singapore.
- Ray, M., A. Rai, V. Ramasubramanian and K. N. Singh (2016), ARIMA-WNN hybrid model for forecasting wheat yield time series data. *Journal of the Indian Society of Agricultural Statistics*, 70(1): 63-70.
- Remus, W. and M. O'Connor (2001), *Neural Networks for Time-Series Forecasting*, New york, Springer.
- Zhang, G., B. E. Patuwo and M. Y. Hu (1998), Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14: 35-62.

ANNEXURE I

```
z=read.csv(file.choose() , header=TRUE) # Uploading data in R
head(z)
zz=data.frame(z)
n = nrow(zz)
size = round(0.90*n) # Dividing 90% data as training data as rest 10% as testing data
train=zz[1:size,]
test=zz[(size+1):n,]
install.packages("forecast") # installation of forecast package
library(forecast)
fit<-nnetar(train$x)
fitv=fitted(fit) # fitted values
ff<-forecast(fit, h=nrow(test))
kk2=ff$mean# Forecasted values
```


Chapter 16

Hybrid Time-Series Models

Ranjit Kumar Paul

INTRODUCTION

Time-series forecasting is an important statistical analysis technique used as a basis for manual and automatic planning in many application domains (Gooijer and Hyndman, 2006). Forecasting plays a crucial role in business, industry, government and institutional planning. Much effort has been devoted over last few decades to develop and improve several time-series forecasting models. There are several linear time-series models available in literature. One of the important and widely used technique for analysis of univariate time-series data is Box Jenkins' Autoregressive integrated moving average (ARIMA) methodology (Box *et al.*, 2007). In addition to ARIMA, various exponential models can also be used to forecast a linear time-series process. But one of the major limitations of these models is the pre-assumed linear form of the models. This assumption of linearity limits the application of ARIMA model to real time-series data. Some of the applications of this model can be found in Paul and Das (2010, 2013), Paul *et al.* (2013, 2014), Paul (2015).

Linear models are not able to describe any changes in the conditional variances present in the real data. To tackle this situation, Engle (1982) defined the Autoregressive conditional heteroscedastic (ARCH) models in which significant presence of autocorrelation in the squared residual series is considered. But the ARCH models give satisfactory forecast only with large number of parameters which has necessitated the emergence of more parsimonious version that is the Generalized ARCH (GARCH) models (Bollerslev, 1986). In GARCH models the unconditional autocorrelation function has slow decay rate.

Unlike the traditional model-based methods, artificial neural network (ANN) is a data-driven, self-adaptive, non-linear, non-parametric method of forecasting. Many nonlinear processes that have unknown functional relationship can be modeled by ANN. There are many empirical evidences that non-linear models perform well for long term forecasting whereas the linear models are suitable for short range forecasting. So there is a need of combining the linear and non-linear models in order to get more accurate forecast. Real world time-series data are hardly pure linear or non-linear in nature. They are often mixed up with both linear and non-linear components in the structure. This phenomena makes it necessary to combine linear and non-linear models in order to capture the existing pattern in the dataset more accurately. It is observed that no single forecasting method will be the best choice in every situation. Most of the real-world problems are complex in nature and any single model is not able to capture several

patterns uniformly. Therefore, combining of different models is important to increase the chance of capturing different patterns and improve the forecasting performance. Paul and Sinha (2016) have compared the performance of ARIMAX and NARX model for forecasting crop yield. Mitra and Paul (2017) have illustrated the performance of hybrid time series models as compared to individual models.

TIME-SERIES FORECASTING MODELS

There are several approaches of time-series modeling. Some of the traditional linear time-series models are moving average, exponential smoothing and ARIMA. To overcome the deficiencies of linear time-series models and to capture certain non-linear pattern in the time-series data several non-linear time-series models are available in the literatures. The most commonly used non-linear time-series models are ARCH and GARCH models.

The ARIMA Model

In an ARIMA model, it is assumed that the future value of a variable is a linear function of past values of the variable itself and random errors also. It is a linear univariate time-series model which expresses a time-series process, say, $\{y_t\}$, $t = 1, 2, \dots, n$ in terms of three sets of parameters as

$$\varphi(L)y_t = (1 - L)^{-d}\theta(L)e_t$$

where y_t and e_t are the actual value and random error at the time t , respectively; $\varphi(L)$ and $\theta(L)$ are the polynomial of lag operator L of order p and q respectively with root outside the unit circle; random errors, e_t are assumed to be independent and identically distributed with mean zero and variance σ^2 . In general an ARIMA model is denoted as ARIMA (p, d, q) where p, d and q represents the order of autoregression, integration (differencing) and moving average respectively. ARIMA is a general class of models which can represent both stationary and non-stationary processes by allowing d value as 0 and 1 or 2 respectively.

The Box-Jenkins methodology of ARIMA modeling includes three iterative steps, viz. model identification, parameter estimation and diagnostic checking. The first step of the process is to check the stationarity of the series since the estimation procedure is only available for stationary process. A series is called stationary if up to 2nd order central moments, that means mean, variance and autocorrelation structure are constant over time. If the series is non-stationary then differencing is required to make the series stationary, while logarithmic transformation is required to stabilize the variance. After suitable differencing and transformation a tentative models is selected based on the autocorrelation function (ACF) and partial autocorrelation function (PACF). The parameters of the tentatively selected models are estimated such that the overall measure of error is minimized or the likelihood function is maximized. In last step

of model building diagnostic checking for model adequacy is done for all candidate models by plotting ACF of residuals and via portmanteau test like Box-Pierce and Ljung-Box test. Finally the most suitable model is selected on the basis of the minimum Akaike Information Criterion (AIC) or Schwarz-Bayesian Criterion (SBC) and lowest root mean square error (RMSE).

The ARCH Model

variance which is present in many real time-series data. To handle such a situation, Engle (1982) has introduced the ARCH models in which significant presence of autocorrelation of squared residuals is considered. The ARCH (q) model for the series $\{\epsilon_t\}$, $t = 1, 2, \dots, n$ is defined by specifying the conditional distribution of $\{\epsilon_t\}$ given the information available up to $t - 1$. The process $\{\epsilon_t\}$ is ARCH (q) if the conditional distribution of $\{\epsilon_t\}$ given the available ψ_{t-1} information is

$$\epsilon_t | \psi_{t-1} \sim N(0, h_t) \text{ and } \epsilon_t = \sqrt{h_t} \epsilon_t, \dots \quad (1)$$

where $\{\epsilon_t\}$ is a white noise process that means $\{\epsilon_t\}$ is a sequence of independent and identically distributed (*i.i.d*) random variables with mean zero and variance 1, i.e., $\epsilon_t \sim iid(0,1)$ and the conditional variance h_t is defined as

$$h_t = a_0 + \sum_{i=1}^q a_i \epsilon_{t-i}^2, \quad a_0 > 0, a_i \geq 0 \quad \forall \text{ and } \sum_{i=1}^q a_i < 1 \dots \quad (2)$$

ARCH model has some drawbacks. Firstly, when the order of ARCH model is very large, estimation of a large number of parameters is required which is really a cumbersome process. Secondly, The ARCH model has the property that the unconditional autocorrelation function of squared residuals; if it exists; decay very rapidly compared to what is typically observed unless maximum lag q is large.

The GARCH Model

To overcome these difficulties Bollerslev (1986) have proposed GARCH model in which conditional variance is also a linear function of its own lags.

A GARCH (p, q) process has the following form

$$h_t = a_0 + \sum_{i=1}^q a_i \epsilon_{t-i}^2 + \sum_{j=1}^p a_j h_{t-j} \dots \quad (3)$$

$$= a_0 + a(L) \epsilon_t^2 + b(L) h_t$$

where $a(L)$ and $b(L)$ are the finite polynomial in the lag operator L of order p and q respectively. The conditional variance defined by (3) has the property that the autocorrelation function of the squared residuals, ϵ_t^2 , if exists, decay slowly. The most popular model is GARCH (1,1). A sufficient condition for the conditional variance to be positive is:

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

The first step of a GARCH process is to check for conditional heteroscedasticity of the squared residual series ϵ_t^2 which is known as the ARCH test. There are two tests available in the literature for ARCH test. The first one is Ljung-Box test where the null hypothesis is that the first m lags of autocorrelation functions of the ϵ_t^2 series are zero. The second test for conditional heteroscedasticity is the Lagrange multiplier (LM) test. After detecting the ARCH effect the parameters of the model is estimated using the Gaussian maximum likelihood estimation (GMLE) method. In the last step the most suitable model is selected on the basis of minimum AIC or SBC value and lowest RMSE Applications of this model in agriculture can be found in Paul *et al.* (2009, 2014) and Ghosh *et al.* (2011).

The ANN approach to time-series modeling

The working principle of ANN is based on the human brain by making the right connections. Like the structure of neuron ANN is decomposed into several nodes which are categorized as input layer that accepts external information, one or more hidden layer that perform simple operations on the data and an output that results the required information. All the nodes are connected through an acyclic arc. The Fig 1 shows the architecture of a simple ANN. There are two Artificial Neural Network topologies – feed-forward and feedback. In feed-forward topology the flow of information is unidirectional and there is no feedback path where as in feedback topology feedback paths are there. The two topologies are demonstrated in Fig 2.

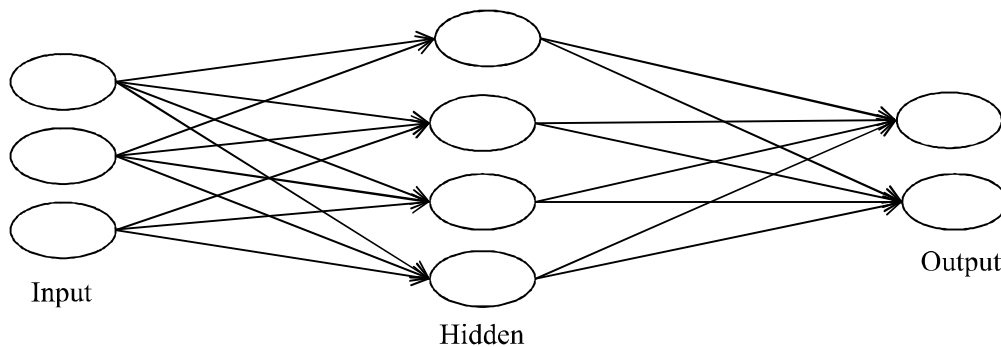


Fig 1: Artificial neural network with one hidden node

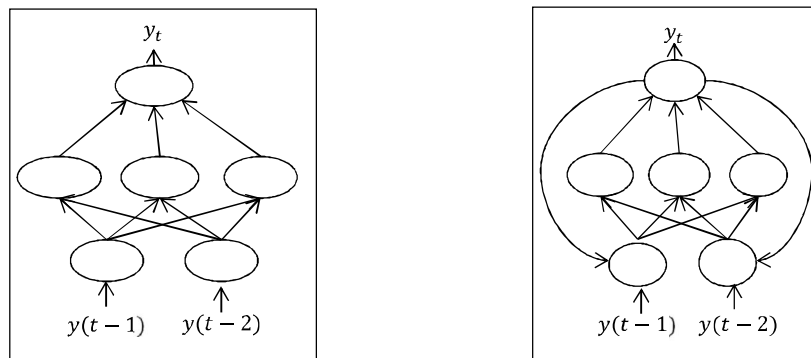


Fig 2: Artificial Neural Network topologies – feed-forward (Left) and feedback (Right)

The application of neural network structure for solving a particular time-series problem involves determination of number of layers and total number of nodes in the structure which is done on experimentation basis. It is established that single hidden layer with sufficient number of nodes at the hidden layer and adequate data for initialization can well approximate any nonlinear function. In neural network determination of number of input nodes which are lagged observations of same variable plays an important role in model building. Determination of output nodes is relatively easy. It is suggested that model with small number of nodes at hidden layer results in improved out-of-sample forecasting performance.

THE HYBRID METHODOLOGY

Zhang (2003) proposed a hybrid approach that decomposes a time-series into its linear and non-linear component. The hybrid model considers the time-series y_t as a combination of both linear and non-linear components. That is,

$$y_t = L_t + N_t$$

where L_t and N_t represent the linear and non-linear component present in the given data, respectively. These two components are to be estimated from the data. This hybrid method of combining forecasting has following steps:

1. First, a linear time-series model, say, ARIMA is fitted to the data.
2. At the next step residuals are obtained from the fitted linear model. The residuals will contain only the non-linear components. Let e_t denotes the residual at the time t from the linear model, then

$$e_t = y_t - \hat{L}_t$$

where, \hat{L}_t is the forecast value for the time t from the estimated linear model.

1. Diagnosis of residuals is done to check if there is still linear correlation structures left in the residuals. The residuals are tested for possible presence of non-linearity by using BDS test.
2. Once the residuals confirm the non-linearity, then the residuals are modeled using a non-linear model, say, ARCH or ANN. And also obtain the forecast values, \hat{N}_t for the residual series.
3. Finally the forecasted linear and non-linear components are combined to obtain the aggregated forecast values as

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

The hybrid approaches can be graphically represented by Fig 3 and Fig 4.

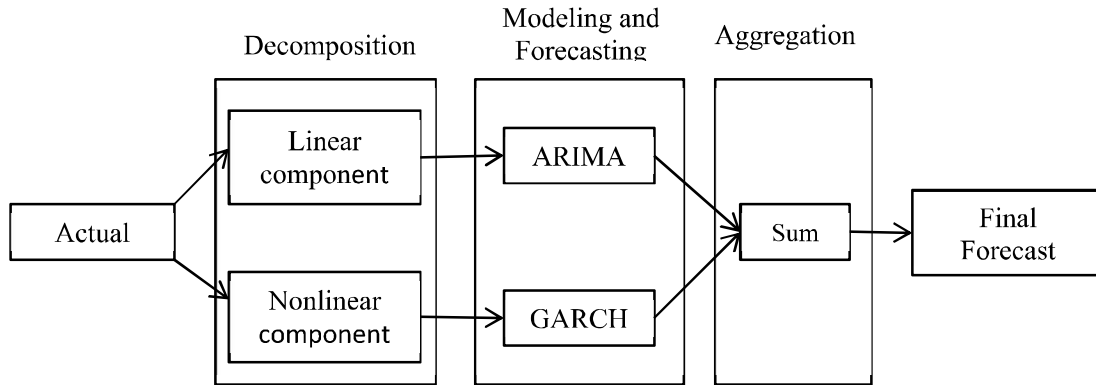


Fig 3: Schematic representation of ARIMA-GARCH hybrid methodology

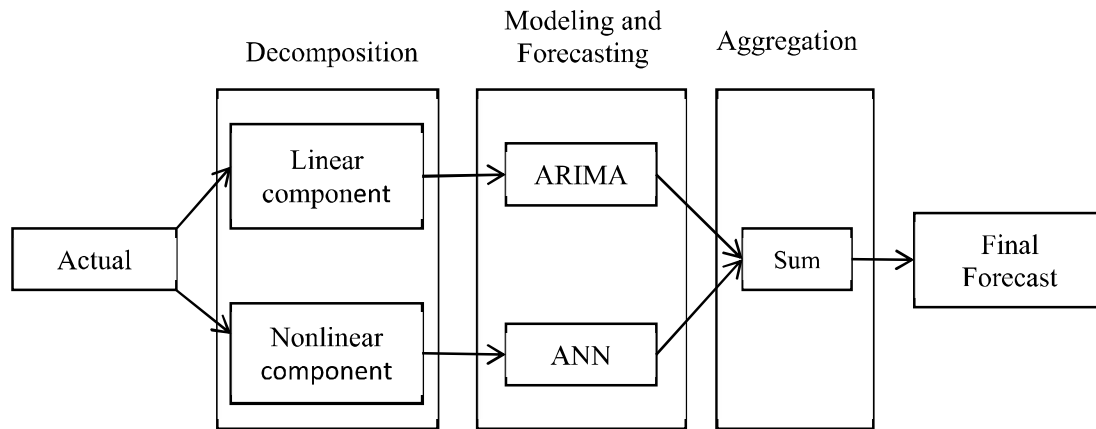


Fig 4 : Schematic representation of ARIMA-ANN hybrid methodology

ILLUSTRATION

For the present study potato price data series belong to Agra market in India for the period January, 2005 to May, 2017, collected from National Horticulture Research and Development Foundation (NHRDF) (the website: <http://nhrdf.org/en-us/>) are used. The data series is divided into two parts: training set and testing set (holdout set). The training data set (January, 2005 to May, 2016) is used for parameter estimation and the last 12 observations i.e. from June, 2016 to May, 2017 considered as testing set is used for validation purpose and also for obtaining out-of-sample forecast (Mitra and Paul, 2017).

Descriptive statistics and seasonal indices

The descriptive statistics of potato price for Agra market are reported in Table 1. A perusal of the Table 1 indicates that average potato price in Agra market is 652. Since the CV is more than 50% it can be concluded that the variability in price of Agra market is slightly in higher sight. The series under consideration is positively skewed

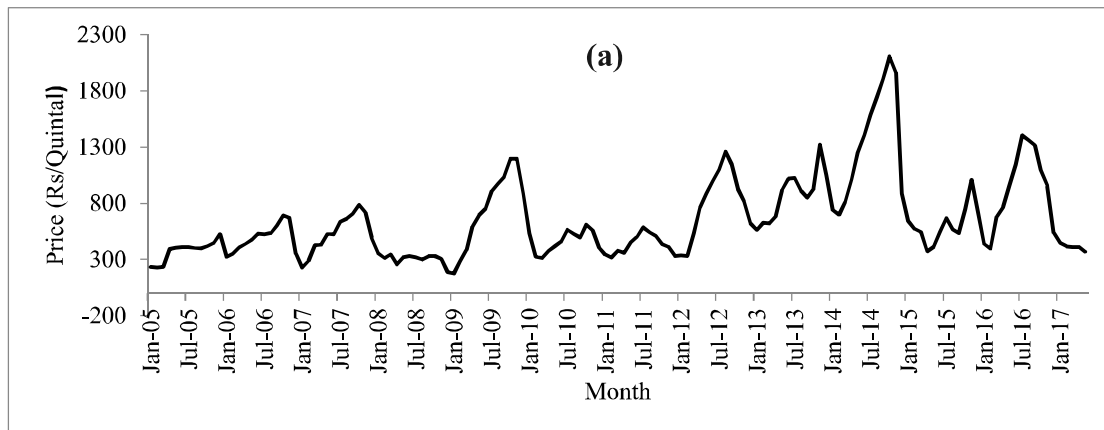
and leptokurtic. Original data is seasonally adjusted to eliminate the influence of seasonality in price. Table 2 shows the seasonal index values. Relatively higher values of seasonal indices are found from June to November. Being a rabi crop, the planting time of potato is 15th September - 15th October and it gets ready for harvesting at the end of November. Fresh arrival starts to reach the market by the end of November onwards.

Table 1: Descriptive statistics of potato prices in Agra market

Statistics	Agra
Observations	149
Mean (Rs./quintal)	652.33
Minimum	175
Median	536.00
Maximum	2107.00
SD	367.86
CV	56.39
Skewness	1.52
Kurtosis	5.51

Note: SD: standard deviation; CV: coefficient of variation

The first and foremost step in time-series analysis is to plot the data and visualize the presence of several time-series components. Figure 5a and figure 5b show the time-series plot of average monthly price of potato for original series and monthly potato price for seasonally adjusted series from January, 2005 to May, 2017 in Agra market. A perusal of this figure indicates that the price attains its higher values during the period June, 2014 to December, every year. Though the highest price has been observed in October, 2014. The time-plot of original price data also indicates that some seasonal pattern is present in the dataset and it is required some kind of seasonal adjustment.



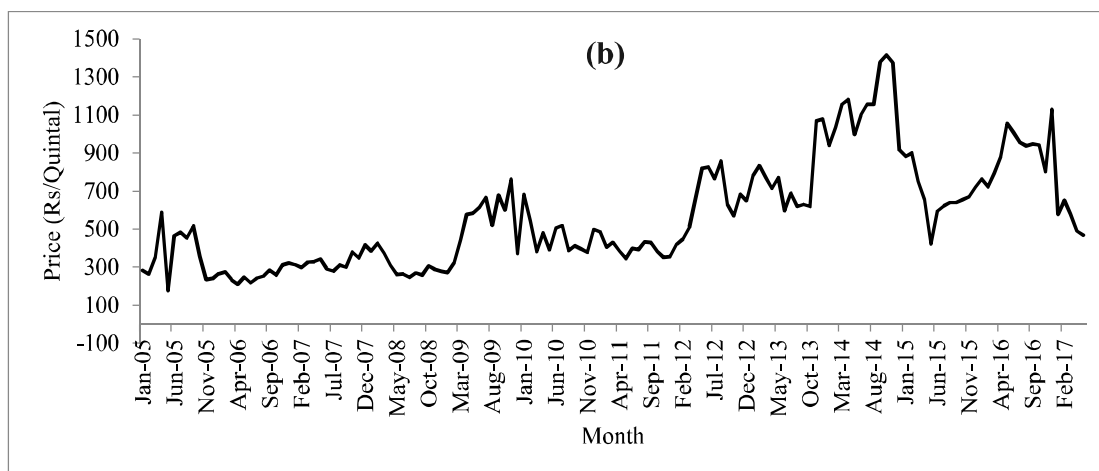


Fig 5: Monthly price of Potato from January 2005 to May 2017 for (a) Original Series (b) Seasonally Adjusted Series in Agra Market

Table 2: Seasonal factors for potato prices in the Agra market

Months	Seasonal factor	Months	Seasonal factor
January	0.68	July	1.21
February	0.65	August	1.20
March	0.76	September	1.22
April	0.83	October	1.29
May	0.97	November	1.32
June	1.08	December	0.79

Test for stationarity

Phillips-Perron (PP) and Augmented Dicky-Fuler (ADF) tests have been applied to see the presence of non-seasonal unit root in the seasonally adjusted series it was found that the null hypothesis of unit root test is not rejected at 5% level of significance indicating seasonally adjusted series are non-stationary in nature and the results are given in the Table 3. Non rejection of the null hypothesis of unit root for both the tests at 5% level of significance indicates that differencing is required to make the seasonally adjusted series stationary for the market. Rejection of null hypothesis of stationarity test for 1st differenced series reveals that no more differencing is required.

Table 3: ADF and PP test for stationarity

Markets	Original series				1 st differenced series			
	ADF test		PP test		ADF test		PP test	
	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
Agra	-2.72	0.07	-2.68	0.09	-10.35	<0.001	-10.35	<0.001

Fitting of ARIMA model

After confirming the stationarity of the price series after one differencing, suitable ARMA model was selected based on minimum AIC and BIC criterion and observing the significance of autocorrelation and partial autocorrelations functions. Accordingly, ARIMA(1,1,0) model is selected for seasonally adjusted price series of potato in Agra market. The parameter estimates of fitted ARIMA model are furnished in Table 4 along with their significance level (p-value).

Table 4: Parameter estimates of the ARIMA (1,1,0) of Agra market

Model	Parameters	Estimate	p-value
ARIMA (1,1,0)	C	5.58	0.47
	AR (1)	-0.26	<0.01

Testing for ARCH effects

The presence of autocorrelation in the squared residuals of best fitted ARIMA model was investigated and reported in Table 5. It was found that the squared residuals are autocorrelated at least up to 12 lags indicating possible presence of ARCH effect. To test the presence conditional heteroscedasticity, ARCH-LM test is performed and it is found that the ARCH effect is significant up to 5 lags.

Table 5: Test for ARCH effects for seasonally adjusted series

Order	Agra			
	Q-statistic	p-value	LM-statistic	p-value
1	9.43	<0.01	9.24	<0.01
2	9.58	0.01	10.61	0.01
3	9.65	0.02	11.12	0.01
4	9.67	0.05	11.34	0.02
5	9.87	0.08	11.66	0.04
6	9.89	0.13	11.79	0.07
7	9.89	0.19	11.81	0.11
8	10.06	0.26	11.90	0.16
9	10.09	0.34	12.18	0.20
10	10.50	0.40	12.88	0.23
11	10.92	0.45	15.00	0.18
12	11.26	0.51	15.00	0.24

Fitting of GARCH model

Accordingly, to capture the non-linearity and heteroscedasticity in conditional variance, GARCH model is applied for modelling and forecasting the price series. The parameter estimates of best fitted ARIMA and GARCH model are furnished in Table 6 along their significance level.

Table 6: Parameter estimates of the ARIMA (1,1,0)-GARCH (1,1) model for Agra Market

Model	Parameters	Estimate	p-value
ARIMA (1,1,0)- GARCH (1,1)	Mean equation		
	C	5.29	0.43
	AR (1)	-0.06	0.62
	Variance equation		
	C	330.19	0.19
	ARCH(1)	0.43	<0.01
	GARCH(1)	0.68	<0.01

Fitting of hybrid models

Once it is confirmed that the residuals of the fitted ARIMA model contains nonlinear part and also the significant ARCH effect is present, the hybrid models namely ARIMA-ANN and ARIMA-GARCH model as discussed in section 4 were employed to investigate the improvement in forecast accuracy as compared to the individual ARIMA and GARCH models.

Evaluation of Forecasting Performances

The prediction abilities of the ARIMA and GARCH models and the hybrid models i.e. ARIMA-ANN and ARIMA-GARCH are compared with respect to mean absolute percentage error (MAPE) and root mean squared error (RMSE) for last twelve observations (i.e. for last twelve months). The formula for computing MAPE and RMSE are given below

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \times 100$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

where y_t is the actual price at time t , \hat{y}_t is the predicted price at time t , and n is the sample size for the hold out data. In the present investigation n is 12. The values of MAPE and RMSE of different models are reported in Table 7.

Table 7: Comparison of prediction performance of different models

Comparison based on MAPE				
Series	ARIMA	GARCH	ARIMA-GARCH	ARIMA-ANN
Agra	16.16	16.54	16.00	11.31
Comparison based on RMSE				
Agra	172.71	189.47	169.44	123.90

The accuracy of a statistical model is the fundamental feature to select that particular model and to take many important decisions. Box-Jenkins's ARIMA methodology is most popular method of forecasting of a linear time-series process. In many of the practical situations, the assumptions of linearity and homoscedastic error variance which are two most crucial assumptions of ARIMA model are violated. In such cases, nonlinear time series models are called for. GARCH family of models is the most widely used nonlinear time series models in literature. The hybrid methodology which decomposes a series into its linear and nonlinear part followed by modelling each part separately before they are combined for getting final forecast is described in detail here. The above algorithm has been applied in forecasting the wholesale price of potato in Agra market. The comparison of forecast performance among the ARIMA, GARCH, ARIMA-GARCH and ARIMA-ANN hybrid models has been carried out. It is seen that the hybrid models perform better than the individual counterpart i.e. ARIMA and GARCH models with respect to minimum MAPE and RMSE value. The residuals from finally fitted hybrid model are examined and it is found that the residuals are independent and normally distributed ensuring the adequacy of model selected. It is also to be noted that if the exogenous variable is included in the model it may increase the accuracy of forecasting provided that the exogenous variable is correlated with the study variable.

REFERENCES

- Bolerslev, T. (1986), Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics*, 31: 307-327.
- Box, G. E. P., G. M. Jenkins and G. C. Reinsel (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 United Kingdom inflation, *Econometrica*, 50: 987-1007.
- Gooijer, J. G. D. and R. J. Hyndman (2006), 25 years of time-series forecasting. *International Journal of Forecasting*. 22(3), 443-73.
- Ghosh, H., R. K. Paul and Prajneshu (2011), Nonlinear time series modeling and forecasting for periodic and ARCH effects. *Journal of Statistical Theory and Practice*. 4(1): 27-44.
- Paul, R. K. and M. K. Das (2010), Statistical modelling of inland fish production in India. *Journal of the Inland Fisheries Society of India*: 42: 1-7.
- Paul, R. K., H. Ghosh and Prajneshu (2009), GARCH Nonlinear Time Series Analysis for Modelling and Forecasting of India's Volatile Spices Export Data. *Journal of the Indian Society of Agricultural Statistics*, 62 (2): 123-132

- Paul, R. K., H. Ghosh and Prajneshu (2014), Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of the Indian Society of Agricultural Statistics*, 68(1): 85-92.
- Paul, R. K. and M. K. Das (2013), Forecasting of average annual fish landing in Ganga Basin. *Fishing chimes*, 33 (3): 51-54
- Paul, R. K., S. Panwar, S. K. Sarkar, A. Kumar, K. N. Singh, S. Farooqi and V. K. Chaudhary (2013), Modelling and Forecasting of Meat Exports from India. *Agricultural Economics Research Review*, 26 (2), 249-256.
- Paul, R. K., Prajneshu and H. Ghosh (2013), Modelling and forecasting of wheat yield data based on weather variables. *Indian Journal of Agricultural Science*, 83: 180-183.
- Paul, R. K., W. Alam and A. K. Paul (2014), Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences*, 84(4): 130-134.
- Paul, R. K. (2015), ARIMAX-GARCH-WAVELET Model for forecasting volatile data. *Model Assisted Statistics and Application*, 10(3): 243-252.
- Paul, R. K. and K. Sinha (2016), Forecasting crop yield: a comparative assessment of ARIMAX and NARX model. *RASHI*, 1(1): 77-85.
- Mitra, D. and R. K. Paul (2017), Hybrid time-series models for forecasting agricultural commodity prices. *Model Assisted Statistics and Applications*, 12: 255–264.
- Zhang, G. P. (2003), Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50: 159-175.
- Zou, H. and Y. Yang (2004), Combining time series models for forecasting. *International Journal of Forecasting*, 20: 69-84.

PART IV

IMPACT ASSESSMENT METHODS

Chapter 17

ECONOMIC SURPLUS APPROACH

Vinayak Nikam, Jaiprakash Bishen, T. K. Immanuelraj, Shiv Kumar and
Abimanyu Jhahria

INTRODUCTION

The role of agricultural projects is crucial in addressing many of the existing socio-economic challenges and to bring out desirable impact on environment and society. While the donor agencies are particularly concerned about the how their projects are effective in generating the positive impacts and the extent which project investments are rationalized. All these necessitate the project impact assessment. The impact of the project is accessed by different impact evaluation/impact assessment methodologies based on the stated objectives of the project. Impact assessment methods essentially evaluate the long term effect of an intervention (deliberate/non-deliberate; planned/unplanned; active/passive and so on) on the intended object/entity under study (man, material, environment, economy, society etc.). Masters *et al.* (1996) advocated different impact evaluation approaches such as econometric approaches, programming approaches, and economic surplus approaches. Each approach has a set of peculiar quantification methods.

Impact assessment framework

The significance of impact assessment relies on the basic framework of research continuum which answers the basic question on when to do impact assessment. The objectives of the research problem and the research framework both define the tools of impact assessment which may be required to get it done (Fig 1).

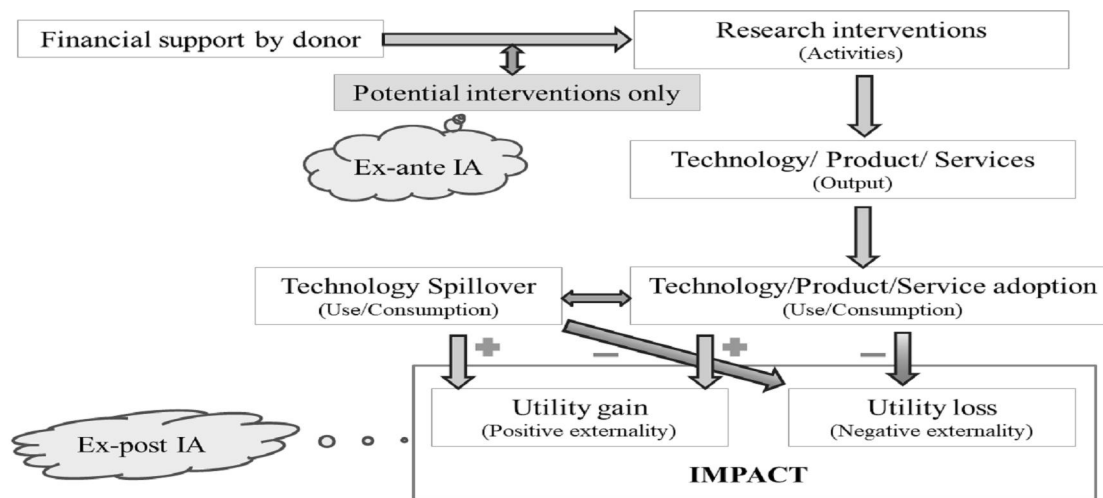


Fig 1: Impact assessment process: an extended framework

Through impact assessment one necessarily measures/quantifies the positive or negative externalities inflicted upon the society or environment by the any cause (such as research interventions). Economic Impact Evaluation (EIE) methods can be employed either pre or post implementation of the project as in *ex-ante* and *ex-post* evaluations respectively. In agriculture, varieties and technologies are two most usual interventions undertaken by any research organizations. As depicted in Fig 1, *ex-ante* impact assessment is done prior to actual implementation of the project which helps in estimating the most likely effect on the target indicator (may be yield, income, employment, poverty etc.) of the activities, which are to be undertaken by the project. *Ex-ante* impact assessment is based on some of the probabilistic assumptions about the most likely activities/causes due to the interventions. However, the *ex-post* impact assessment reckon its foundation on the actual activities/causes due to the interventions undertaken. The tools for impact assessment differ according to the level of impact. For example, if one would like to study the impact of a training program on the participants, IA tools like propensity score matching (PSM) or difference in difference (DID) are some of the powerful tools. However, tools like economic surplus model is a promising tool for measuring a macro level impact like total monetary benefits to the society owing to dissemination and adoption of a high yielding crop variety. Before discussing the economic surplus approach, let's have a look on different approaches to measure the impact of a programme. The following approaches are available at our disposal for impact assessment.

Different Approaches for Impact Assessment

Econometric approach: The econometric approach employs tools such as production and cost functions, or a total factor productivity analysis to attribute the productivity change due to research only. Regression tools (like Probit, Logit, Tobit, and Two-stage least squares (2SLS) regressions) are also used to assess the impact of investment on socio-economic and agro-ecosystem services. However, the availability of quality database in the developing countries poses significant challenges to the use of econometric approaches.

Project approach: It uses the tools like Benefit-Cost ratio (B: C), Net Present Worth (NPW/NPV) of project, Internal Rate of Returns (IRR), Sensitivity Analysis etc. to assess the project impact. But, this approach is purely economic in nature and overlooks the social impact.

Economic surplus approach: Most prominent and copious method employed in evaluating the impacts of investments in agricultural research (Griliches, 1958)

Bio-economic model: It's a synthesis between biophysical and economic information into a single composite model which has the ability to simultaneously addressing various aspects of agriculture and natural resource management (NRM) technological changes which results in trade-offs among objectives of economic sustainability and

environmental protection. (Barbier, 1998; Barbier and Bergerson, 2001; Holden and Shiferaw, 2004; Holden *et.al.* 2004)

Meta-analysis: is effectively an analysis of analyses. Meta-analysis is a modern tool available with the researchers for collating the findings from the past studies, and derives broad conclusions from them. It is helpful to policymakers, who may be confronted by numerous conflicting conclusions

Programming methods: This approach finds the optimal solution to a problem under certain restrictions or constraints. It iterates and find out the most optimal solution from a set of available solutions to the same problem. Thus, it tries to maximize one objective, i.e. farmers' profit/ minimize costs subjected to constraints like availability of land, labour and other inputs.

However, each approach is associated with few merits and demerits which are depicted in Table 1.

Table 1: Merits and demerits of different impact assessment approaches.

IA Approaches	Merits	Demerits
Econometric approach	Can be applied in all the sectors.	Non- availability of time series data on many variables of interest.
Project approach	Give precise results and are quick to estimate.	Only economic impacts are measured. Sensitive to discount rate and life of the project.
Economic surplus approach	Can be used in all the sectors and it requires limited information than any other method.	Sensitive to demand and supply elasticity.
Bio-economic model	More broader approach as whole system is included in this approach	Economic valuation of non-tradable good and services
Meta-analysis	Provides a macro picture	Aggregation bias is associated.

Source: Palanisami *et al.* (2011)

ECONOMIC SURPLUS APPRAOCH

The Economic surplus approach is most exhaustively used tool for evaluating the impact of technology on the economic welfare of households (Moore *et al.*, 2000; Wander *et al.*, 2004; Maredia *et al.*, 2000; Swinton, 2002). It measures the cumulative social gains due to research project/technology. Using economic surplus model (ESM) it is feasible to appraise the return on investments by maneuvering the change in consumer and producer surplus through a technological change due to the research. ESM can also be used along with the research costs to estimate the net present value

(NPV), internal rate of return (IRR), or benefit-cost ratio (BCR) (Maredia *et al.*, 2000). An economic surplus approach takes into account the reduction in per unit cost and price responses owing to research-induced supply shifts and computes the distributary effects of research benefits. The model indicates the extent of research-induced decline in per unit cost of production and in adoption by farmers, may reduce market prices (Norton and Dey, 1993).

Concept of economic surplus

Economic surplus indicates the difference between the monetary values of the units produced and consumed at the equilibrium price and quantity. Mathematically, economic surplus represents the summation of total consumer surplus and total producer surplus ($ES=CS+PS$). The triangle PeD represents the total consumer surplus which is the difference between the willingness to pay of the consumer for a product/service without forgoing it and what he actually pays in exchange of getting the product/service (Fig 2).

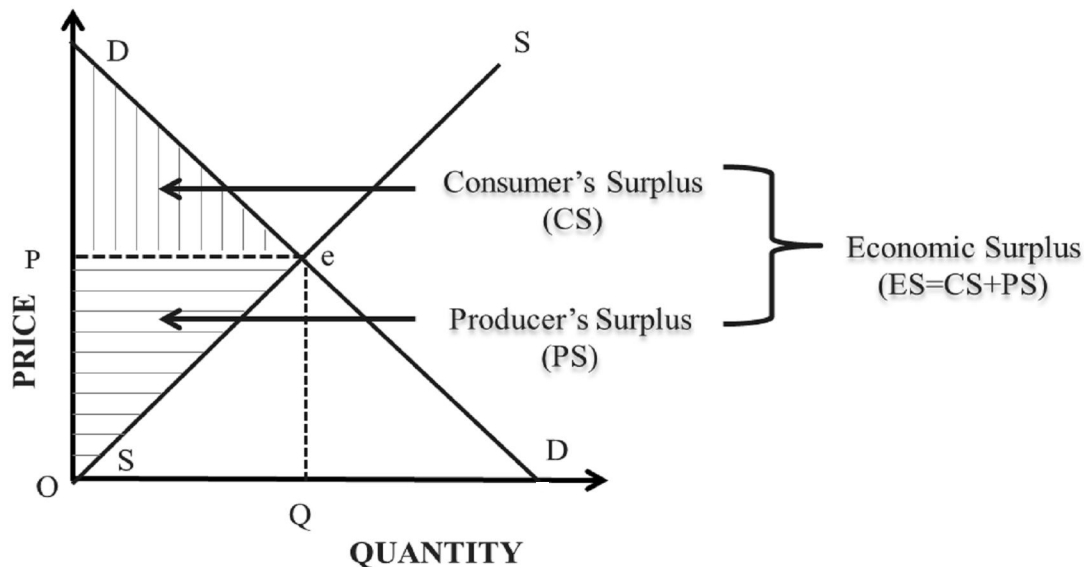


Fig 2: Concept of economic surplus

Similarly the triangle OeP represents the total producer surplus which is the difference between what a producer receives after selling the product/service in the market and his willingness to sell the same product/service at a particular price. However, the triangle OeD represents the total economic surplus accrued to the society due to the product/service which is equivalent to area of the triangle PeD and OeP . However, the three different surpluses (producer surplus, consumer surplus and economic surplus) are sensitive to any changes in demand or supply elasticity (Fig 3).

Masters *et al.* (1996) reported that the elasticity of supply and demand do not have significant influence on estimation of economic surplus in comparison to other variables, i.e. price, productivity, quantity etc. However, they have a pronounced effect

on distribution of economic benefits of among the producer as well as user groups. Here we have discussed the three cases and distribution of benefits of research among the producer and user groups (Table 2).

Table 2: Effect of changes in elasticity of demand and supply on distribution of benefits

Cases	Relation between E_s & E_d	Distribution of Research Benefits
Case-I	Elasticity of demand and supply are same i.e. $E_s = E_d$	$\Delta CS = \Delta PS$
Case-II	Elasticity of demand is more than supply elasticity i.e. $E_s < E_d$	$\Delta CS < \Delta PS$
Case-III	Elasticity of demand is less than supply elasticity i.e. $E_s > E_d$	$\Delta CS > \Delta PS$

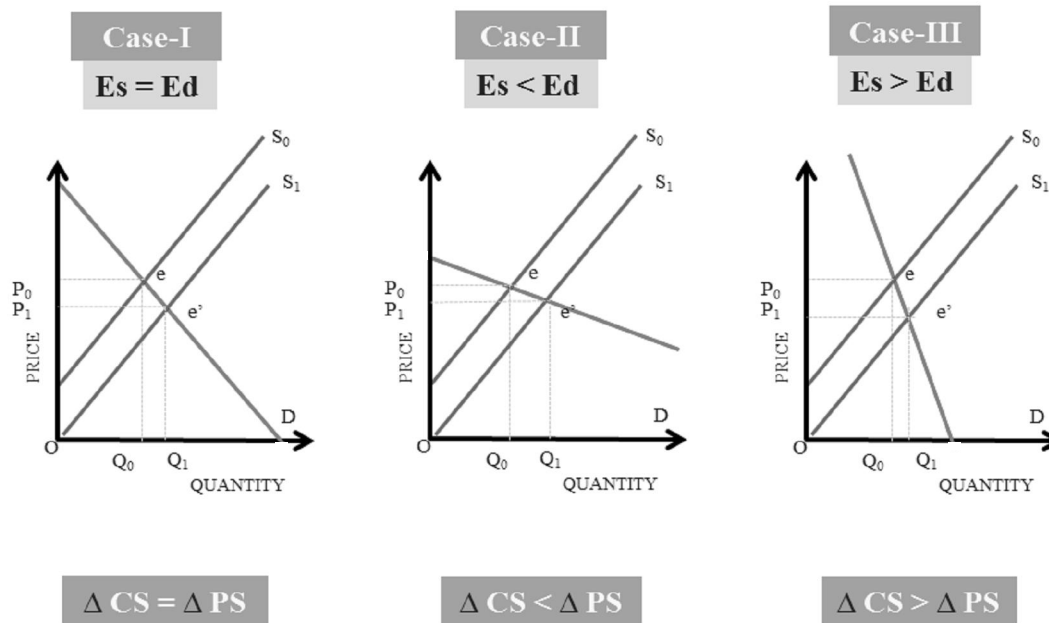


Fig 3: Effect of changes in elasticity of demand and supply on distribution of benefits

Economic surplus model

The economic surplus model derives its root from the Marshallian theory of economic surplus that indeed has roots in shifts in the supply and demand curves over the time (Palanisami *et al.*, 2009). The economic surplus approach aims at measurement/quantification of the total economic surplus accrued to the society due to an intervention (technological change or any other which shifts the supply curve). Measurement of economic surplus is also based on some of the tools of welfare economics which look at how the interaction of demand and supply determine the value of economic transaction. However, given the constant demand function the measurement of

economic surplus indicates the effect of technological change on the supply. Economic surplus approach is a proven tool for *ex-ante* impact assessment of any research/ technological intervention.

Assumptions of economic surplus model

Ex-ante impact assessment methods are based on some probabilistic assumptions about the deliberated interventions and their anticipated adoption. These, assumptions transform the complex real life situations into a simple one which are amenable to analyze and measure. According to Harbenger (1971) and Alston (1995) the economic surplus model is based on following assumptions -

- 1) A parallel shift of the supply curve following the adoption of technology/ investment
- 2) The functional form of the supply curve is unknown
- 3) Assumption of closed economy
- 4) In context of open economy, country under consideration is the only exporter of product and rest of world do not adopt the technology
- 5) The competitive demand price for a given unit measures the value of that unit to the consumer
- 6) The competitive supply price for a given unit measures the value of that unit to the producer
- 7) The costs and benefits accruing to each member of the relevant group should be added

Model specifications

The economic surplus approach is applicable to the closed and open economies both. This section emphasizes on the mathematical model for the two variants (parallel and proportionate shift in the supply curve) for the closed economy conditions. The parallel shift in supply curve is result of an improved technology adoption which shifts the supply curve to the right side and resultant increase in quantity supplied and reduced price (Fig 2). For example, adoption of a new high yielding variety/new technology results in increases in production and thereby the increased quantum of output in the market (Variant 1). However, the knowledge up gradation or imparting skills will strengthen the management practices in the field which translate into proportionate shift in supply curve (Variant 2).

Case I: Close economy: In the closed economic condition, country is assumed to be self-sufficient for its needs i.e. the economy neither produces surplus nor deficit and therefore the case of external trade is ruled out.

Closed economy with parallel supply shift

For a closed economy with a parallel supply shift that results from an improved technology/ new technology, the annual changes in market level total economic surplus

(ΔTS) can be measured as (Fig 4).

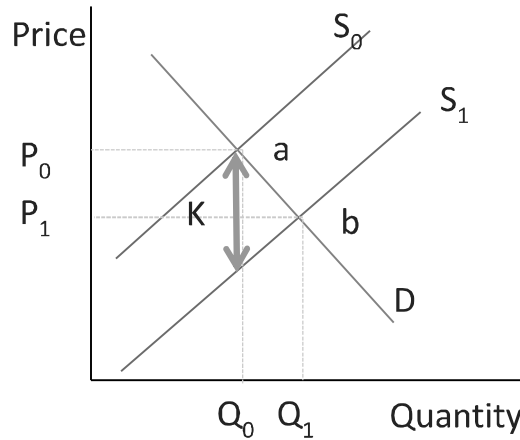


Fig 4: Impact of research on closed economy with a parallel supply shift

Economic surplus can be calculated using following formula.

$$\Delta CS = P_0 Q_0 Z (1 + 0.5 Z \eta)$$

$$\Delta PS = P_0 Q_0 (K - Z) (1 + 0.5 Z \eta)$$

$$\Delta TS = \Delta CS + \Delta PS = P_0 Q_0 K (1 + 0.5 Z \eta)$$

Where,

P_0 = Base price of the commodity

Q_0 = Base quantity,

η = Absolute value of the price elasticity of demand

$Z = K \varepsilon / (\varepsilon + \eta)$ or the proportionate price reduction in the market, where ε is the elasticity of supply

K_t = Proportionate reduction in cost per ton of production in time t or research induced shift in supply and it can be calculated as-

$$K_t = ((E(Y)/\varepsilon) - (E(C)/(1+E(Y))) A t (1-d)^t$$

Where,

$E(Y)$ = Proportionate yield increase per hectare for technology adopters

ε = Price elasticity of supply

$E(C)$ = Proportionate variable input cost change per hectare

A = Proportion of the area affected by the technology

d = Depreciation rate of the technology

(Alston *et al.*, 1995)

Closed economy with proportionate supply shift

For a closed economy with a proportionate supply shift that results from an improved knowledge/ management practices, the annual changes in market level total economic surplus (ΔTS) can be measured as (Fig 5):

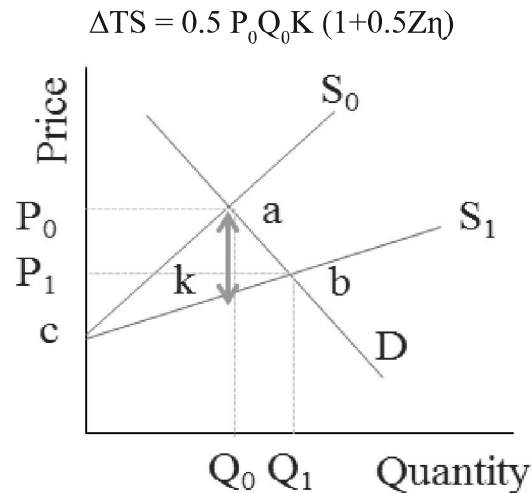


Fig 5: Impact of research on closed economy with a proportional supply shift

Data requirement

Estimation of economic surplus requires data on production, real prices, price elasticities of supply and demand, expected yield increases and reduction in cost, probability of research success, time to complete the research, adoption rates, and discount rate. To calculate net benefits, information on research and development costs is also required (Nikam *et al.*, 2019).

Merits and demerits of economic surplus model

The popularity of economic surplus model as a prudent tool for *ex-ante* impact assessment is due to segregation of total economic benefits due to a technology, among major stakeholders (producers and consumers). It also provides an estimate for gross benefits to the society from a research or technological development. Nderim (2008) reported its significance as a tool for both *ex-ante* as well as *ex-post* impact assessment. The other merits of ESM include its requirement of minimum data to provide the most precise results and ease in using the model. Despite these merits, some demerits of ESM approach includes the ignorance about the transaction cost which results in the overestimation of benefits; partial equilibrium nature of analysis which ignores the effect of any relationship with other product and factor in the market and it provides only the gross benefit of any intervention and ignores the net benefit.

Improvement over economic surplus model

There are some improvements over the methodology of impact assessment over time. And some of these methodologies give more accurate results than the approach of economic surplus. These improvements are:

- 1) Equivalent Variation (EV) - It takes into account the income effect of the price change,
- 2) Econometric Approach - Gives most reliable results

Tools for computation of economic surplus

Some of the important tools for computation of economic surplus and their developers are indicated in Table 3.

Table 3: Tools for computation of economic surplus

S. No.	Tools	Developed By
1.	MODEXC	International Centre for Tropical Agriculture (CIAT)
2.	RE4	Australian Centre for International Agricultural Research (ACIAR)
3.	Dynamic Research Evaluation for Management (DREAM)	International Food Policy Research Institute (IFPRI)
4.	Spreadsheet Approach	-

ILLUSTRATION

Explanation of economic surplus using practical example

Case of mobile app by ICAR-NRC Grapes: mobile based app that provides real time information as well as forecasting about weather, pest and diseases of the grape crop in Maharashtra state of India. Individualized timely and accurate forecasting about weather and pests and diseases, decision support system and information about grape cultivation etc. through the app helps the farmers to save the inputs as well as increase in the yield. The project for its development was started in 2008 and it was completed developed and commercialized in the year 2012. Till the year 2017 it was adopted by 15% of the total grape growers in Maharashtra.

Closed economy model was assumed in this case only about 5 per cent of grapes are exported and remaining consumed domestically. To calculate total economic surplus generated by the app at macro level (here at state level), we need following different types of data. (Example is given for understanding purpose only and may not necessarily reflect the actual values).

a) Value of increase in yield and change in cost of cultivation

This can be obtained by primary survey of the grape growers. It can be done using two approaches. Before and after- in which data would be collected from the farmers

before and after adoption of the app. Increase in yield after and decrease in cost of cultivation after the adoption of mobile app would be obtained in percentages. Suppose the increase in yield is 11 per cent and decreasing the cost of cultivation is 0.12 per cent.

b) Elasticity of demand and supply

Elasticity of demand and supply in relation to prices may be calculated using standard formulae. The other way by which we can obtain these values using review of literature. Example- In Indian scenario, Kumar *et al.* (2011) have estimated the demand elasticity for various agriculture commodities. From this paper we obtained the demand elasticity of fruit crop which is -0.595. A demand elasticity of -0.595 implies that a one percent price reduction increases the quantity demand of grapes by 0.595 percent. Similarly supply elasticity of grapes was obtained by review of literature which was considered as 0.40 per cent. A supply elasticity of 0.4 implies that a one percent increase in price only increase the quantity supplied by only 0.4 percent.

c) Adoption rate of the technology

Adoption rate of the mobile app is given in Fig 4. We also project the economic surplus over the period of time like for next five years or 10 years. For that purpose, we need to assume the adoption rate of the technology in future. This can be obtained by discussion with various stakeholders involved in the technology (Mobile app developer, Scientists from NRC Grapes, grape growers etc.). While projecting the adoption rate care should be taken that the adoption rate follows standard adoption curve given by (Rogers, 2003). It shows that over certain period of time for any technology, adoption rate increases up to certain points and starts declining thereafter. This may be because of technology becomes obsolete, or some other similar competitive product has been developed. In this case with discussion with various stakeholders, adoption rate was assumed to be 20 per cent in 2020 and decline thereafter (Fig 6).

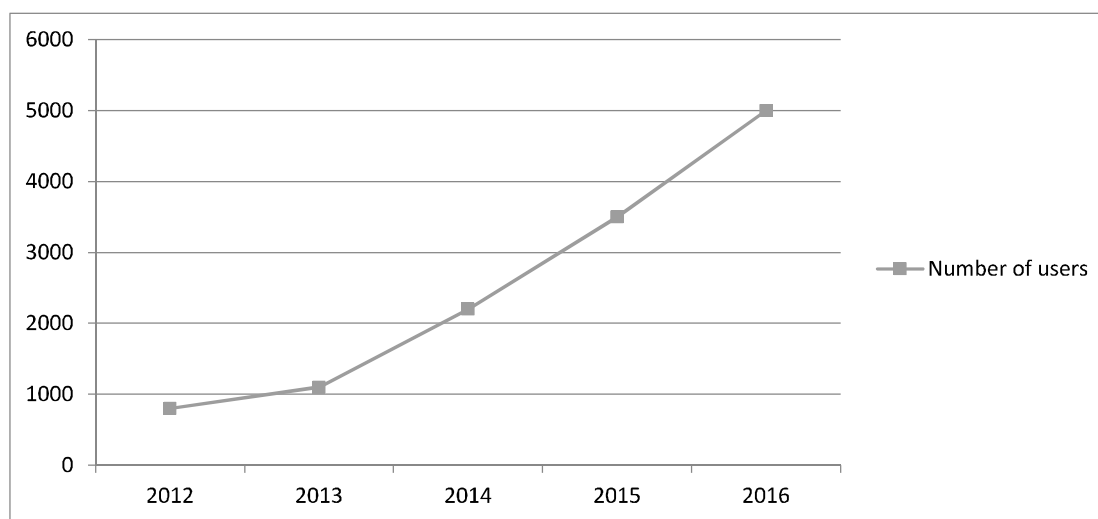


Fig 6: Adoption of mobile app over time

d) Data about production of grapes and prices.

Data about the production and prices of grapes in Maharashtra state can be obtained in time series from 2008 to 2017 from government official website like (www.data.gov.in). Production quantity can be converted into kilogram. In case of prices obtained year wise in time series from the government data, we need to adjust it as per inflation and convert it into the real prices. For this we can obtain the index of wholesale prices from the website of Directorate of Economics and Statistics and convert the year wise price into real price. Remember that we also need to extrapolate the production and prices (based on previous data) for the next five to ten year for estimating the economic surplus likely to be generated by the technology in near future. Using regression equation, we can extrapolate the figures for next five to ten years as per requirement.

e) Probability of success

Probability of success of the technology ranges from 0 to 1. Considering changing nature of ICT technology, we have considered the probability of success at 50 per cent

f) Depreciation factor of technology

Depreciation factor implies that how soon this technology become depreciate or obsolete. Mathematically, it can be calculated as one minus annual rate of depreciation. We have assumed that for initial six years from introduction this technology does not depreciate, after then it starts depreciating due to emergence of new competitors who can provide better service. It is also possible that with increasing scale of operation, quality of the service may deteriorate.

g) Cost of research, development and extension

Cost of the research could be obtained from the Research Project Proposal-I (RPP-I) of the project from which the technology was developed. Here, cost of research from 2008 was obtained from RPP-I of NRC Grapes. Development cost in terms of developing app, hiring the server, cost of forecasting, cost of establishing the automatic weather stations at farmers' field and maintenance cost like hiring of staffs to do all these activities was considered and was obtained from the company operating the mobile app.

h) Discount rate

It refers to the interest rate used to determine the present value. Mruthyunjaya *et al.* (2004) have used interest rate of eight per cent to evaluate the returns on research investment by National Agricultural Technology Project (NATP) on various agricultural projects. As it is time preference concept, we have used 10 per cent of interest rate to calculate the Net Present Value in analysis.

Various formula used for calculation of NPV, IRR, consumer surplus, producer surplus, and total surplus are given in Fig 7. Results obtained are give in Fig 8.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2																						
3	year	Price elasticity of supply	Price elasticity of demand	Maximum yield change	Red. in marginal cost	Input cost change per ha	Input cost change per kg	Net cost change per	Prob. of success	Adoption rate	Dep. rate of Tech.	K	Z	Grapes Price per kg unit current	real price	Quantity of grapes production in kg	Change in consumer surplus	change in producer surplus	Change in total surplus	Research Cost	Net benefit	Results
4																						
5																						
6	2007	0.4	0.6	0.1	=D6B6	-0.1	=B7E(-D20)	=B5E4G6	0.5	0.0	1.0	=B8E3B6A6K6	=E6F8B6B6C6	35.3	40.3	1.29E-09	=A3B6D6E6F6H6J6K6	=D6E6F6G6H6I6J6K6	=B3A6D6E6F6H6J6K6	3.19E+05	=(S6-T6)	
7	2008	0.4	0.6	0.1	=D7B7	-0.1	=B7F(-D20)	=B5E4G7	0.5	0.0	1.0	=B8E3B7A7K7	=E7F8B7B7C7	35.5	30.6	1.42E-09	=A3B7D7E7F7H7J7K7	=D7E7F7G7H7I7J7K7	=B3A7D7E7F7H7J7K7	3.67E+05	=(S7-T7)	
8	2009	0.4	0.6	0.1	=D8B8	-0.1	=B7F(-D20)	=B5E4G8	0.5	0.0	1.0	=B8E3B8A8K8	=E8F8B8B8C8	30.5	30.4	4.40E-08	=A3B8D8E8F8H8J8K8	=D8E8F8G8H8I8J8K8	=B3A8D8E8F8H8J8K8	3.74E+05	=(S8-T8)	
9	2010	0.4	0.6	0.1	=D9B9	-0.1	=B7F(-D20)	=B5E4G9	0.5	0.0	1.0	=B8E3B9A9K9	=E9F8B9B9C9	37.1	43.8	7.74E-08	=A3B9D9E9F9H9J9K9	=D9E9F9G9H9I9J9K9	=B3A9D9E9F9H9J9K9	3.75E+05	=(S9-T9)	
10	2011	0.4	0.6	0.1	=D10B10	-0.1	=B7F(-D20)	=B5E4G10	0.5	0.0	1.0	=B8E3B10A10K10	=E10F8B10B10C10	45.2	46.6	1.81E-09	=A3B10D10E10F10H10J10K10	=D10E10F10G10H10I10J10K10	=B3A10D10E10F10H10J10K10	3.79E+05	=(S10-T10)	
11	2012	0.4	0.6	0.1	=D11B11	-0.1	=B7F(-D20)	=B5E4G11	0.5	0.0	1.0	=B8E3B11A11K11	=E11F8B11B11C11	39.0	31.4	2.05E-09	=A3B11D11E11F11H11J11K11	=D11E11F11G11H11I11J11K11	=B3A11D11E11F11H11J11K11	7.24E+06	=(S11-T11)	
12	2013	0.4	0.6	0.1	=D12B12	-0.1	=B7F(-D20)	=B5E4G12	0.5	0.0	1.0	=B8E3B12A12K12	=E12F8B12B12C12	31.5	31.5	2.16E-09	=A3B12D12E12F12H12J12K12	=D12E12F12G12H12I12J12K12	=B3A12D12E12F12H12J12K12	5.74E+06	=(S12-T12)	B-C
13	2014	0.4	0.6	0.1	=D13B13	-0.1	=B7F(-D20)	=B5E4G13	0.5	0.1	1.0	=B8E3B13A13K13	=E13F8B13B13C13	47.2	63.0	2.29E-09	=A3B13D13E13F13H13J13K13	=D13E13F13G13H13I13J13K13	=B3A13D13E13F13H13J13K13	7.14E+06	=(S13-T13)	=B15V17
14	2015	0.4	0.6	0.1	=D14B14	-0.1	=B7F(-D20)	=B5E4G14	0.5	0.1	1.0	=B8E3B14A14K14	=E14F8B14B14C14	44.9	26.8	2.29E-09	=A3B14D14E14F14H14J14K14	=D14E14F14G14H14I14J14K14	=B3A14D14E14F14H14J14K14	6.50E+06	=(S14-T14)	NPB
15	2016	0.4	0.6	0.1	=D15B15	-0.1	=B7F(-D20)	=B5E4G15	0.5	0.2	1.0	=B8E3B15A15K15	=E15F8B15B15C15	31.1	35.4	2.64E-09	=A3B15D15E15F15H15J15K15	=D15E15F15G15H15I15J15K15	=B3A15D15E15F15H15J15K15	6.66E+06	=(S15-T15)	=B16W18J1521
16	2017	0.4	0.6	0.1	=D16B16	-0.1	=B7F(-D20)	=B5E4G16	0.5	0.2	1.0	=B8E3B16A16K16	=E16F8B16B16C16		33.5	2.78E-09	=A3B16D16E16F16H16J16K16	=D16E16F16G16H16I16J16K16	=B3A16D16E16F16H16J16K16	7.06E+06	=(S16-T16)	NPC
17	2018	0.4	0.6	0.1	=D17B17	-0.1	=B7F(-D20)	=B5E4G17	0.5	0.2	0.9	=B8E3B17A17K17	=E17F8B17B17C17		33.5	2.97E-09	=A3B17D17E17F17H17J17K17	=D17E17F17G17H17I17J17K17	=B3A17D17E17F17H17J17K17	6.80E+06	=(S17-T17)	=B16W18J1621
18	2019	0.4	0.6	0.1	=D18B18	-0.1	=B7F(-D20)	=B5E4G18	0.5	0.2	0.9	=B8E3B18A18K18	=E18F8B18B18C18		33.5	3.17E-09	=A3B18D18E18F18H18J18K18	=D18E18F18G18H18I18J18K18	=B3A18D18E18F18H18J18K18	6.96E+06	=(S18-T18)	
19	2020	0.4	0.6	0.1	=D19B19	-0.1	=B7F(-D20)	=B5E4G19	0.5	0.3	0.8	=B8E3B19A19K19	=E19F8B19B19C19		33.5	3.36E-09	=A3B19D19E19F19H19J19K19	=D19E19F19G19H19I19J19K19	=B3A19D19E19F19H19J19K19	7.06E+06	=(S19-T19)	
20	2021	0.4	0.6	0.1	=D20B20	-0.1	=B7F(-D20)	=B5E4G20	0.5	0.3	0.8	=B8E3B20A20K20	=E20F8B20B20C20		33.5	3.55E-09	=A3B20D20E20F20H20J20K20	=D20E20F20G20H20I20J20K20	=B3A20D20E20F20H20J20K20	7.16E+06	=(S20-T20)	
21	2022	0.4	0.6	0.1	=D21B21	-0.1	=B7F(-D20)	=B5E4G21	0.5	0.3	0.7	=B8E3B21A21K21	=E21F8B21B21C21		33.5	3.75E-09	=A3B21D21E21F21H21J21K21	=D21E21F21G21H21I21J21K21	=B3A21D21E21F21H21J21K21	7.26E+06	=(S21-T21)	CS
22																						=B16W18J1621
23																						PS
24																						=B16W18J1621
25																						TS
26																						=B16W18J1621

Fig 7: Screenshot of spreadsheet showing the formulae used for calculation of surplus

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	year	Price elasticity of supply	Price elasticity of demand	Maximum yield change	Red. in marginal cost	Input cost change per ha	Input cost change per kg	Net cost change per	Prob. of success	Adoption rate	Dep. rate of Tech.	K	Z	Grapes Price per kg unit current	real price	Quantity of grapes production in kg	Change in consumer surplus	change in producer surplus	Change in total surplus	Research Cost	Net benefit	Results
2																						
3																						
4																						
5																						
6	2007	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	35.3	40.3	1.29E+09	0.00E+00	0.00E+00	0.00E+00	3.19E+05	-3.19E+05	
7	2008	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	35.5	30.6	1.42E+09	0.00E+00	0.00E+00	0.00E+00	3.67E+05	-3.67E+05	
8	2009	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	30.5	30.4	4.40E+08	0.00E+00	0.00E+00	0.00E+00	3.74E+05	-3.74E+05	
9	2010	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	37.1	43.8	7.74E+08	0.00E+00	0.00E+00	0.00E+00	3.75E+05	-3.75E+05	
10	2011	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	45.2	46.6	1.81E+09	0.00E+00	0.00E+00	0.00E+00	3.79E+05	-3.79E+05	
11	2012	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	39.0	31.4	2.05E+09	1.24E+08	1.85E+08	3.09E+08	7.24E+06	3.01E+08	
12	2013	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.0	1.0	0.0	0.0	31.5	31.5	2.16E+09	1.78E+08	2.65E+08	4.44E+08	5.74E+06	4.38E+08	B-C
13	2014	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.1	1.0	0.0	0.0	47.2	63.0	2.29E+09	7.57E+08	1.13E+09	1.88E+09	7.14E+06	1.88E+09	542.3
14	2015	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.1	1.0	0.0	0.0	44.9	26.8	2.29E+09	4.74E+08	7.05E+08	1.18E+09	6.56E+06	1.17E+09	NPB
15	2016	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.2	1.0	0.0	0.0	31.1	35.4	2.64E+09	1.13E+09	1.67E+09	2.80E+09	6.66E+06	2.79E+09	1.57E+10
16	2017	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.2	1.0	0.0	0.0		33.5	2.78E+09	1.23E+09	1.83E+09	3.06E+09	7.06E+06	3.05E+09	NPC
17	2018	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.2	0.9	0.0	0.0		33.5	2.97E+09	1.46E+09	2.17E+09	3.62E+09	6.86E+06	3.61E+09	2.89E+07
18	2019	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.2	0.9	0.0	0.0		33.5	3.17E+09	1.67E+09	2.49E+09	4.16E+09	6.96E+06	4.16E+09	
19	2020	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.3	0.8	0.0	0.0		33.5	3.36E+09	1.81E+09	2.69E+09	4.51E+09	7.06E+06	4.50E+09	
20	2021	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.3	0.8	0.0	0.0		33.5	3.55E+09	1.93E+09	2.88E+09	4.81E+09	7.16E+06	4.80E+09	
21	2022	0.4	0.6	0.1	0.3	-0.1	-0.1	0.4	0.5	0.3	0.7	0.0	0.0		33.5	3.75E+09	2.04E+09	3.03E+09	5.07E+09	7.26E+06	5.07E+09	CS
22																						3.91E+09
23																						PS
24																						5.82E+09
25																						TS
26																						9.74E+09

Fig 8: Screenshot of spreadsheet showing the values of the various parameters in estimation of economic surplus

REFERENCES

- Alston, J. M., G. W. Norton and P. G. Pardey (1995), Science under scarcity: principles and practice for agricultural research and evaluation and priority setting, Cornell University Press, Ithaca, New York.
- Barbier, B. (1998), Induced innovation and land degradation: Results from a bio-economic model of a village in West Africa. *Agricultural Economics*, 19: 15-25
- Barbier, B. and Bergeron wander (2001), Natural resource management in the hillsides of Honduras: bio-economic modelling at the micro watershed level. Research Report No. 123, International Food Policy Research Institute (IFPRI), Washington D.C, U.S.A. 59 p.
- Dey, M. and G. Norton (1993), Analysis of Agricultural Research Priorities in Bangladesh. BARC, Dhaka, Bangladesh.
- Griliches, Z. (1958), Research costs and social returns: hybrid corn and related innovations. *Journal of Political Economy*, 66 (5): 419-431
- Harberger, A. C. (1971), Three basic postulates for applied welfare economics: an interpretive essay. *Journal of Economic Literature*, 9:785-797
- Holden, S. and B. Shiferaw (2004), Land degradation, drought and food security in a less-favoured area in the Ethiopian highlands: a bio-economic model with market imperfections. *Agricultural Economics*, 30: 31-49.
- Holden, S., B. Shiferaw and J. Pender (2004), Non-farm income, household welfare, and sustainable land management in a less-favoured area in the Ethiopian Highlands, *Food policy*, 29: 369-392.
- Kumar, P., Anjani Kumar, P. Shinoj and S. S. Raju (2011), Estimation of Demand Elasticity for Food Commodities in India. *Agricultural Economics Research Review*, 24 (1).
- Masters, W. A., B. Coulibaly, D. Sanogo, M. Sidibé and A. Williams (1996), The economic impact of agricultural research: a practical guide. Department of Agricultural Economics, Purdue University, West Lafayette, IN. [Available by e-mail: Masters@AgEcon. Purdue.edu]
- Moore, R. Michael, Gollehon Noel R. and M. Hellerstein Daniel (2000), Estimating producer's surplus with the censored regression model: an application to producers affected by Columbia river basin salmon recovery. *Journal of Agricultural and Resource Economics*, 25(2): 325-346.
- Mruthyunjaya, S. Pal, L. M. Pandey and A. K. Jha (2004), Impacts of selected technologies refined under NATP. *Agricultural Economics Research Review*, Agricultural Economics Research Association (India), vol. 17 (Conference issue).
- Nderim, R. (2008), Annex ante economic impact analysis of developing low cost technologies for pyramiding useful genes from wild relatives into elite progenitors

- of cassava, M.Sc. Thesis, Virginia Polytechnic Institute and State University, pp 28-34
- Nikam, V., S. Kumar and I. T. Kingsley (2019), Evaluation of watershed development programmes in India using economic surplus method. *Indian Journal of Agricultural Sciences*, 89 (6): 1039–43
- Palanisami, K., D. Suresh Kumar and S. Nedumaran (2011), Methodology in evaluation of watershed programmes, NCAP, proceedings 15, pp.10
- Palanisami, K., D. Suresh Kumar, S. P. Wani and M. Giordono (2009), Evaluation of watershed development programmes in India using economic surplus method. *Agricultural Economics Research Review*, 22: 197-207
- Rogers, E. M. (2003), Diffusion of innovations (5th ed.). New York: Free Press.
- Swinton, S. M. (2002), Integrating sustainability indicators into the economic surplus approach for NRM impact assessment. In: Methods for Assessing the Impacts of Natural Resources Management Research. A summary of the proceedings of the ICRISAT-NCAP/ICAR International Workshop, Eds: B. Shiferaw, H. A. Freeman, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad, 6-7 December.
- Wander, Alcido Elenor, Magalhaes, Marilia Castelo. Vedovoto, Graciela Luzia and Martins, Espedito Cezario (2004), Using the economic surplus method to assess economic impacts of new technologies — Case studies of EMBRAPA, Rural Poverty Reduction through Research for Development — Conference on International Agricultural Research for Development, Deutscher Tropentag, Berlin, 5-7 October.

Chapter 18

INTRODUCTION TO CAUSAL INFERENCE

Arathy Ashok

INTRODUCTION

Causal inference is a science of inferring the presence and magnitude of cause-effect relationship from the data. Causal inference as a discipline, investigates how to design and analyze empirical studies in order to infer the effects of intervention or policies. In the evaluation research, it helps to identify the effect of an exposure, treatment or intervention on an outcome variable through experimental or observational studies. The basic concepts of impact evaluation stem from the causal inference theories in statistics and it has got widespread application in medical research, sociology, economics etc. Understanding the formal language of causal inference, causal assumptions and the analysis methods can help the researchers to appropriately design and analyze an evaluation study. In this chapter an attempt had been made to give a general understanding on

1. Association Vs. causation
2. Confounding
3. Graphical methods to identify causal effects
4. Counterfactuals in causal inference
5. Techniques for estimating causal effect

Association Vs. causation

Causal inference is completely different from the standard statistical inference which tries to establish the associations. For example, a study among school students showed that there is a positive correlation between height and performance in the class. This association does not indicate that increase in height make the students a better performer in the class. Rather, it is a spurious association.

Similarly, if we consider a simple linear regression which can be indicated as

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

where Y, dependent variable; X, independent variable and β_1 is the regression coefficient. Here β_1 indicates the average change in the dependent variable (Y) with unit change in independent variable X. The regression coefficient β_1 may not indicate causation, as there are chances of existence of other factors/variables (omitted variables) which is associated with the dependent variable Y, rather than only X. If those omitted/missing

variables affect both independent and dependent variables simultaneously, then the simple linear regression will not give good estimate of causal effect of X on Y.

Suppose we add some more variables to the previous model, resulting in a multiple regression of the following form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \dots \quad (2)$$

There are chances that the regression coefficient β_1 in equation (1) may be different from β_1 in equation (2). The estimation bias in both equation 1 and 2 mentioned above is mainly due to confounding effect. So in general, we can say that association is not always equal to causation. To be causal, the association between an independent variable (treatment/exposure variable) and dependent variable (outcome variable) must be supplemented by additional assumptions or data.

Confounding

Confounding is a kind of systematic error/bias which occurs due to distortion of exposure/treatment by some other factors. It can also be explained as the effect of a third variable that accounts for all or some of the association between exposure and outcome (Fig 1). An uncontrolled confounder is considered as one of the common causes of endogeneity.

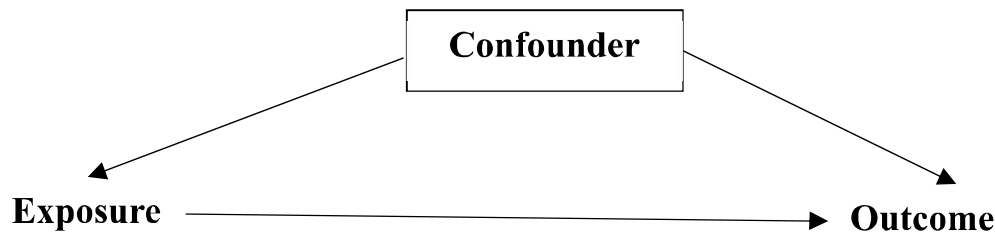


Fig 1: Confounders in cause effect association

A confounder(s) could be a variable or group of variables which is

- Associated with exposure/treatment
- Associated with outcome but not a consequence of exposure

To have a better understanding of the direct causal effects, indirect causal effects and confounders in causal inference, we will use graphical methods in the next session.

Graphical methods to identify causal effects

During late 1990s, pioneering efforts were made to explain the complex models of causal effects in empirical research using graphical methods. Pearl (1995) used the directed acyclic graphs (DAGs) which gave a non-parametric representation of association between different variables, based on expert knowledge of how the data were generated. DAGs consists of two elements:

- Variables/vertices/ nodes
- Unidirectional arrows/edges/ paths

A simple DAG looks like the following where X, I and Y are nodes and the unidirectional arrows/edges represent the association paths (Fig 2).



Fig 2: Directed acyclic graph

Theoretically DAGs are absolute mathematical representations with two distinct functions; sets of probability distributions and causal structures. DAGs that are interpreted causally are called causal DAGs. To be causal, the DAG incorporates two main assumptions;

1. Exclusion restriction: when there is no arrow directly from X to Y, manipulating X will not change Y, unless all parents of Y are manipulated.
2. No omitted confounders assumption: All common causes of any two variables are included as a variable on the DAG.

Some possible associations between the exposure and outcome variables are explained with helps of DAGs is given in Table 1.

Table1: Representation of different types of association between exposure and outcome in DAG

1.	<pre> graph LR X --> I I --> Y X --> Y </pre>	DAG representing direct effect of X on Y and indirect effect of X on Y through a mediating variable I
2.	<pre> graph LR C --> X C --> Y X --> Y </pre>	DAG representing the effect of treatment X on outcome Y, in the presence of confounder C
3.	<pre> graph LR U --> I U --> Y X --> I I --> Y X --> Y </pre>	DAG where the unmeasured confounder, U affecting the mediating variable I and the outcome Y

To understand which path contributes to the causal association and which do not, the graphical rule of separation can be used. It helps to verify the concept of independence based on a DAG. To understand this better, we can consider the DAG as an electric circuit. Then in case of the DAG in Fig 3, as both X and Y collide on I, there will not be any association between X and Y or we can say that X and Y are independent.



Fig 3: Colliders/inactive/independent

Similarly, if there are no colliders present in the DAG, it represents an active path and there may be association between X and Y along all active paths (Fig 4).

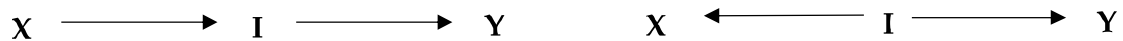


Fig 4: Non colliders/ active path

In case of Fig 4, the former DAG, representing non colliders, variable I blocks the directed path between X and Y and hence X and Y are d separated by I, the mediating variable. Considering the probability distribution in the above mentioned DAG, we can say that X is independent of Y conditional on Y.

With the above mentioned conditions we can identify different possible cause effect association between X and Y from the DAG given in Fig 5:

1. Direct causal effect between X and Y
2. Indirect causal effect between X and Y through mediating variable I
3. But there is no causal effect due to spurious association through X-I-U-Y (I is a collider in this path)

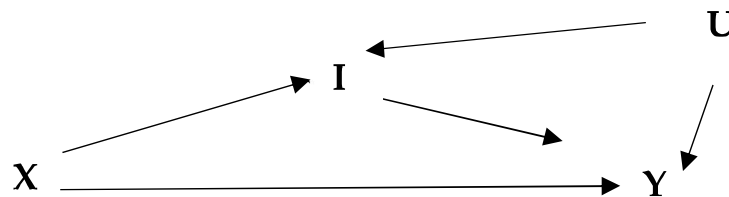


Fig 5: different causal pathways in a DAG

We can adjust for non-colliders (active to inactive) or the colliders and their descendants (inactive to active) and the d separation helps to identify for which confounder we need to adjust/condition when estimating the causal effect.

Counterfactuals in causal inference

Consider the following example of farmers in a drought prone area where they suffer from severe yield losses due to drought conditions. Suppose, among them, few farmers adopted a specific drought tolerant crop variety and observed an increase in yield and

we want to assess the actual effect of the drought tolerant variety on crop yield. A small portion of the data is given in Table 2.

Table 2: Yield outcomes based on technology adoption among farmers

Farmers	X Adopted drought tolerant variety? (Yes=1, No=0)	Y Whether there is an increase in yield? (Yes=1, No=0)
Farmer1	0	0
Farmer2	1	1
Farmer3	1	1
Farmer4	0	0
Farmer5	0	1
Farmer6	1	0
Farmer7	1	1
Farmer8	0	0
Farmer9	0	1

Take the case of Farmer 2,3 and 7 in Table 1. They had adopted the drought tolerant variety and observed an increase in yield. Is it the technology adoption that cause an increase in crop yield? We cannot answer this question unless we know what would had happened those farmers (farmer 2, 3, and 7), if they did not adopt the technology. Here the outcome has been observed in situations that did not actually happen (counter to the fact situation) and it is called a counterfactual.

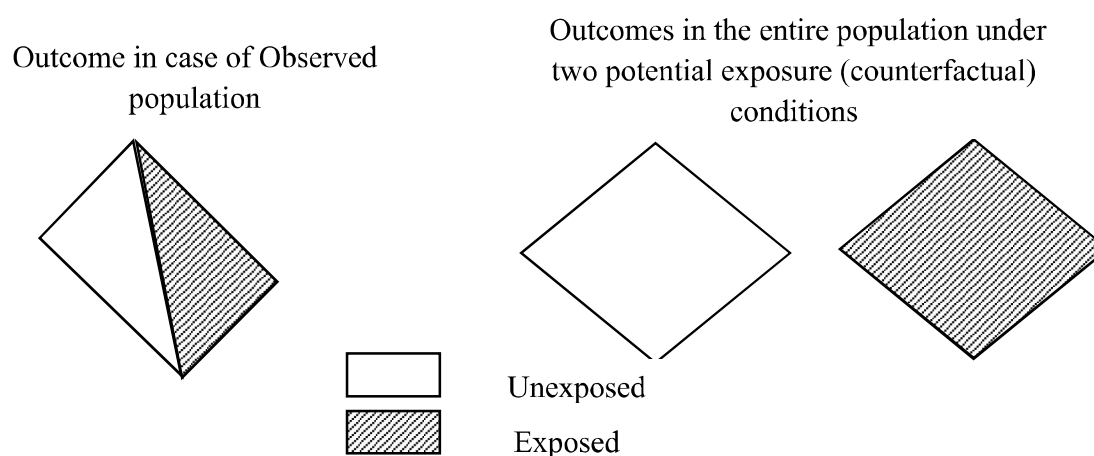


Fig 6: Conceptual difference between outcomes in observed population and counterfactual

In order to estimate causal effect, we need to have counterfactuals. The difference between the actual outcome (what happened to the treated/exposed group) and the

counterfactual outcome (what would have happened if they had not been treated) will give the causal effect of the treatment/exposure on outcome. As we cannot directly observe the counterfactual, we need to estimate it in some way.

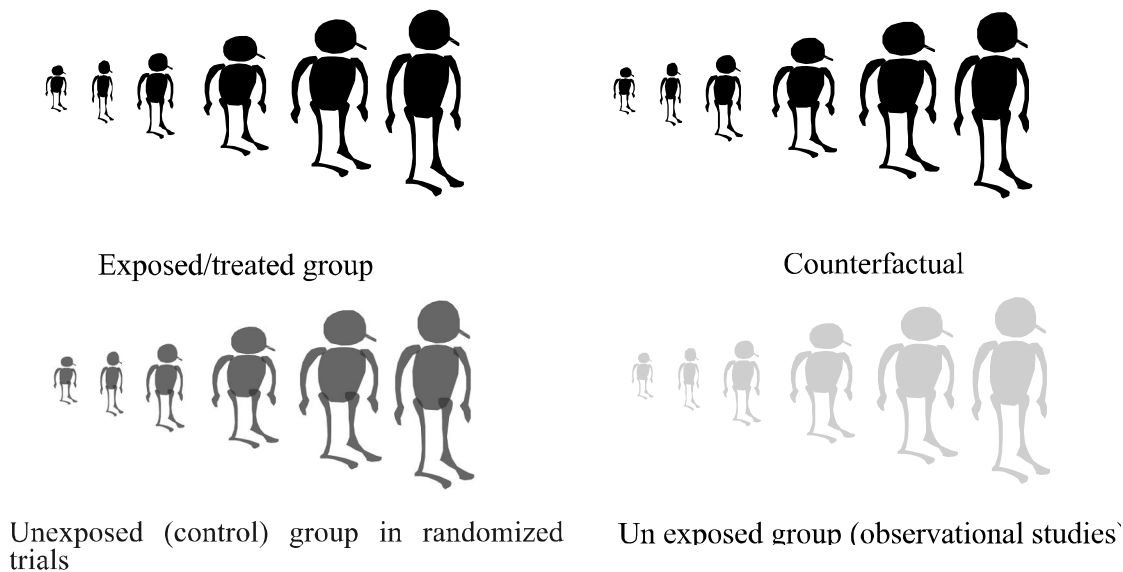


Fig 7: Counterfactual Vs. observed unexposed group in causal inference

In case of randomized controlled trials (RCTs), through randomization we can get the unexposed/control group as a close substitute to counterfactual (Fig 7). Whereas in case of most of the observational studies, as there is no randomization, the difference between the treated and non-treated groups may not be purely due to the treatment effect, rather it can be a mixed effect of treatment and the confounders. Until we condition the confounders, the estimated counterfactual could not be a close substitute to the counterfactual and the estimates of the treatment effect may be biased.

Techniques for estimating causal effect

Several techniques can be applied to get unbiased estimates of treatment effect. As mentioned earlier, the best way to estimate the causal effect is doing controlled experiments. In such experiments, the control group provides an estimate of the counterfactual. However, such experiments are costly and not always feasible. Hence we need to opt other methods for estimating treatment effect in case of observational studies where one of the major caveat is related to the estimation/prediction of the counterfactual when there are confounders. Different statistical techniques used to overcome the confounding bias in causal effect estimation are summarized in Table 3 (Imbens and Rubin, 2015 and Varian, 2016). These techniques are used in the cases where the confounders are high dimensional.

Table 3: Techniques for estimating causal effect in observational studies

S. No.	Technique	Description
1.	Propensity score (PS)	<ul style="list-style-type: none"> Indicates the estimates of probability of treatment as a function of observed characteristics. Estimated PS is used as if it were the only confounding covariate for causal effect estimation through matching. Estimation of PS through logit/ probit regression. Matching/ regression adjustment/ weighting based on PS to estimate the average treatment effect
2.	Instrumental variable (IV)	<ul style="list-style-type: none"> IV can be considered as a proxy for exposure or treatment variable which associated with exposure, but has no direct effect on outcome. Using IV can give more robust estimates when there is unobserved confounder problem. Two stage least square to estimate the treatment effect where the first stage regresses the treatment on the IV along with other covariates and in the second stage the outcome is regressed using the predicted treatment from the first stage regression.
3.	Regression discontinuity	<ul style="list-style-type: none"> When a treatment/ intervention is done based on some threshold Causal effect is estimated by comparing outcomes for experimental units on each side of the threshold
4.	Difference in difference	<ul style="list-style-type: none"> Used when we are measuring the outcomes of the treated and non-treated groups at two time periods (before and after treatment)

REFERENCES

- Pearl, J. (1995), Causal Diagrams for Empirical Research. *Biometrika*, 82: 669-710
- Imbens, G. and D. L. Rubin (2015), Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction. Cambridge University Press, New York.
- Varian, H. R. (2016), Causal Inference in Economics and Marketing. *Proceedings of the National Academy of Sciences*, 113 (27):7310-7315.

Chapter 19

PROPENSITY SCORE MATCHING

K. S. Aditya and Subash S. P.

INTRODUCTION

In observational studies where, the researcher has no control over the variables unlike experiments, it is difficult to make causal claims due to lack of suitable counterfactual outcomes. For a credible impact assessment, the treatment group and control group should be similar with respect to pre-treatment covariates. However, in social science research involving impact assessment, this condition is rarely met due to non-random assignment of treatment. Owing to non-random assignment, the treatment and control group mostly differ from each other and hence constructing counterfactual outcome is difficult.

Matching techniques aim at comparing treatment and control units, which are similar with respect to some observable characters. Matching is easy and straight-forward when the dimensionality is small, i.e., matching with respect to one or two characteristics. When number of variables with respect to which matching is to be done increases, as in many cases of its application, it becomes difficult to decide on which dimensions to match and this is termed as ‘curse of dimensionality’. Propensity score matching (PSM) provide a natural weighting scheme that overcome the problem and yields reliable estimator of treatment effect.

In PSM, we estimate the probability of a unit being in the treatment group based on all relevant observable characteristics, which is called as propensity score. Variables which affect either program participation or the outcome must be included in the estimation of propensity scores. Participants and non-participants are then matched based on the propensity scores after satisfying the assumptions that two units having similar propensity scores are also similar with respect to variables used to estimate the pscore. The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups. The validity of PSM depends on two conditions: (a) conditional independence (namely, that unobserved factors do not affect participation) and (b) sizable common support or overlap in propensity scores across the participant and non-participant samples.

To make the point clear, let us consider an example. Suppose we want to measure the impact of adoption of improved varieties of wheat. Theoretically, we know that adopters of improved varieties are usually well educated, cosmopolitan and belong to

higher social strata. Yield of such farmers will be higher than the non-adopters, even in absence of the improved variety. Hence, comparing the outcomes of adopters and non-adopters results in selection bias. In other words, we do not have a proper counterfactual outcome to measure the impact of new variety. One approach could be to obtain a composite score of probability of adoption (propensity score) conditional upon set of observable characteristics for both adopters and non-adopters and matching adopters and non-adopters on propensity score to create a comparable, artificial counterfactual. Once we have a counterfactual, we can proceed to measure the average difference in outcome across two groups.

PSM conditions

Conditional independence

Conditional independence states that given a set of observable covariates ‘X’ that are not affected by treatment, potential outcomes Y are independent of treatment assignment T. If unobserved characteristics determine program participation, conditional independence will be violated, and PSM is not an appropriate method. Unfortunately, there is no straight forward way to test this assumption. As a robustness check, sensitivity analysis suggested by Rosenbaum (also called as R- Bounds) can be performed to see how sensitive the estimates are in presence of bias due to unobserved variables.

Common support

A second condition is assumption of the common support or overlap condition. This condition ensures that treatment observations have comparable observations “nearby” in the propensity score distribution. Specifically, the effectiveness of PSM also depends on having a large number of participant and nonparticipant observations so that a substantial region of common support can be found (Fig 1 and 2).

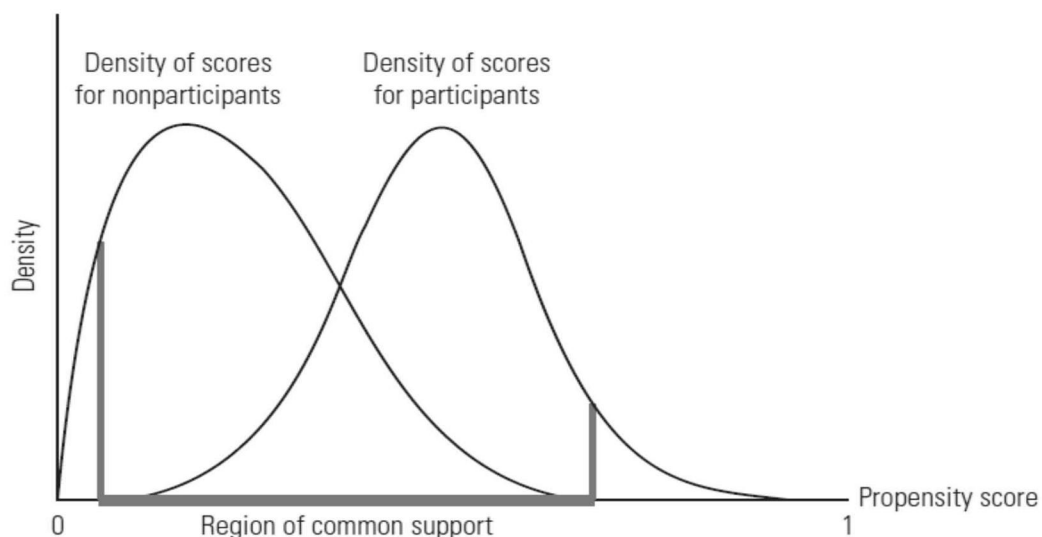


Fig 1: Example of common support

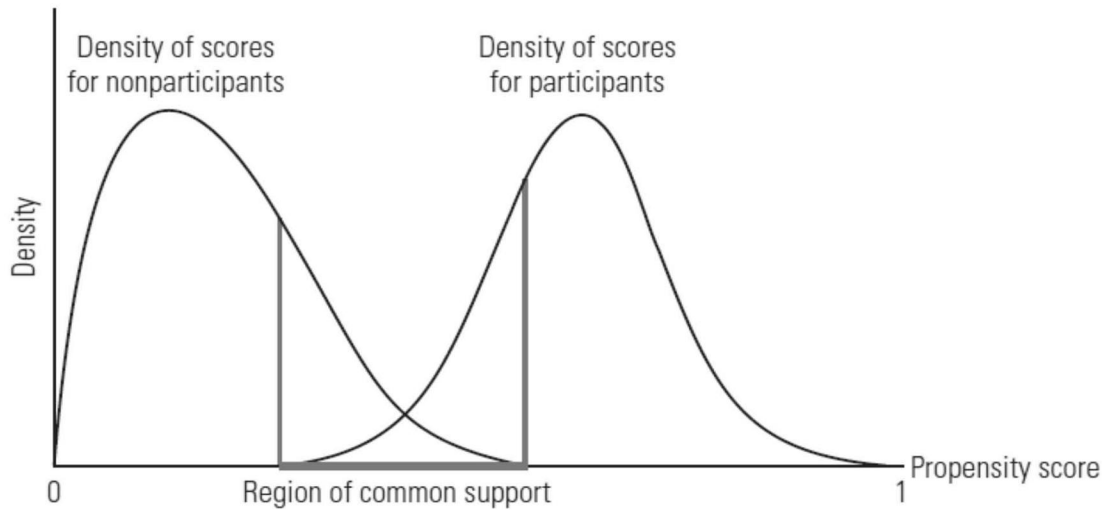


Fig 2: Weak common support

Source: Khandker *et al.* (2010)

STEPS IS PROPENSITY SCORE MATCHING

Step 1. Estimating propensity score; Modeling the program participation

To calculate the program treatment effect/ impact, one must first calculate the propensity score $P(X)$ on the basis all observed covariates X that jointly affect participation and outcome of interest. The aim of matching is to find the closest comparison group from a sample of non-participants to the sample of program participants. First, the samples of participants and non-participants are pooled, and then participation ‘ T ’ is to be estimated observed covariates ‘ X ’ in the data that are likely to determine participation or outcome of interest.

When the purpose is to compare the outcomes for those participating ($T = 1$) with those not participating ($T = 0$), this estimate can be constructed from a probit or logit model of program participation. Caliendo and Kopeinig (2008) also provide examples of estimations of the participation equation with a non-binary treatment variable, based on work by Brysonet *et al.*, 2002), Imbens (2000) and Lechner and Michel (2001). In this situation, one can use a multinomial probit (which is computationally intensive but based on weaker assumptions than the multinomial logit) or a series of binomial models.

Selection of variables is the most crucial thing and great care must be exercised. All those variables, which can influence the assignment of treatment or influence the outcome should be included in the model. As for the relevant covariates X , PSM will be biased if covariates that determine participation are not included in the participation equation. After the participation equation is estimated, the propensity scores can be estimated. Every sampled participant and non-participant will have an estimated propensity score. Note that the participation equation is not a deterministic model, and hence usual model evaluation criteria’s like adjusted R square and significance of individual factors does not matter much.

Step 2: Defining the Region of Common Support and Balancing Tests

Next, the region of common support needs to be defined where distributions of the propensity score for treatment and comparison group overlap. The observations for which there are no close observations with respect to propensity scores (out of common support) are pruned from the dataset. Once the common support assumption is satisfied, balancing property of the propensity score needs to be tested. The observations are arranged into different strata based on propensity scores and within each stratum, the observations in treated and control groups must have similar values for variables used to estimate propensity score (pscore). If the average value of the variables within a pscore strata are different across treatment and control group, then propensity scores cannot be used for matching. If the balancing property is not satisfied, we have to try different model specification. However, there is no specific guidelines on steps to achieve the balanced propensity scores. Researchers have to try different model specifications and re-estimate the propensity scores. This is also one of the limitation of the PSM that there are no clear guidelines on how to address the problem of ‘not balanced propensity scores’.

Step 3: Matching

Different matching criteria can be used to assign participants to non-participants on the basis of the propensity score. As discussed here, the choice of a particular matching technique may therefore affect the resulting program estimate through the weights assigned. Often, researchers report results from different matching estimators as a robustness check.

1. *Nearest-neighbour matching*: One of the most frequently used matching techniques is NN matching, where each treatment unit is matched to the comparison unit with the closest propensity score. Within NN matching, there are different types such as NN1, NN3 and NN%, where the number represents how many matches are used for each unit. Matching can be done with or without replacement. One important thing to remember here is that the bootstrapped standard errors cannot be used while using NN matching (Abadie and Imbens, 2008).
2. *Caliper matching/Radius matching*: One commonly reported problem with the NN matching is that the nearest neighbor still can be at a distance in terms of pscore. This results in poor matching and biased estimates. This can be avoided by imposing the maximum propensity score distance (caliper) for matching. Austin (2011) recommends that a caliper of 0.2 standard deviation of propensity score is ideal. However, one must exercise great caution while using calipers due to a problem termed as “PSM paradox”, where using narrow calipers leads to increase in bias (King and Nielsen, 2016)
3. *Stratification/interval matching*: This matching method divides the data into different strata within the region of common support. Further, within each stratum, the program effect is measured as mean difference in outcomes

between treated and control observations weighted by share of participants to non-participants.

4. *Kernel matching/local linear matching:* The major drawback with the methods explained so far is that there are possibilities that too few observations from the non-participants might qualify the imposed criteria. As an alternative, non-parametric matching estimators such as kernel matching and LLM use a weighted average of all non-participants to construct the counterfactual match for each participant.

Advantages and disadvantages

The main advantage (and drawback) of PSM relies on the degree to which observed characteristics drive program participation. If selection bias from unobserved characteristics is likely to be negligible, then PSM may provide a good comparison with estimates from the completely randomized experiments. To the degree participation variables are incomplete, the PSM results can be suspect. However, this particular assumption cannot be directly tested. Another advantage of PSM is that it does not necessarily require a baseline or panel survey, although in the resulting cross-section, the observed covariates entering the logit model for the propensity score would have to satisfy the conditional independence assumption.

One major criticism against PSM is that when the observations are pruned from the data due to lack of common support, the bias could increase with respect to one or two variables resulting in increase in bias. See King and Neilson (2016) for a detailed discussion on this.

Best practices in using PSM for measuring impact

- Use PSM only if you have large sample size (King and Nielson, 2016)
- Ensure that ‘Common support assumption’ is satisfied
- All the relevant covariates/ controls used (only those variables which influences program participation/ value of outcome variable must be used in the analysis)
- Ensure that balancing property is satisfied
- Try different methods of matching (nearest neighbor, caliper etc.)
- Perform a sensitivity analyses for the estimates for hidden bias. One of the major criticisms of PSM is that the treatment participation not only depends on observables but also on variables that are not observed/ measured. We measure impact using PSM based on the assumption that the observations having similar pscore have similar probabilities of being in treated group. However, if due to hidden bias, a particular unit has ‘delta’ times higher probability in treated group, what will happen to our estimates of impact? This can be tested using Rosenbaum bounds. This sensitivity analysis will indicate at what level of ‘delta’ the estimates of impact cease to be unbiased

- ‘Rosenbaum Bounds’ can be used for sensitivity analysis
- Diagnose data imbalance before and after matching (one case use multivariate L1 distance suggested by King and Neilsen, 2016)
- When using nearest neighbor matching, the bootstrapped standard errors are not valid. It is suggested that analytical standard errors be used for checking the significance of the estimates.

Couple of research papers using PSM

Aditya, K. S., Khan T and A. Kishore (2018), Adoption of crop insurance and impact: insights from India. *Agricultural Economics Research Review*, 31(347-2019-565): 163-174.

Abebaw, Degnet, and Mekbib G. Haile (2013), The Impact of Cooperatives on Agricultural Technology Adoption: Empirical Evidence from Ethiopia. *Food Policy*, 38(1): 82–91. <http://dx.doi.org/10.1016/j.foodpol.2012.10.003>.

Fischer, Elisabeth, and Martin Qaim (2012), Linking Smallholders to Markets: Determinants and Impacts of Farmer Collective Action in Kenya. *World Development*, 40(6): 1255–68. <http://dx.doi.org/10.1016/j.worlddev.2011.11.018>.

Hellin, Jon (2012), Agricultural Extension, Collective Action and Innovation Systems: Lessons on Network Brokering from Peru and Mexico. *The Journal of Agricultural Education and Extension*, 18(2): 141–59.

Mendola, Mariapia (2007), Agricultural Technology Adoption and Poverty Reduction: A Propensity-Score Matching Analysis for Rural Bangladesh. *Food Policy*, 32(3): 372–93.

Shiferaw, Bekele A., Tewodros A. Kebede and Liang You (2008), Technology Adoption under Seed Access Constraints and the Economic Impacts of Improved Pigeon Pea Varieties in Tanzania. *Agricultural Economics*, 39: 309–23. <http://doi.wiley.com/10.1111/j.1574-0862.2008.00335.x> (September 4, 2013).

Wollni, Meike, and Manfred Zeller (2007), Do Farmers Benefit from Participating in Specialty Markets and Cooperatives? The Case of Coffee Marketing in Costa Rica. *Agricultural Economics*, 37(2–3): 243–48.

ILLUSTRATION

PSM in stata using pscore and psmatch2

We will use “pscore”, and “psmatch2” user written packages to implement PSM in stata. There are other advanced packages like nnmatch2. In this document we would be discussing how to do PSM using pscore and psmatch2. For the practical, we will be using the hh_98.dta provided together with Khander *et al.* (2010) for analysis (read

about the study from the book). The command for estimating psm is given in box below. This could be copied and pasted in command window or used in a do file.

Box 1

```
*PSM analysis of Bangladesh microfinance data
*Log transformation of land area data
gen lnland =ln(1+hhland/100)
*PSM command
pscore dmmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg
[pw=weight], pscore(ps98) blockid(blockf1) comsup level(0.001).
```

Note: Read more about the structure of the command using **help pscore**.

The results include probit regression output (shown below), the estimation and description of the propensity scores, the number of blocks and stratification using propensity scores, and the balancing property test. The area of common support is those propensity scores within the range of the lowest and highest estimated values for households in the treatment group.

The following output shows that the identified region of common support is [.00180123, .50022341], the final number of blocks is 4, and the balancing property is not satisfied. The most important element to look for in the output is the list of variables that cause the balancing property not to be satisfied. The output shows the “egg” variable is not balanced in block 2. The solution to this problem is to use a different set of covariates and rerun the “pscore” command.

After a few iterations, you will find that dropping “egg” and “lnland” allows the “pscore” command to be rerun with the balancing property satisfied. So “pscore” on “dfmfd” is run again, this time excluding the “egg” and “lnland” variables. Before rerunning the “pscore” command, it is important to drop the “ps98” and “blockf1” variables that were created as a result of the earlier run. Because female program participation is of more interest, the “pscore” command is shown here with female participation only.

Results from box 1 are given below

```
*****
Algorithm to estimate the propensity score
*****
```

The treatment is dmmfd

```
HH has male |
microcredit |
participant |
: 1-Y, 0-N |      Freq.      Percent      Cum.
-----+-----
          0 |          909          80.51          80.51
          1 |          220          19.49         100.00
-----+-----
        Total |         1,129         100.00
```

Estimation of the propensity score

```
(sum of wgt is 1.1260e+03)
Iteration 0: log pseudolikelihood = -424.61883
Iteration 1: log pseudolikelihood = -390.85321
Iteration 2: log pseudolikelihood = -389.10243
Iteration 3: log pseudolikelihood = -389.05511
Iteration 4: log pseudolikelihood = -389.05501
```

```
Probit estimates                                Number of obs   =    1129
                                                Wald chi2(11)   =    64.36
                                                Prob > chi2     =    0.0000
Log pseudolikelihood = -389.05501              Pseudo R2      =    0.0838
```

```
-----+-----
          |      Robust
dmmfd    |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
sexhead  |   .915108    .2432905    3.76  0.000   .4382675   1.391949
agehead  |  -.0036952   .0046186   -0.80  0.424  -.0127475   .005357
educhead |   .0161662   .0170125    0.95  0.342  -.0171777   .04951
lnland   |  -.3341691   .1113146   -3.00  0.003  -.5523417  -.1159965
vaccess  |  -.0752904   .1770457   -0.43  0.671  -.4222935   .2717128
pcirr    |   .2088394   .1753383    1.19  0.234  -.1348174   .5524961
rice     |   .145771    .0384417    3.79  0.000   .0704268   .2211153
wheat    |   .0465751   .0648087    0.72  0.472  -.0804475   .1735977
milk     |  -.0017358   .023861    -0.07  0.942  -.0485026   .045031
oil      |  -.0249797   .0135856   -1.84  0.066  -.051607   .0016476
egg      |  -.7687454   .2311995   -3.33  0.001  -1.221888  -.3156028
_cons    | -1.188481    .8358266   -1.42  0.155  -2.826671   .4497088
-----+-----
```

Note: the common support option has been selected
The region of common support is [.00180123, .50022341]

Description of the estimated propensity score
in region of common support

Estimated propensity score				

	Percentiles	Smallest		
1%	.0055359	.0018012		
5%	.0170022	.0020871		
10%	.0346036	.0026732	Obs	1127
25%	.069733	.0028227	Sum of Wgt.	1127
50%	.1206795		Mean	.1339801
		Largest	Std. Dev.	.0850809
75%	.1811405	.4698302		
90%	.2527064	.472444	Variance	.0072388
95%	.2965199	.4735467	Skewness	.8931864
99%	.3903884	.5002234	Kurtosis	3.942122

Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output

The final number of blocks is 4

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks

Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output

Variable egg is not balanced in block 2

The balancing property is not satisfied

Try a different specification of the propensity score

	HH has male		
Inferior	microcredit		
of block	participant: 1=Y, 0=N		
of pscore	0	1	Total

0	380	49	429
.1	382	97	479
.2	140	70	210
.4	5	4	9

Total	907	220	1,127

Note: the common support option has been selected

End of the algorithm to estimate the pscore

With the propensity scores generated, the outcomes of interest (such as total per capita expenditure) between the treatment group and the matched control group are now compared to see whether the microcredit programs affect the outcome of interest. The following sections estimate the treatment effect of microcredit program participation, using different matching techniques that are available.

Results from matching

For estimating impact, `pscore` or `psmatch2` or `nnmatch2` commands is used.

Box 3

Psmatch2 dependant_variable indepnet_variables, **outcome**(outcome_variable)

For our case, the command will be

```
psmatch2 dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil,
outcome(lexptot)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
lexptot	Unmatched	8.41274124	8.45984398	-.047102735	.038602113	-1.22
	ATT	8.41274124	8.42471704	-.011975798	.047629103	-0.25

In this case, -0.11 is the Average Treatment Effect on Treated (ATT. ATT in general explains the impact of participation of male members on the expenditure of the households which are in treated group. On the other hand, Average Treatment Effect (ATE) estimates impact on the entire sample (both treated and control). So, we can say, male SHG participation has no impact on household expenditure. However, in our case, the ATE estimates are statistically not significant. Hence we conclude that the participation of member from household has no impact on the expenditure of participating households.

REFERENCES

- Abadie, A. and G. W. Imbens (2008), On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6): 1537-1557.
- Austin, P. C. (2011), An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioural Research*, 46(3): 399-424.
- Bryson, A., D. Richard, and S. Purdon (2002), The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies. Working Paper 4, Department for Work and Pensions, London.
- Caliendo, M. and Sabine Kopeinig (2008), Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22 (1): 31-72.

- Imbens, G. (2000), The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*, 87 (3): 706–10.
- Khandker, R., B. Shahidur, G. Koolwal, A. Hussain and Samad (2010), Handbook on impact evaluation: quantitative methods and practices. The International Bank for Reconstruction and Development / The World Bank. Washington DC.
- King, G. and R. Nielsen (2016), Why propensity scores should not be used for matching. *Political Analysis*, 1-20.
- Lechner and Michael (2001), Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In *Econometric Evaluation of Labor Market Policies*, ed. Michael Lechner and Friedhelm Pfeiffer, 43–58. Heidelberg and New York: Physica-Verlag.

Chapter 20

DIFFERENCE-IN-DIFFERENCE MODEL

M. Balasubramanian and Gourav Kumar Vani

INTRODUCTION

The Difference-in-Difference (DiD) approach is a research design for estimating causal effects. It is a popular choice for those assessing the impact of policy interventions when panel data is available. Its utility is also enhanced when suitable instrumental variables are not found. It is also useful when there is need for controlling confounding variables in quasi experimental setting. The DiD design is usually based on comparing *de facto* two different groups of objects, pre- and post-implementation of programme/intervention. 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, then constant difference in trend of outcome variable between treatment and control prior to programme implementation is added to the outcome of the control group to find the counterfactual needed to assess the impact of the programme.

The first scientific study based on DiD approach was conducted by Snow (1854). Snow was trying to answer the question “whether cholera was transmitted by either (bad) air or (bad) water”. He used quasi experimental setting where water supply to one district of London changed from polluted water taken from the Thames River in Central London to a supply of cleaner water from upriver. Rose (1952) investigated the impact of a regime of ‘mandatory mediation’ on work stoppages by a DiD design. The research question involved was that “whether mediation mandatory was effective in preventing work stoppages”. For this purpose, the study made comparison between (1) states having law and states not having law; (2) states before and after the law is put into operation. The first comparison was achieved by taking percentages of each of the three states to the total United States, for the measures used. The second comparison was 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 on employment levels in retail industry for the state of Oregon (U.S.A.). The comparison was made between for a particular group of employees in Portland, the largest city, with the rest of the state. Lester (1946) was concerned with the effects of wages on employment. He conducted a survey of firms that had operations in both the northern and the southern states of the United States. His idea was to compare employment levels of groups of firms with low average wages to groups of firms with higher wage levels before and after various minimum wage rises. There was mild effect on latter’s wage bills, possibly through increase in the minimum wage. The early applications of this technique highlighted that it does not require high computational powers for estimation of programme effects, as long as covariates are not required. DiD as a

research design had been used in several context to address important policy issues, for example, work of Card and Krueger (1994) on the effects of minimum wages on employment, Ashenfelter and Card (1985), Heckman and Hotz (1989) and Blundell *et al.* (2004) on impact of training and other labour market interventions on labour market outcomes for unemployed, work of Card (1990) on the impact of immigration on the local labor market.

ILLUSTRATION WITH EXAMPLE

Farm household samples under Krishna river basin of Nagarjuna Sagar Project (NSP) were purposively selected for collecting the data from the selected mandals in order 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. The total sample size for the study was 178, out of which 138 are adopters and 40 are non-adopters.

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 Fig 1, where curve AA refers to the traditional irrigation technology production function, curve AB refers to the improved technology production function and curve AC refers to the improved technology with capacity building. With X units of water, traditional technology produces Y_1 units of output, water saving interventions (DSR, AWD, MSRI) produces Y_2 units of output whereas improved technology with capacity building produces Y_3 units of output. The difference between Y_2 and Y_3 is the additional output due to capacity building/trainings (Palanisami *et al.*, 2014).

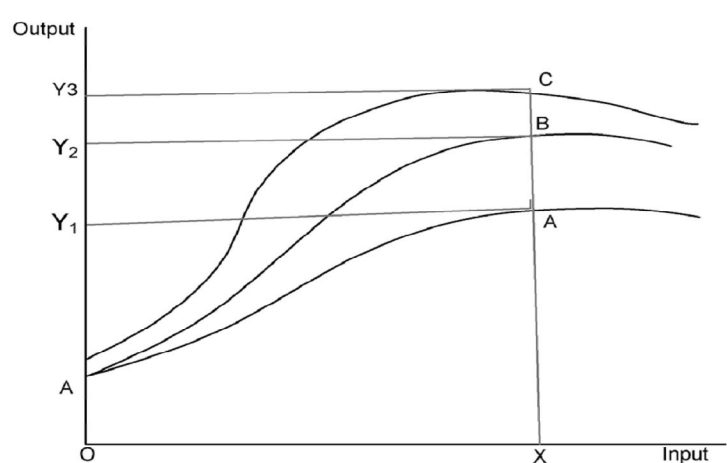


Fig 1: Technology adoption and crop yield

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 discounted capital budgeting measures like the benefit-cost (B-C) ratio and internal rate of return (IRR). However, these measures had failed to account for all associated factors. These discounted measures can assess the economic viability of the proposed technology either *ex-ante* or *ex-post*. There is no way to account for the impact of the technology by showing that difference in *ex-ante* and *ex-post* measurement of these discounted measures is purely because of technological intervention. This is on account of many variables which change during the adoption of technology or implementation of programme. These variables are called as time variant variables.

Some bias in measuring impact can also occur because of time invariant individual effects, for instance, when large proportion of the adopters of technology were the individuals with better social networks which enabled them in adoption of technology, while many of the individuals with poor network could not gain information required to adopt the same. Similarly, gender, caste and managerial capability can also be the time invariant factors which do not change with individual before and after adoption of the technology. The Difference-in-Difference (DiD) approach accounts for very well for time invariant as well as time variant variables and thus overcomes the shortcomings associated with discounted measures. DiD has both components, with and without as well as pre and post. This enables accounting for time fixed individual effects (which are taken care of in with and without approach) and time variant effects (which are taken care of in pre and post approach). The impact of the project is then estimated using the interaction of time variant and time invariant approach. Hence, this a combination of both with and without and before and after approach i.e. double 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 have participated 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 Fig 2.

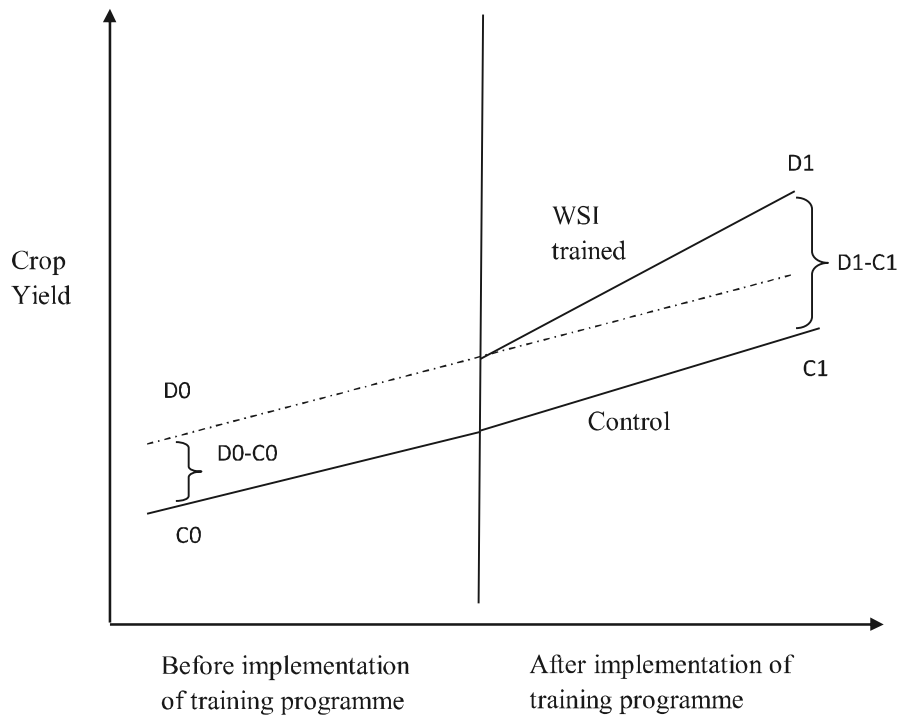


Fig 2: Illustration of Programme effect in DiD approach

Double Difference

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

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

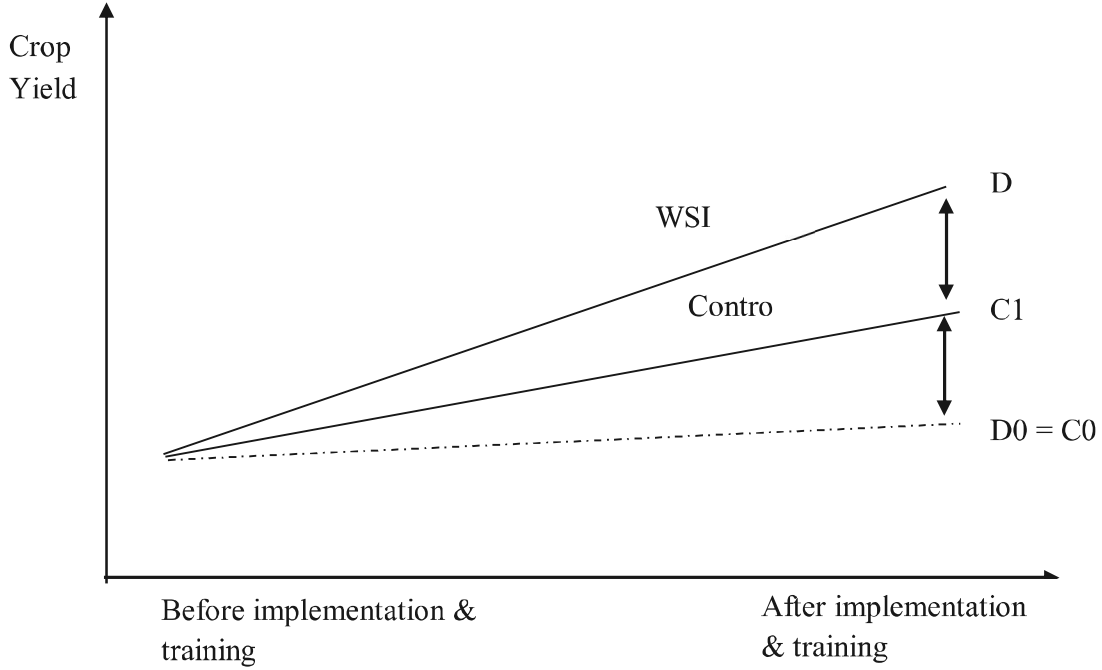


Fig 3: Illustration of impact of capacity building program by double difference method.

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

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

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

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

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

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

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

Observation belongs to	Δ	T	y_i
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)

$$\text{Impact of the program} = ((a+b+c+d) - (a+b)) - ((a+c)-a) = d \dots \quad (6)$$

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

S. 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 (qtl/ha) 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. 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

S. 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. MSRI adoption has also significant impact at 1 per cent level in the interaction of capacity building program and technology transfer.

The R² 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.

REFERENCES

Ashenfelter, O. and D. Card (1985), Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67: 648–660.

- Blundell, R., C. Meghir, M. Costa Dias, and J. van Reenen (2004), Evaluating the employment impact of a mandatory job search program'. *Journal of the European Economic Association*, 2:569–606.
- Card, D. (1990), The impact of the mariel boatlift on the miami labor market. *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.
- Heckman, J. J. (1996), Comment on eissa: Labor supply and the economic recovery act of 1981. In: M. Feldstein and J. Poterba (eds.): *Empirical Foundations of Household Taxation*. 5–38.
- Heckman, J. J. and V. J. Hotz (1989), Choosing among alternative non experimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, 84: 862–880.
- Lester, R. A. (1946), Shortcomings of marginal analysis for the wage employment problems'. *American Economic Review*, 36: 63–82.
- Obenauer, M. and B. von der Nienburg (1915), Effect of minimum wage determinations in oregon. *Bulletin of the U.S. Bureau of Labor Statistics*, 176, Washington, D.C. U.S. Government Printing Office.
- Palanisami, K., C.R.Ranganathan, D. Sureshkumar and R.P.S. Malik (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*, 5 (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, *Journal of the American Statistical Association*, 96: 210-219.
- Snow, J. (1854), The cholera near golden square, and at deptford. *Medical Times and Gazette*, 9: 321–322.
- Wing, C., K. Simon and R. A. Bello-Gomez (2018), Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39.

Chapter 21

REGRESSION DISCONTINUITY DESIGN

Subash S. P. and Aditya K. S.

INTRODUCTION

In observational studies, the lack of counterfactual outcome is the major concern for a credible impact assessment. Non-random allocation of the program makes it difficult to find suitable counterfactual for the study. Quasi experimental methods like propensity score matching aims to artificially construct a counterfactual group and then estimate the counterfactual outcome to estimate the impact. However, in some cases, discontinuity arising out of an eligibility criteria/ other exogenous factor can be used to estimate the counterfactual outcome and impact. The basic intuition behind using discontinuity is that, units just above and below a threshold will be similar to each other, except for the treatment and hence, are good counterfactuals. This method of exploiting the discontinuity to estimate impact is similar to usage of Instrumental Variable (IV).

Regression Discontinuity Design (RDD)

RDD is a quasi-experimental method for assessing impact. Suppose the eligibility for a particular problem is decided by the size of land holding (hypothetically 7 ha), such that all the farmers above 7 ha of land are eligible to be the beneficiaries of the program. Estimating causal relationship is difficult due to pretreatment difference between beneficiaries and non-beneficiaries, with respect to land and other covariates. However, RDD method assumes that the units just above and below the cut off (7ha) of the forcing variable (land) are similar to each other and are good counterfactuals. Based on the assumption that the units around the cutoff point are homogenous, except for treatment, the RDD method estimates impact. The estimated impact assessed using this approach is Local Average Treatment Effect (LATE) (Compared to Average Treatment Effects estimated using other approaches).

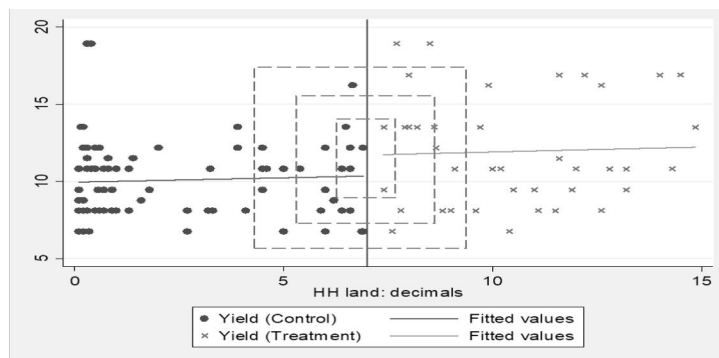


Fig 1: Regression discontinuity experiment with a treatment (intervention)

LATE of the intervention is estimated by comparing the beneficiary's outcome just below the cut off, with that of 'potential beneficiaries' just above the cut off. RDD would be suitable in a context where there are many observations around the cut-off and would result in a narrower bandwidth for comparison. We will discuss various types of RDD and steps in estimating RDD.

Types of RDD

Based on how strictly the cut-off point of the forcing variable divides the beneficiaries and non-beneficiaries, there are two methods of estimating RDD.

1. *Sharp RDD*: This approach can be used when the forcing variable determines the status of treatment allocation with 100% success. For example, consider a situation where the eligibility condition is such that units with forcing variable less than the cut off are eligible to be beneficiaries of the program. In this case, if all the units with value of forcing variable more than cut off are non-beneficiaries and units to the left of cutoff are beneficiaries, we can use the Sharp RDD. In this case, treatment status is a deterministic and discontinuous function of the forcing variable (with discontinuity at the cutoff point).
2. *Fuzzy RDD*: In some cases, the forcing variable cannot precisely determine the treatment allocation status. Continuing with the last example, some units with forcing variable more than cut off might also end up receiving the treatment and vice versa. Further, the fuzzy RDD can be divided into categories; Type 1 fuzzy and Type 2 fuzzy. Type 1 fuzzy is a situation when some units which were supposed to get treatment based on eligibility condition, do not get the treatment and vice versa. Type 2 Fuzzy is when some units which were not supposed to be beneficiaries of the program end up being beneficiaries. Standard non-parametric regression can be used to estimate the treatment effect of interest, in case of both sharp and fuzzy RDD.

Advantages and disadvantages

The advantage of RDD is that most projects/ programs have some kind of eligibility/ exclusion criteria, which can be used to model the discontinuity and estimate the impact by minimizing bias. Another attractive feature of RDD is that its graphical representation of impact which helps in better comprehension. However, RDD is criticised mainly for its lack of external validity and for utilizing very less information from the available sample. As discussed earlier, the RDD uses information only from those households which are near to the cut off region of forcing variable, pruning out the rest of the units. This also limits the generalizability of the results.

Steps in applying RDD approach

Step 0: Graphical analysis

Before doing the RDD analysis we could graphically visualise the data to inspect discontinuity (Fig1). The steps involved are

1. Divide forcing variable into bins
2. Plot forcing variable and average outcome variable in each bin
3. Superimpose a flexible regression line
4. Inspect whether there is discontinuity at the threshold

We could use binplots package in stata software.

Step 1: Check whether the conditions given below are satisfied

Two conditions are to be met before employing RDD approach.

1. Continuous eligibility index- assignment variable /forcing variable
2. Clearly defined cut-off score

Step 2: Check for internal validity assumption

This includes two set of testing

1. Covariate balance: Test whether other covariates jump at the cut-off (Fig 2).
Check: Re-estimate the RD model with covariates as the dependent variable.

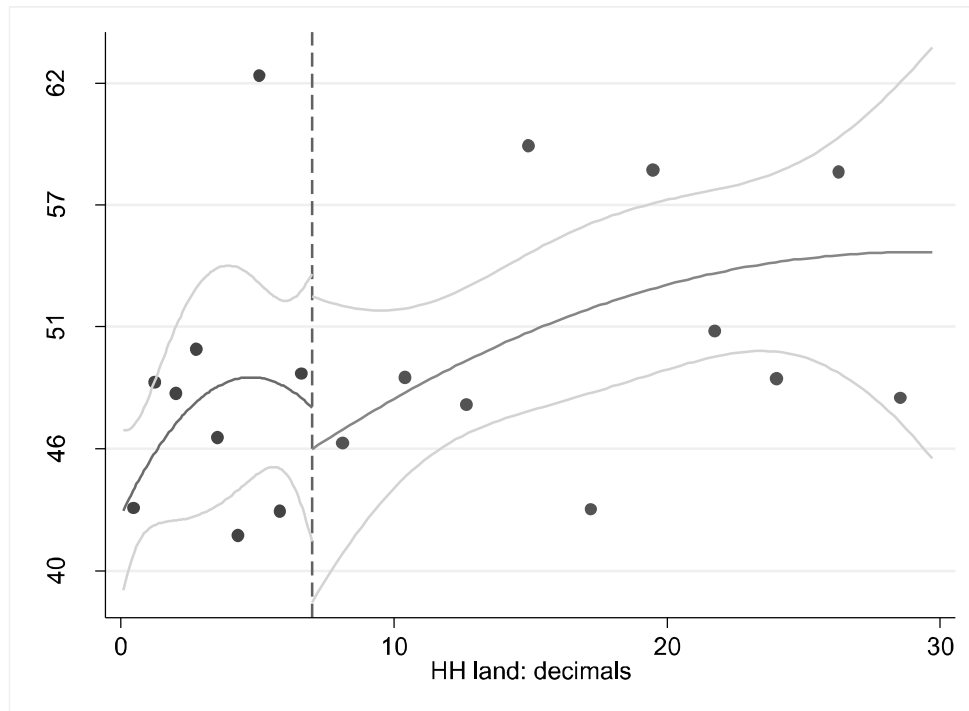


Fig 2: Checking covariates jump at the cut-off point

2. Continuous density of the forcing variable: In some cases, if the perceived benefit from the program is large, few units can manipulate the value of outcome variable and self-select into treatment group. In that case, distribution of forcing variable will have a discontinuity around the cut off. If there is a discontinuity in forcing variable itself, the foundation on which RDD is built will collapse.

We could use `kdenisty` command in stata to check for discontinuity. If there is a sudden shift in distribution of the forcing variable around the cut-off that would constitute discontinuity (Fig 3).

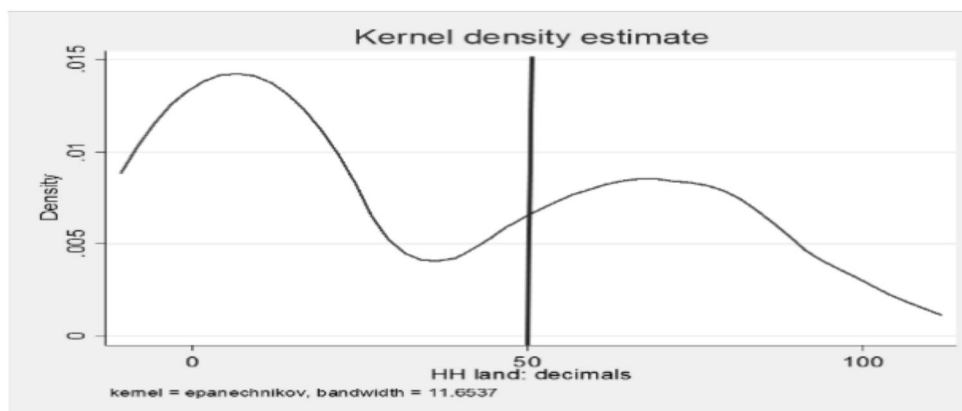


Fig 3: Continuous density of forcing variable at cut off point

Step 3: Bandwidth selection

Optimal bandwidth is estimated by different methods provided in literature Calonico *et al.* (2014), Imbens and Kalyanaraman (2012), Ludwig and Miller (2007). There is recent revision in this approach. `mserd` [mean squared error (MSE)] is new default option the optimal bandwidth selector using 'rdrobust' (Calonico *et al.*, 2016).

Step4: Analysis and interpretation

'**rdrobust**' package in stata software can be used to estimate the models as well as plotting the discontinuity in outcomes for ease of comprehension (Calonico *et al.*, 2015, 2016). Similarly we could also use 'rdd' package in R (<https://cran.r-project.org/web/packages/rdd/rdd.pdf>). In this chapter we discuss implementation of RDD using 'rdrobust' package in stata software.

Also see the article below for understanding of how the studies are done and inferences are drawn.

Asher, S. and P. Novosad (2019), Rural roads and local economics development. *American Economics review* (Forthcoming). Available online <https://www.dartmouth.edu/~novosad/asher-novosad-roads.pdf>.

Calonico, Sebastian, D. Matias, Cattaneo and Max H Farrell (2016), Rdrobust : software for regression discontinuity designs. *The Stata Journal*, (2): 1–30.

Lee, D. S. and T. Lemieux (2010), Regression discontinuity designs in economics. *Journal of Economic Literature*, 48: 281–355.

Sekhri and Sheetal (2014), Wells, Water, and welfare: the impact of access to groundwater on rural poverty and conflict. *American Economic Journal: Applied Economics* 6(3): 76–102.

ILLUSTRATION

Impact assessment of “Save the farmer” programme¹

“Save the farmer” is a project launched by Country XXX in 2017. The project was implemented by Ministry of Agriculture in XXX. The objective of the project was to eradicate poverty of farm households in the country. In the project, improved varietal seeds were supplied to farmers. The selection of beneficiary was based on landholding. Farm households which had less than 7 acres of lands were considered as beneficiary. Ministry of Agriculture did not plan for an impact evaluation while the programme was rolled out. In 2018, the Ministry decided to assess impact of the programme. At the end of 2017 a farm household level survey was carried out by the National Statistical Organization in XXX. Several farm and household level data were collected during the survey. The details of the variable collected during the survey is given in Table 1.

Table 1: Description of variables collected during the survey

S. No.	Variable name	Description	Type
1	HHID	Household ID (Unique)	
2	Agea	Age of household head in years	Continuous
3	Education	Number of years of schooling	Continuous
4	Family_size	Number of members in family	Continuous
5	Land_holding	Total land holding (Acres)	Continuous
6	Assets	Total value of asset in #	Continuous
7	Yield	Yield of the crop (tonnes/acre)	Continuous
8	Expenditure	Total household expenditure in #	Continuous
9	Treatment	1=Households which received treatment, 0=Otherwise	Binary

Note: Few additional variables are provided in the data file for explaining the theory. Assume there is only one crop. Expenditure is in money terms assumed as #.

Task: You are asked to conduct an impact assessment of the programme using farm household data collected by National Statistical Organization. Design the methodology based on the available secondary data and analyze the data. The data is provided in a Stata file.

Tips: You may need to create new variables for analysis (poverty_line). The poverty line in the county is # 3000.

Solution for Case study for RDD

The analysis given below is based on the training data provided in the class. Open the RDD training dataset in Stata.

¹ The case given below is a hypothetical case developed for classroom teaching. The data used in the case is a modified data from Khandker, Shahidur R., Gayatri B., Koolwal, Hussain A. and Samad (2010). “Handbook on impact evaluation: quantitative methods and practices”. The International Bank for Reconstruction and Development / The World Bank. Washington DC.

Install rdrobust package in stata

Start with descriptive

Summarize the treatment and outcome variables.

tab Treatment

```
tab Treatment
```

Treatment	Freq.	Percent	Cum.
0	150	61.73	61.73
1	93	38.27	100.00
Total	243	100.00	

sum Yield

sum Expenditure

gen poverty_final= 0

replace poverty_final= 1 if Expenditure<3000

sum poverty_final

or sum Yield Expenditure poverty_final

```
. sum Yield Expenditure poverty_final
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Yield	243	10.88538	3.29018	6.766816	20.30045
Expenditure	243	2177.878	3668.965	162.8096	43411.15
poverty_final	243	.81893	.3858709	0	1

Step 1: Check whether the conditions given below are satisfied

Two conditions are to be met before employing RDD approach.

1. Continuous eligibility index- assignment variable /forcing variable; the forcing variable should be a continuous variable. The assignment variable is Land_holding.
2. Clearly defined cut-off score

To check if there is a clearly defined cut-off score

sum Land_holding if Treatment = 1

sum Land_holding if Treatment = 0

Regression Discontinuity Design

```
. sum Land_holding if Treatment==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Land_holding	93	2.361618	2.546822	.1	6.9

```
. sum Land_holding if Treatment==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Land_holding	150	33.64473	45.14577	7	420.8

You could see that there is a clear cut-off. In such case we could use sharp RDD. If it's not the case for example

sum Land_holding if Treatment_fuzzy1==1

sum Land_holding if Treatment_fuzzy==0

```
. sum Land_holding if Treatment_fuzzy==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Land_holding	119	23.13418	33.37695	.1	255

```
. sum Land_holding if Treatment_fuzzy==1
```

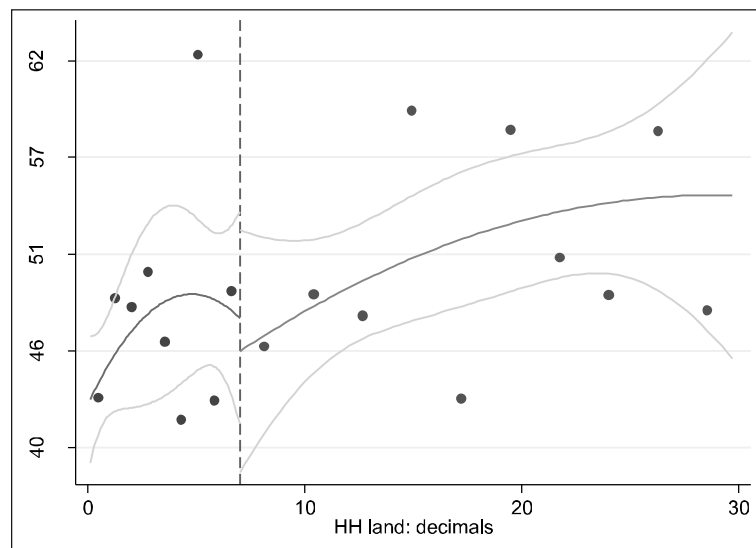
Variable	Obs	Mean	Std. Dev.	Min	Max
Land_holding	124	20.26913	43.10203	.1	420.8

In such cases we use Fuzzy RDD.

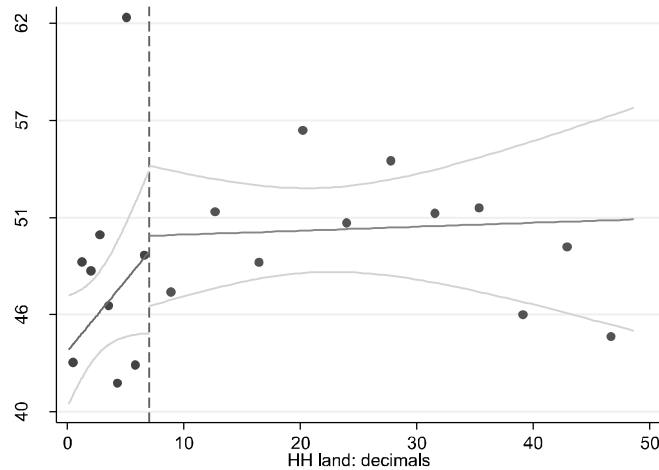
Step 2: Check for internal validity assumption

1. Covariate balance:

cmogram Age Land_holding if Land_holding<30, cut(7) scatter line(7) qftci



cmogram Age Land_holding if Land_holding<50, cut(7) scatter line(7) lfitci

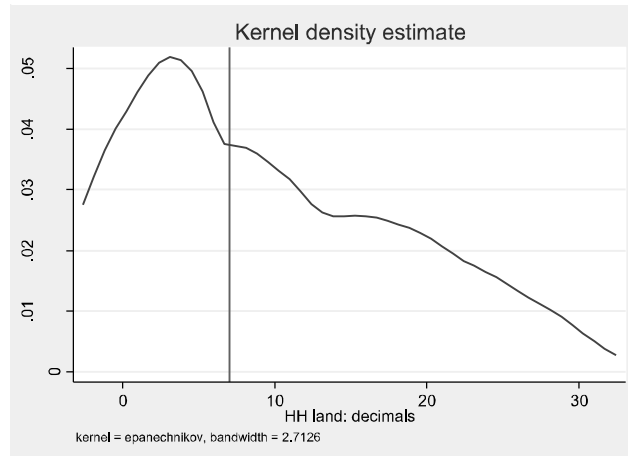


*You could do for all other covariates and see if there is a shift in the cut-off region. So the co-variate is included in the final model. Similarly check for other co-variables.

2. Continuous density of the forcing variable:

Plot a kdensity graph. There should not be any jump around the continuous variable.

kdensity Land_holding if Land_holding<30, xline(7)



From the plotted graph it seems the data is not ideal for RDD.

kdenisty Land_holding

Step 3: RDD analysis using rdrobust

The next step is to run the local linear regression for outcome (household per capita expenditure) against household's landholding for both eligible (participants) and ineligible (non-participants) households. Local polynomial regression allows estimated outcomes to be stored for both participants and non-participants. The next step is to take means of those outcomes at the cut-off point. Because the cut-off point is a single

value (7 acres), it is better to specify a range of landholding values and take means of outcomes for households that are within that range. Use the package `rdrobust` to run the programme.

rdrobust Yield Land_holding, c (7)

Sharp RD estimates using local polynomial regression.

Cutoff c = 7	Left of c	Right of c	Number of obs =	243
Number of obs	93	150	BW type	= mserd
Eff. Number of obs	35	30	Kernel	= Triangular
Order est. (p)	1	1	VCE method	= NN
Order bias (q)	2	2		
BW est. (h)	4.735	4.735		
BW bias (b)	6.570	6.570		
rho (h/b)	0.721	0.721		

Outcome: Yield. Running variable: Land_holding.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	3.674	1.6506	2.2259	0.026	.438966 6.9091
Robust	-	-	2.1134	0.035	.303852 8.06478

rdrobust Expenditure Land_holding, c (7)

Sharp RD estimates using local polynomial regression.

Cutoff c = 7	Left of c	Right of c	Number of obs =	243
Number of obs	93	150	BW type	= mserd
Eff. Number of obs	35	27	Kernel	= Triangular
Order est. (p)	1	1	VCE method	= NN
Order bias (q)	2	2		
BW est. (h)	4.498	4.498		
BW bias (b)	5.821	5.821		
rho (h/b)	0.773	0.773		

Outcome: Expenditure. Running variable: Land_holding.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	929.39	1605.5	0.5789	0.563	-2217.4 4076.17
Robust	-	-	0.2490	0.803	-2780.62 3589.95

rdrobust poverty_final Land_holding, c (7)

Sharp RD estimates using local polynomial regression.

Cutoff c = 7	Left of c	Right of c	Number of obs =	243
Number of obs	93	150	BW type	= mserd
Eff. Number of obs	35	27	Kernel	= Triangular
Order est. (p)	1	1	VCE method	= NN
Order bias (q)	2	2		
BW est. (h)	4.432	4.432		
BW bias (b)	7.220	7.220		
rho (h/b)	0.614	0.614		

Outcome: poverty_final. Running variable: Land_holding.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	-.15568	.17307	-0.8995	0.368	-.494897 .183542
Robust	-	-	-0.8635	0.388	-.550326 .213702

You could also visualise the RDD graph using the following commands

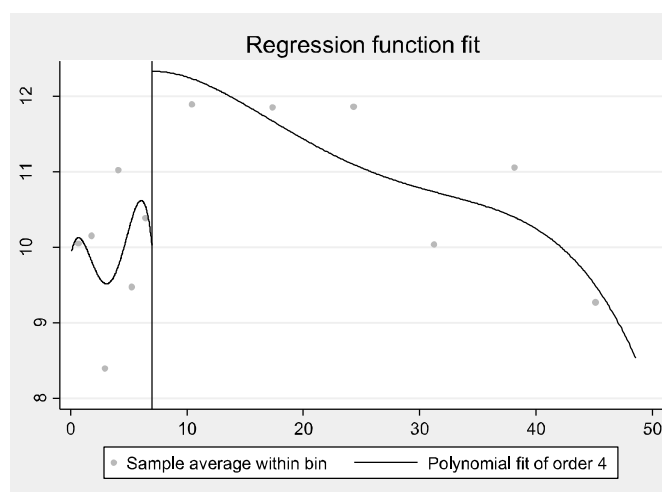
rdplot Yield Land_holding, c (7)

3D Plot with evenly spaced mimicking variance number of bins using spacings estimators.

Cutoff c = 7	Left of c	Right of c	Number of obs =	223
			Kernel =	Uniform
Number of obs	93	130		
Eff. Number of obs	93	130		
Order poly. fit (p)	4	4		
BW poly. fit (h)	6.900	41.600		
Number of bins scale	1.000	1.000		

Outcome: Yield. Running variable: Land_holding.

	Left of c	Right of c
Bins selected	6	6
Average bin length	1.150	6.933
Median bin length	1.150	6.933
IMSE-optimal bins	3	3
Mimicking Var. bins	6	6
rel. to IMSE-optimal:		
Implied scale	2.000	2.000
WIMSE var. weight	0.111	0.111
WIMSE bias weight	0.889	0.889



Similarly, we could use the same command for other outcome variables.

rdplot Expenditure Land_holding, c (7)

rdplot poverty_final Land_holding if Land_holding < 30, c (7)

The coefficients need to be interpreted in the reverse order. Negative sign indicate increase in mean value over control. The results² shows that the program participation has a positive effect on yield and expenditure but non-significant impact on expenditure.

² Covariates not included

For further details regarding the package use ‘help rdrobust’ command. Further the model could be improved controlling for covariates.

rdrrobust Yield Land_holding , c (7) covs(Age Education Assets)

rdrrobust Expenditure Land_holding, c (7) covs(Age Education Assets)

rdrrobust poverty_final Land_holding, c (7) covs(Age Education Assets)

Covariate-adjusted sharp RD estimates using local polynomial regression.

Cutoff c = 7	Left of c	Right of c	Number of obs =	223
			BW type =	mserd
			Kernel =	Triangular
			VCE method =	NN
Number of obs	93	130		
Eff. Number of obs	33	27		
Order est. (p)	1	1		
Order bias (q)	2	2		
BW est. (h)	4.251	4.251		
BW bias (b)	5.848	5.848		
rho (h/b)	0.727	0.727		

Outcome: Yield. Running variable: Land_holding.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	3.6635	1.6968	2.1591	0.031	.337885 6.98905
Robust	-	-	2.0195	0.043	.121807 8.143

Covariate-adjusted estimates. Additional covariates included: 3

REFERENCES

- Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2016), Regression discontinuity designs using covariates. Working Paper, University of Michigan. http://www-personal.umich.edu/~cattaneo/papers/Calonico-Cattaneo-Farrell-Titiunik_2016_wp.pdf.
- Calonico, S., M. D. Cattaneo and R. Titiunik (2014), Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82: 2295-2326.
- Calonico, S., M. D. Cattaneo and R. Titiunik (2015), Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110: 1753-1769.
- Imbens, G., Kalyanaraman and Karthik (2012), Optimal Bandwidth Choice for the Regression Discontinuity Estimator, *The Review of Economic Studies*, 79(3): 933–959.
- Khandker, S., R. Gayatri, B. Koolwal, Hussain and A. Samad (2010), Handbook on Impact Evaluation: Quantitative Methods and Practices”. The International Bank for Reconstruction and Development / The World Bank. Washington DC.
- Ludwig, J. and D. Miller (2007), Does head start improve children’s life chances life chance? Evidence from a Regression Discontinuity Design, *The Quarterly Journal of Economics*, 122(1): 159-208.
- Ravallion and Martin (2008), Evaluating Anti-Poverty Programs. In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3787–846. Amsterdam: North-Holland.

Chapter 22

SYNTHETIC CONTROL METHOD

Prabhat Kishore

INTRODUCTION

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 units 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 unit of considered population may have some differences either observable or unobservable. In reality, it is 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 is without intervention at earlier and same unit has 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 had not taken 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 the help of affected individual unit.

However, in some cases intervention takes place at state or country and policy makers are interested to know the 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 some times a few. So there is a lack of sufficient number of treated and control units for inferences. Major policy intervention occurred at macro level like Government of Bihar has repealed APMCs Act in 2006 with motif to remove restriction in agricultural marketing. To know its impact on agricultural GDP or prices of agricultural commodities, there is need to have robust method which can provide inference at macro level. Here intervention occurred at macro level as it was targeted

to affect whole agriculture of Bihar state and most of the other states still continued with APMCs Act. In this context, Synthetic Control Method provides new insight to tackle stated problem 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 yet to be seen. 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. Divergence in outcome values between the synthetic and treated unit in the post-treatment period if the intervention has a significant impact.

Econometric model

Suppose there is $S+1$ states in India where one state got intervention and remaining non intervention states are 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 $i= 1, 2, \dots, S+1$ and time $t=1, 2, \dots, T$. 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 that outcome of untreated unit does not get 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.

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

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

$$\sum_{m=1}^k V_m (X_{1m} - X_{0m}W)^2$$

by selecting the optimal values of W and V_m where X_{jm} is the value of the m^{th} attribute of the treated unit; X_{0m} is a $1 \times j$ 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 V_m 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 as $Y_1 - Y_0W$, 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

Case of APMC in Bihar

In the 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.

Following stata command used to create synthetic Bihar from donor pool:

xtset stcode year

synth agrigdp agriarea gca gia pvillelectric cropintensity, trunit(4) trperiod (2006) xperiod(1984(1)2005) nested fig

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	9932.49	6806.34
Net cropped area (thousand hectare)	6994.77	6981.73	4415.73
Net irrigated area (thousand hectare)	3342.52	3341.92	1854.18
Electrified villages (Per cent)	65.58	65.28	85.66
Cropping intensity (Per cent)	135.80	135.41	136.6

Fig 1 shows agricultural GDP of Bihar (bold line) and its synthetic (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 turn out to be noticeable divergence between actual and synthetic Bihar in post treatment revealing toward the insignificant impact of repeal on agricultural GDP. Table 2 present weights assigned to each state in order to construct counterfactual of Bihar agricultural GDP.

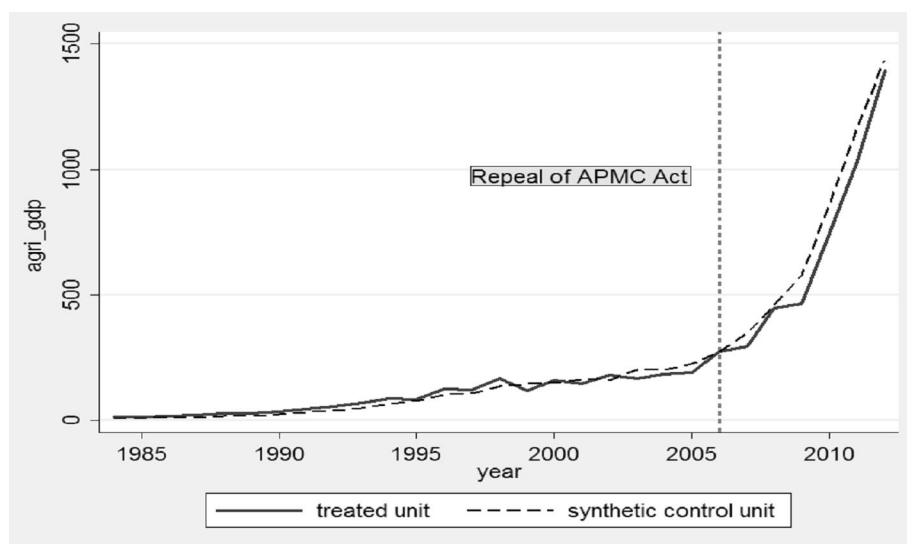


Fig 1: Trends in Agricultural GDP: Bihar Vs. Synthetic Bihar

Table 2: State weight in Synthetic Bihar

State	Assigned weight	State	Assigned weight
Andhra Pradesh	0.012	Mizoram	0.004
Arunachal Pradesh	0.403	Nagaland	0.003
Assam	0.008	Odisha	0.017
Delhi	0.003	Puducherry	0.001
Goa	0.003	Punjab	0.001
Gujarat	0.008	Rajasthan	0.148
Haryana	0.002	Sikkim	0.004
Himachal Pradesh	0.001	Tamil Nadu	0.063
Karnataka	0.009	Tripura	0.005
Madhya Pradesh	0.018	Uttar Pradesh	0.189
Maharashtra	0.013	West Bengal	0.007
Meghalaya	0.062		

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 the rest of the states 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 2 makes 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 21st 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.

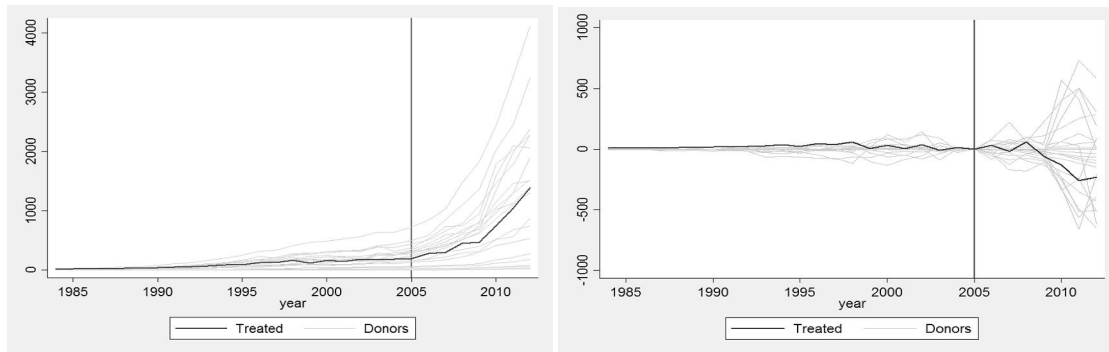


Fig 2: Agricultural GDP gap between Bihar & synthetic Bihar

Acknowledgment: This study is part of work done at IFPRI, SAO, New Delhi under professional attachment training with Dr Avinash Kishore and Dr. Devesh Roy.

REFERENCES

- Abadie, A., A. Diamond and J. Hainmueller (2015), Comparative politics and the synthetic control method. *American Journal of Political Science*, 59 (2): 495-510.
- Abadie, A. and J. Gardeazabal (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 J. Hainmueller (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.
- Galiani, S. and B. Quistorff (2016), The synth_runner Package: Utilities to Automate Synthetic Control Estimation Using synth, University of Maryland.
- Kreif, N., R. Grieve, D. Hangartner, S. Nikolov and M. Sutton (2016), Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units, *Health Economics*, 25: 1514–1528.

Chapter 23

INSTRUMENTAL VARIABLE ESTIMATION

Anuja A. R., K. N. Singh, Shivaswamy G. P., Rajesh T. and Harish Kumar H. V.

INTRODUCTION

In general, research issues in the social sciences are causal. Impact assessment studies focus on the influence of treatment on outcome. For example, while assessing the impact of a welfare initiative on poverty reduction, the welfare program is the treatment and poverty reduction is the intended outcome. Here, allotting treatment randomly to the experimental units is not feasible. Estimation of a causal relationship under such circumstances is problematic as it is difficult to establish that the treatments are exogenous to the investigated system.

One of the basic assumptions of Ordinary Least Square (OLS) is that there is no correlation between independent variables and residuals. When the predictor variable X is correlated with the error term U , the estimation of the causal effect using observational data will be biased. The problem can be addressed by adding additional exogenous variables to the model. In social science, Instrumental Variable (IV) technique is helpful to estimate the causal effect when there exists endogeneity. The Wu-Hausman test can be used to check endogeneity of treatment variable. IV can be used to solve the problem of omitted variable bias and the classic errors-in-variables problem.

Endogeneity occurs when there exists a correlation between independent variables and the error term. Let us take an example to explain the situation. Suppose we want to assess the impact of years of schooling on the earning of individuals. We observe correlation between years of schooling and the outcome variable i.e. earnings of individuals. But this correlation not necessarily indicates a causal relationship. Suppose, there is some unobservable variable that influences the outcome here such as IQ of the individual. There is a possibility that a better IQ of the individual is positively influencing both the treatment (years of schooling) and outcome variables (earnings of the individual). Fig 1 depicts the situations where causal inference in observational studies will be valid. The instrumental variable technique is an important tool used in the impact assessment studies in agriculture.

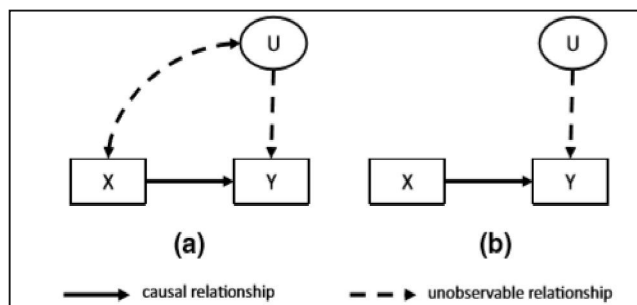


Fig 1: Examples of a situation where the modeling of causal relationships using observational data will be biased (a) and a situation where it will be valid (b)

(Adopted from Pokrope, 2016)

WHAT ARE THE INSTRUMENTAL VARIABLES?

Instrumental variable (IV) methods allow for endogeneity. An instrumental variable Z is an exogenous variable employed to assess the causal effect of variable X on Y (Fig 2).

A variable Z is an instrumental (relative to the pair (X, Y)) if

- (i) Z is independent of all variables (including error terms) that have an influence on Y that is not mediated by X and
- (ii) Z is not independent of X (Pearl 2000).

The first clause is referred to as the ‘exclusion’ and the second as the ‘relevance’.

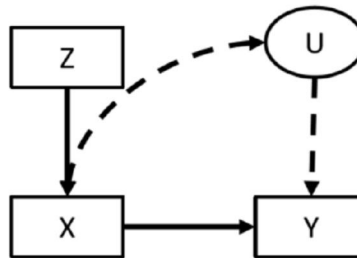


Fig 2: Situation where Z is a valid instrument (Pokrope, 2016)

Illustrating the application of instrumental variable technique in the agriculture

Birthal *et al.* (2015) employed IV technique to assess the impact of crop diversification on farm poverty in India. Unobserved features such as skill, motivation, etc. may lead to bias in the estimated coefficient. Using OLS regression to assess the impact may capture this unobserved heterogeneity and hence the estimates can suffer from bias. An instrumental variable was introduced into the model to mask unobserved heterogeneity at household level. As explained earlier, an ideal IV will not influence the outcome but will influence the treatment variable. In the study, the neighborhood effect based on geographical and social proximity was the IV. The logic of choosing the IV was that if the number of farmers growing high-value crops in the neighborhood is high it would positively influence the treatment variable i.e. area share of high-value crops. At the same time, the said IV would not affect the outcome variable of the model (farm poverty).

Selection of the instrumental variable

The selection of IV is of at most importance for the proper estimation of the causal effect. Finding a suitable instrumental variable for a large-scale database is a difficult task. Knowledge, experience and thorough understanding of the research issue can guide the researcher in finding proper IV for a situation. Weak instruments may worsen the bias in estimation (Khandker *et al.*, 2010). A value greater than 10 for the first stage F statistic indicates a strong instrument. This does not necessarily rule out a weak instrument issue.

Disadvantages of instrumental variable

There are many challenges associated with the application of IV variables in impact assessment. The very difficulty in finding a suitable IV following all the assumptions

is a major challenge. The poor performance of IV in small samples is another issue (Baum, 2008). The strength of the IV determines the precision. In comparison with the OLS estimates, IV estimates suffer from severe precision loss, if the instrument is weak. IV approaches are not immune from selection bias and the issue can be addressed by using the inverse probability of selection weights (Canan *et al.*, 2017)

IV technique using Two-Stage Least Squares (2SLS) regression

In the OLS regression, there is a basic assumption that all independent variables are uncorrelated with the error term. Two-Stage least squares (2SLS) regression analysis is employed when there exists problem of endogeneity (Gujarati *et al.*, 2012)

Problematic causal variable: This is the independent variable that is correlated with the error term or it is the variable that is influenced by other variables in the model. This endogenous causal variable is replaced with an instrumental variable in the first stage of the analysis.

Instrument variable: An instrumental variable is a new variable used in 2SLS to account for unobserved behavior between variables.

Estimation stages

First stage: A new variable is created using the instrument variable

Second stage: Instead of actual values of the problematic predictors, estimated values from the earlier stage is used in an OLS model to estimate the impact of the treatment variable

First stage regression:-

$$x_i = I\alpha + Zv + \delta_i \dots \quad (1)$$

x_i – Vector of the endogenous variable i (where $i = 1, \dots, N$)

I - Matrix for Instrumental variables

Z - Matrix of the covariates

δ_i - Error term

The role of the instrumental variables finishes at the first stage of 2SLS. Covariates are included in the first stage of the estimation to ensure that there is no direct influence of IV on the outcome. More than one IV can be employed in the first stage considering the appropriateness of the variables.

Second stage regression: -

$$y = \hat{x}_i\beta_i + Z\beta + e \dots \quad (2)$$

y - Vector of the outcome variable

- \hat{x}_1 - Vector of predicted values of x based on first stage regression
- β_i - Parameter estimate of the causal effect of X on Y
- Z - Matrix of the covariates
- β - Vector of slope parameters for the covariates from Z
- e - Error term.

Interpretation

The IV estimates indicate the local average treatment effect (LATE) instead of the average treatment effect (ATE). The ATE is the expected average effect of the treatment on outcome. The LATE provides information about the units that are likely to get the treatment if it is in the treatment group, but otherwise not take the treatment. The estimated LATE can be generalized for the population if there is no striking difference between the individuals influenced by the instrument and the population (Pokrope, 2016).

ILLUSTRATION

Suppose we want to study the impact of having health insurance on medical expenses. In the given example, the dependent variable is ‘medical expenses’ (y_1), the endogenous regressor is ‘having health insurance’ (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. In this example, social security income (ssi) ratio of the individual is used as an instrument (x_2). The IV represents variables assumed to affect ‘the choice of having health insurance or not’ but to have no direct effect on the outcome i.e. medical expenses. Table 1 indicates the sample data.

Table 1: Sample data

Number	Medical expenses	Health insurance	Age	Female	Income	Illnesses	ssi ratio
1	595	1	74	1	95	0	0.15
2	1783	1	73	0	36	3	0.40
.
.
.
n-1	720	0	69	1	29	1	0.15
n	809	1	90	1	21	1	0.36

Note: The data used in the illustrative example is a modified data from Katchova, A. (2013). Instrumental Variables in STATA. <https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables>.

OLS regression in STATA

First, define the dependent variable, independent variables, endogenous variable and instrumental variable. Command used for OLS regression in STATA – ‘**regress**’. Here the dependent variable is medical expenses (y_1). The endogenous regressor is ‘having

health insurance' (y_2) and exogenous regressors are illness, age, and income (x_1) of the individuals. Table 2 illustrates the results of OLS regression. The results indicate that for individuals with health insurance, the medical expenses are 7.5% higher than those for individuals without health insurance.

STATA Command: regress $y_1 y_2 x_{list}$

Table 2: Result of OLS regression

y_1 : log of medical expenses	Coef.	SE	t	P>t	[95% Conf. Interval]	
Health insurance (y_2)	0.075*	0.026	2.880	0.004	0.024	0.126
Illnesses (x_1)	0.441*	0.010	46.040	0.000	0.422	0.459
Age (x_1)	-0.003	0.002	-1.380	0.167	-0.006	0.001
Log of income (x_1)	0.017	0.014	1.250	0.211	-0.010	0.044
Constant	5.780*	0.151	38.310	0.000	5.484	6.076

* $p < 0.01$

2SLS estimation: - Command used for 2SLS regression using IV in STATA is as follows. ‘

Command: ivregress 2sls $y_1 (y_2 = x_2) x_{list}$

Table 3: Result of 2 SLS estimation

y_1 : log of medical expenses	Coef.	SE	t	P>t	[95% Conf. Interval]	
Health insurance (y_2)	-0.852*	0.198	-4.300	0.000	-1.241	-0.463
Illnesses (x_1)	0.449*	0.010	43.590	0.000	0.428	0.469
Age (x_1)	-0.012*	0.003	-4.230	0.000	-0.017	-0.006
Log of income (x_1)	0.098*	0.022	4.350	0.000	0.054	0.142
SS incomer ratio (instrument x_2)	-					
Constant	6.590*	0.235	28.090	0.000	6.130	7.050

* $p < 0.01$

X_{list} – Indicates list of exogenous variables

Table 3 explains the results of 2SLS with IV model. After instrumentation, for individuals with health insurance, their medical expenses are 85.2% lower than those for individuals without health insurance. It is evident from the results that the 2SLS coefficient turned out quite different from the OLS coefficient.

The following tests can be employed to ascertain the strength and suitability of the instruments.

Durbin-Wu-Hausman test for endogeneity

The endogeneity in the model can be tested using the Durbin-Wu-Hausman test for endogeneity. The Null hypothesis of the Durbin-Wu-Hausman test is that the independent variables are exogenous in nature. Rejection of null-hypothesis indicates the presence of endogeneity. The presence of endogeneity necessitates the usage of IV approach.

In the given example *test for endogeneity* was performed using the following command in STATA.

```
quietly ivregress 2sls y1 (y2=x2) x1list, first
estat endogenous
quietly regress y2 x2 x1list
quietly predict vhat, resid
quietly regress y1 y2 x1list vhat
testvhat
```

```
F( 1, 10083) = 25.14
Prob > F = 0.0000
```

```
Tests of endogeneity
Ho: variables are exogenous
```

```
Durbin (score) chi2(1) = 25.0914 (p = 0.0000)
Wu-Hausman F(1,10083) = 25.139 (p = 0.0000)
```

The rejection of null hypothesis confirmed the presence of endogeneity.

Correlation

The correlation between ‘having health insurance’ (endogenous variable) and ssi (IV) was tested and there was a negative correlation of -0.21. Here the correlation is weak and this may lead to biased estimates.

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,10084)	Prob > F
healthinsu	0.0684	0.0680	0.0194	68.881	0.0000

Weak instrument test -F statistics

As a thumb rule, if the value of F statistics of the model is greater than 10, instruments are not weak. Following commands were used to estimate the F statistics.

quietly ivregress 2sls y_1 ($y_2 = x_2$) x_{list} , vce (robust)

estat first stage, forcenonrobust

As the value is 69 (which is greater than 10 as per thumb rule), the given instrument is not weak.

Validity of multiple instruments.

The test for over-identifying restriction can be used to check the validity of multiple instruments. In the given example we have employed a single instrument.

REFERENCES

- Baum, C. F. (2008), Using Instrumental Variables Techniques in Economics and Finance. Boston College and DIW Berlin. German Stata Users Group Meeting, Berlin, June 2008. Online available at <https://www.stata.com/meeting/germany08/Baum.DESUG8621.beamer.pdf>
- Birthal, P. S., D. Roy and D. S. Negi (2015), Assessing the Impact of Crop Diversification on Farm Poverty in India. *World Development*, Elsevier, 72(C), 70-92.
- Canan, C., C. Lesko and B. Lau (2017), Instrumental Variable Analyses and Selection Bias. *Epidemiology (Cambridge, Mass.)*, 28(3); 396–398. doi:10.1097/EDE.0000000000000639
- Gujarati, D. N., D. C. Porter and S. Gunasekar (2012), *Basic Econometrics*. McGraw Hill Education (India) Private Limited.
- Katchova, A. (2013), Instrumental Variables in STATA. Online available at <https://sites.google.com/site/econometricsacademy/econometrics-models/instrumental-variables>.
- Khandker, S. R., G. B. Koolwal and H. A. Samad (2010), *World Bank: Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank.
- Pearl, J. (2000), *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.
- Pokrope, A. (2016), Introduction to Instrumental Variables and their Application to Large-Scale Assessment Data. *Large-scale Assessments in Education* 4:4 DOI 10.1186/s40536-016-0018-2.

PART V

GROWTH ANALYSIS

Chapter 24

COMPUTABLE GENERAL EQUILIBRIUM MODELS

Balaji S. J.

INTRODUCTION

“Everything depends on everything else”. The linkages among economic sectors invariably prove all economic events are related with each other *i.e.* a hike in oil prices raises not just the prices of commodities traded but might lead to a) a contraction in investment that a farmer makes on his land as he is supposed to spend more for running his tractor, thus affecting his production; b) would cut-down his family’s budget on health or entertainment expenditure; or c) might lead the workers to demand for more wages due a rise in food and transport expenditure. In short, an economic shock gets transmitted to an entire economy having multiple implications. These interlinkages form the fundamental structure of CGE models. They attempt to incorporate all economic activities like agriculture, industries and services, the agents involved such as farmers, agricultural laborers, industrial workers, the self-employed entrepreneurs, Government, private enterprises, export and import activities etc. More formally, and technically, one would define CGE models as “a system of equations that describes an economy as a whole and the interactions in the parts within”. In a general framework, these equations describe producer and consumer behaviour and impose market clearing constraints and they are solved for the set of prices at which the quantities of supply and demand are in equilibrium (Burfisher, 2011 and Dixon, 2013).

For simplicity, assume that there exist one representative household that reflect rest of the households’ behaviour in an economy. Further, assume that there are only two type of commodities produced, namely a) a food commodity and b) a non-food commodity, labor and capital are the only factors used in producing these goods. For producing these commodities, firms demand labor and capital services from the households. In return, household are paid to these services in the form of wages and rent. The income earned this way are used by the households for purchasing the food and non-food articles produced by the firms. For time being, let’s assume there exist to trade with rest of the regions and no Government exists to tax on income from the households and no financial institutions exist to attract their savings. These set of activities cover a circular form of economy which one would have observed in macroeconomic subjects. Though the events are macroeconomic in nature, behaviour of households and industries are essentially derived from the core of microeconomic structure. Households are assumed to maximize their consumption utility subject to their budget constraints. Firms are assumed to maximize their profit subject to their production technologies. Prices are assumed to move freely to respond demand-supply interactions. Given, their behaviour,

the entire set of actions are solved simultaneously so as equilibrium is achieved in both factor and product market simultaneously.

Representing the economic behaviour of different stakeholders in a mathematical form would help to represent the system more clearly. For the simple case that was discussed above, the mathematical equations³ of different agents and market clearing conditions are shown below.

Household behaviour

The utility maximization behaviour of households could be represented as

$$\max_{X_i} UU = \prod_i X_i^{\alpha_i}$$

given the budget constraint

$$\sum_i p_i^x X_i = \sum_h p_h^f F_h$$

where,

i, j : goods

h, k : factors

UU : utility

X_i : consumption of i^{th} good ($X_i \geq 0$)

FF_h : endowment of h^{th} factor for the household

p_i^x : demand price of i^{th} good ($p_i^x \geq 0$)

p_h^f : price of h^{th} factor ($p_h^f \geq 0$)

α_i = share parameter in the utility function ($0 \leq \alpha_i \leq 1, \sum_i \alpha_i = 1$)

For the i^{th} good, the demand function would then be

$$X_i = \frac{\alpha_i}{p_i^x} \sum_h p_h^f F_h \quad \forall i$$

Firm behaviour

The profit maximizing behaviour of firm could be given as

$$\max_{Z_j, F_{h,j}} \pi_j = p_j^z Z_j - \sum_h p_h^f F_{h,j}$$

for their technology function

³ The equations and notations are based on Hosoe *et al.* (2010)

$$Z_j = b_j \prod_h F_{h,j}^{\beta_{h,j}}$$

where,

i, j : firm

h, k : factor

π_j : profit of j^{th} firm

Z_j : output of j^{th} firm

$F_{h,j}$: the h^{th} factor used by j^{th} firm

p_j^z : supply price of j^{th} good

p_h^f : price of h^{th} factor

$\beta_{h,j}$: share coefficient in the production function ($0 \leq \beta_{h,j} \leq 1, \sum_h \beta_{h,j} = 1$)

b_j : scale coefficient in the production function

The demand function would be

$$F_{h,j} = \frac{\beta_{h,j}}{p_h^f} p_j^z Z_j \quad \forall h, j$$

Market clearance

The market clearing equations would be

$$\begin{aligned} X_i &= Z_i \quad \forall i \\ \sum_j F_{h,j} &= F_h \quad \forall h \\ p_i^z &= p_i^x \quad \forall i \end{aligned}$$

Solving these equations simultaneously provides equilibrium level of prices at which both demand and supply of commodities are matched.

SOCIAL ACCOUNTING MATRIX

In practice, the Social Accounting Matrix (SAM) is used to construct and solve a CGE model. Part of the coefficients displayed in the former equations arise from independent econometric exercises. The SAM depicts flow of all transfers and transactions in a given time among all economic agents. It represents the 'real' value of economic transactions hence the structure refers to an 'actual' economy. A major advantage of employing CGE model is that the data requirements are relatively smaller. And since most of the models deal with a wider geographical region such as country, group of countries or the entire world, one would expect most of the data aggregates could be compiled from their respective national accounts or data repositories, which saves time and cost.

In case of national models, one could construct the Social Accounting Matrix (SAM) using the estimates published in National Accounts Statistics (NAS). Let's look into the case of India. The NAS estimates show Gross Value Added (GVA) estimate for the year 2015-16 as Rs. 105 trillion at 2011-12 prices. The contribution of agriculture (Food) and non-agriculture (Non-food) were Rs. 16 trillion and Rs. 89 trillion respectively. Assuming this value addition is generated using labor and capital as the only factors and there were no intermediate inputs involved, one would estimate the factor contribution if the coefficients for such estimates are available. For example, if we assume that the contribution of labor and capital are 49% and 51% respectively for the overall economy, the entire value added output can be distributed among these factors using these estimates. In the present case, the shares assumed for labor were 48% in the total value of food produced and 55% for the non-food commodities. Employing these shares against corresponding output values, we obtain the factor contribution estimates for both food and non-food production activities.

Table 1: Simple Social Accounting Matrix for India (2015-16)

		Activity		Factor		Final demand	Total
		Food	Nonfood	Capital	Labor	Household	
Activity	Food (F)					16	16
	Nonfood (N)					89	89
Factor	Capital (K)	7	46				53
	Labor (L)	9	43				52
Final demand	Household (H)			53	52		105
Total		16	89	53	52	105	

Unit: Rs. In trillion

It is implicit from the Table 1 that one would obtain factor earnings using the output estimates. Since households are the only agent in the present context, entire factor is derived through agents. They earn 53 trillion by offering their labor services and 52 trillion by providing capital for food and non-food production. One would note that the total value product equals with factor earnings. These earnings are used by the households for food and non-food consumption. These demand estimates emerge from the supply block in which reported the total value of food and non-food items produced in the country. Estimates in different rows and columns satisfy demand supply equality criterion *i.e.* row sum should equal the column sum. To note, the Table 1 presented is based on the assumption of absence of intermediate inputs, trade, savings and investment, and the Government.

The following section provides the codes to implement the simple CGE structure described above. Further, it uses the SAM we constructed using few NAS estimates for India for a preliminary understanding. The summary of results one would observe in GAMS is given at the end.

GAMS Code

Set

u 'SAM entry' / F, N, K, L, H /

i(u) 'goods' / F, N /

h(u) 'factor' / K, L /;

Alias (u,v), (i,j), (h,k);

Table SAM(u,v) 'social accounting matrix'

	F	N	K	L	H
F					16
N					89
K	7	46			
L	9	43			
H			53	52	

Parameter

X0(i) 'household consumption of the i-th good'

F0(h,j) 'the h-th factor input by the j-th firm'

Z0(j) 'output of the j-th good'

FF(h) 'factor endowment of the h-th factor';

X0(i) = SAM(i,"HOH");

F0(h,j) = SAM(h,j);

Z0(j) = sum(h, F0(h,j));

FF(h) = SAM("HOH",h);

display X0, F0, Z0, FF;

Parameter

alpha(i) 'share parameter in utility function'

beta(h,j) 'share parameter in production function'

b(j) 'scale parameter in production function';

alpha(i) = X0(i)/sum(j, X0(j));

beta(h,j) = F0(h,j)/sum(k, F0(k,j));

b(j) = Z0(j)/prod(h, F0(h,j)**beta(h,j));

display alpha, beta, b;

Variable

X(i) 'household consumption of the i-th good'

F(h,j) 'the h-th factor input by the j-th firm'

Z(j) ‘output of the j-th good’
 px(i) ‘demand price of the i-th good’
 pz(j) ‘supply price of the i-th good’
 pf(h) ‘the h-th factor price’
 UU ‘utility [fictitious]’;

Equation

eqX(i) ‘household demand function’
 eqpz(i) ‘production function’
 eqF(h,j) ‘factor demand function’
 eqpx(i) ‘good market clearing condition’
 eqpf(h) ‘factor market clearing condition’
 eqZ(i) ‘price equation’
 obj ‘utility function [fictitious]’;

eqX(i).. $X(i) = e = \alpha(i) * \sum(h, pf(h) * FF(h)) / px(i)$;
 eqpz(j).. $Z(j) = e = b(j) * \prod(h, F(h,j) ** \beta(h,j))$;
 eqF(h,j).. $F(h,j) = e = \beta(h,j) * pz(j) * Z(j) / pf(h)$;
 eqpx(i).. $X(i) = e = Z(i)$;
 eqpf(h).. $\sum(j, F(h,j)) = e = FF(h)$;
 eqZ(i).. $px(i) = e = pz(i)$;
 obj.. $UU = e = \prod(i, X(i) ** \alpha(i))$;

X.l(i) = X0(i);
 F.l(h,j) = F0(h,j);
 Z.l(j) = Z0(j);
 px.l(i) = 1;
 pz.l(j) = 1;
 pf.l(h) = 1;

X.lo(i) = 0.001;
 F.lo(h,j) = 0.001;
 Z.lo(j) = 0.001;
 px.lo(i) = 0.001;
 pz.lo(j) = 0.001;
 pf.lo(h) = 0.001;
 pf.fx(“LAB”) = 1;
 Model splcge / all /;
 solve splcge maximizing UU using nlp;

The above syntax displays the results in GAMS in detail. For space, only the solution report is displayed for discussion.

Solution Report SOLVE splcge Using NLP From line 107

S O L V E S U M M A R Y

MODEL splcge OBJECTIVE UU
 TYPE NLP DIRECTION MAXIMIZE
 SOLVER CONOPT FROM LINE 107

**** SOLVER STATUS 1 Normal Completion
 **** MODEL STATUS 2 Locally Optimal
 **** OBJECTIVE VALUE 68.5212

RESOURCE USAGE, LIMIT 0.000 1000.000
 ITERATION COUNT, LIMIT 4 2000000000
 EVALUATION ERRORS 0 0
 CONOPT 3 25.0.3 r65947 Released Mar 21, 2018 WEI x86
 64bit/MS Windows

C O N O P T 3 version 3.17G
 Copyright (C) ARKI Consulting and Development A/S
 Bagsvaerdvej 246 A
 DK-2880 Bagsvaerd, Denmark
 Pre-triangular equations: 0
 Post-triangular equations: 1
 Definitional equations: 4

** Optimal solution. There are no superbasic variables.

CONOPT time Total 0.002 seconds
 of which: Function evaluations 0.000 = 0.0%
 1st Derivative evaluations 0.000 = 0.0%

---- EQU eqX household demand function

	LOWER	LEVEL	UPPER	MARGINAL
BRD	.	.	.	0.653
MLK	.	.	.	0.653

---- EQU eqpz production function

	LOWER	LEVEL	UPPER	MARGINAL
BRD	.	.	.	0.653
MLK	.	.	.	0.653

---- EQU eqF factor demand function

	LOWER	LEVEL	UPPER	MARGINAL
CAP.BRD	.	.	.	0.653

CAP.MLK	.	.	. 0.653
LAB.BRD	.	.	. 0.653
LAB.MLK	.	.	. 0.653

---- EQU eqpx good market clearing condition

	LOWER	LEVEL	UPPER	MARGINAL
BRD	.	.	.	EPS
MLK	.	.	.	EPS

---- EQU eqpf factor market clearing condition

	LOWER	LEVEL	UPPER	MARGINAL
CAP	53.000	53.000	53.000	.
LAB	52.000	52.000	52.000	EPS

---- EQU eqZ price equation

	LOWER	LEVEL	UPPER	MARGINAL
BRD	.	.	.	-10.441
MLK	.	.	.	-58.080

	LOWER	LEVEL	UPPER	MARGINAL
---- EQU obj	.	.	.	1.000
obj utility function [fictitious]				

---- VAR X household consumption of the i-th good

	LOWER	LEVEL	UPPER	MARGINAL
BRD	0.001	16.000	+INF	.
MLK	0.001	89.000	+INF	.

---- VAR F the h-th factor input by the j-th firm

	LOWER	LEVEL	UPPER	MARGINAL
CAP.BRD	0.001	7.000	+INF	.
CAP.MLK	0.001	46.000	+INF	.
LAB.BRD	0.001	9.000	+INF	.
LAB.MLK	0.001	43.000	+INF	.

---- VAR Z output of the j-th good

	LOWER	LEVEL	UPPER	MARGINAL
BRD	0.001	16.000	+INF	.
MLK	0.001	89.000	+INF	.

---- VAR px demand price of the i-th good

	LOWER	LEVEL	UPPER	MARGINAL
BRD	0.001	1.000	+INF	.
MLK	0.001	1.000	+INF	.

---- VAR pz supply price of the i-th good

	LOWER	LEVEL	UPPER	MARGINAL
BRD	0.001	1.000	+INF	.
MLK	0.001	1.000	+INF	.

---- VAR pf the h-th factor price

	LOWER	LEVEL	UPPER	MARGINAL
CAP	0.001	1.000	+INF	.
LAB	1.000	1.000	1.000	EPS

	LOWER	LEVEL	UPPER	MARGINAL
---- VAR UU		-INF	68.521	+INF .

UU utility [fictitious]

```

**** REPORT SUMMARY : 0          NONOPT
                      0          INFEASIBLE
                      0          UNBOUNDED
                      0          ERRORS
    
```

EXECUTION TIME = 0.000 SECONDS 2 MB 25.0.3 r65947 WEX-WEI

USER: GAMS Development Corporation, USA G871201/0000CA-ANY
Free Demo, +1 202-342-0180, support@gams.com, www.gams.com DC0000

One should observe no syntax error at the right side when the above result is displayed. The appearance of “** Optimal solution” for the above analysis shows the optimal solution was found and absence of problems due to infeasibility or unboundedness.

REFERENCES

- Hosoe, N., K. Gasawa and H. Hashimoto (2010), Handbook of Computable General Equilibrium Modeling. University of Tokyo Press, Tokyo, Japan.
- Burfisher, M. (2011), Introduction to Computable General Equilibrium Models. Cambridge University Press, New York, USA.
- Dixon, P. B. and D. W. Jorgenson (2013), Handbook of Computable General Equilibrium Modelling-Volume 1A. North-Holland Publishers, Oxford, UK.

Chapter 25

DECOMPOSITION OF TOTAL FACTOR PRODUCTIVITY: DEA APPROACH

Dharam Raj Singh, Suresh Kumar, Venkatesh P. and Philip Kuriachen

INTRODUCTION

Productivity indicates the efficiency of a production system; it is the ratio of units of output per unit of input. In other words, it is a relationship between the quantity of output and the quantity of input used to generate an that output. In productivity analysis, when we have multi-output and multi-inputs models, then the output could be of different forms viz., goods or provided services. Further, outputs can be measures in physical (quantities) or financial (value) terms. On the other side, most common forms of inputs are labor, capital, and intermediate inputs. The most commonly used measures of productivity are single (partial) factor productivity (SFP) and total (multifactor) factor productivity (TFP). TFP is defined as the portion of output not explained by the amount of inputs used in production. Its value represents how efficiently and intensely the inputs are utilized in production. In case of multiple inputs model, inputs are heterogeneous in nature, therefore, these inputs are aggregated in an index using price indices.

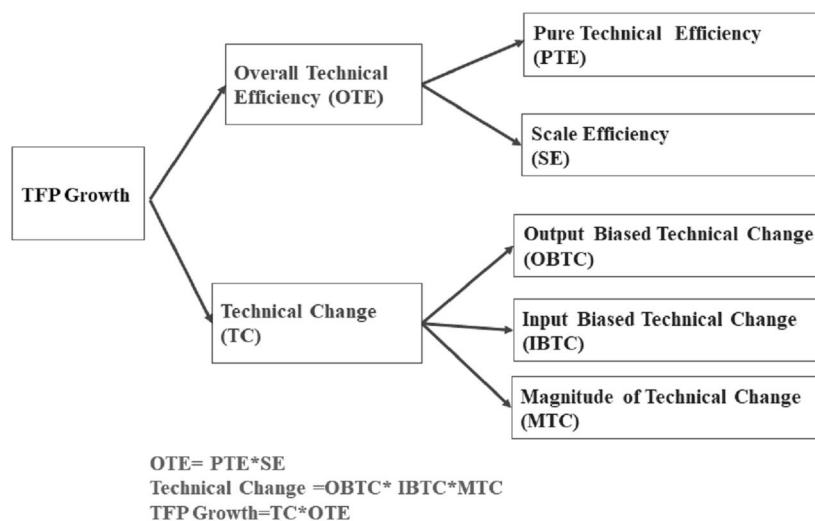


Fig 1: Decomposition of TFP growth

The agricultural output growth is usually due to three types of factors: area growth, yield growth, and prices change. Area growth induces a growth in the quantity of input use in addition to land use. On the other side, the yield growth is generated by both input use growth and productivity growth. Further, the TFP growth is the result

of both technical change and better technical efficiency of the used factors. Technical efficiency (overall) can also be decomposed into pure technical and scale efficiencies. Similarly, technical change can be decomposed into input-biased technical change, output biased technical change and magnitude of technical change (Figure 1). Since TFP measures account for the use of a number of factor inputs in production and, therefore, are more suitable for performance measurement and comparisons across firms and for a given firm over time (Coelli *et al.*, 2005). And TFP can be defined as a ratio of aggregate output produced relative to aggregate input used. Following are some of the definitions:

$$\begin{aligned}
 \text{TFP Growth} &= \text{Output Growth} - \text{Input Growth} \\
 &= \text{Embodied (or endogenous) and Disembodied (or exogenous)} \\
 &\quad \text{Technical Change} \\
 &= \text{Changes in Technical efficiency} + \text{Technological Progress}
 \end{aligned}$$

Measures of TFP Growth

Approaches to measure the TFP growth can be broadly categorized into groups, namely frontier and non-frontier approaches. Non-frontier approaches include growth accounting methods (non-parametric index-based methods) and econometric parametric approaches. On the other side, frontier approaches include the non-parametric Malmquist index methodology and parametric approach such as stochastic production frontier method. In frontier approach, the objective is to estimate the best obtainable positions based on the estimation of a bounding function, given inputs and prices levels. For example, a cost frontier traces the minimum attainable cost given input prices and output while a production frontier traces the set of maximum attainable output for a given set of inputs and technology. This approach differs from the parametric non-frontier approaches where an average function is often estimated by the ordinary least square regression. Further, non-frontier approaches assume that firms are technically efficient whereas frontier approaches identify the role of technical efficiency in overall firm performances. This leads to different interpretation of TFP growth estimated from both approaches. This ensures that output elasticity is equal to input shares in the total cost represents total factor productivity (*TFP_t*) or that part of output growth which cannot be explained by growth in any of the inputs.

TFP growth as obtained from frontier approach consists of two components: (i) outward shifts of the production function resulting from technological progress, and (ii) technical efficiency related to the movements towards the production frontier. The non-frontier approach considers technological progress as a measure of TFP growth. Both frontier and non-frontier approaches can be estimated through parametric and non-parametric techniques. Parametric estimations need the specification of a functional form for the frontier and parameters are estimated through econometric techniques using sample data, however, the accuracy of the estimates is sensitive to the specification of functional form. In non-parametric methods (such as data envelopment

analysis (DEA), there is no need of specifying the functional forms. However, the estimates of non-parametric approaches cannot test statistically.

NON-FRONTIER APPROACHES

A common feature of the TFP index number is that the empirical estimation of different TFP indexes is based on different weighting methods of inputs and outputs. In most commonly used growth accounting methods are: Divisia, Solow, and the Tornqvist indexes.

Solow index method: Solow uses a Cobb-Douglas production function in order to estimate the TFP growth. Estimation of this production function is based on the assumptions such as constant return to scale, autonomous Hick's neutral technical change, and that the factor payments are equal to their marginal products. The production function form is:

$$\ln\left(\frac{TFP_t}{TFP_{t+1}}\right) = \ln\frac{y_t}{y_{t+1}} - \sum_i \beta_{i,t} \ln\left(\frac{x_{i,t}}{x_{i,t+1}}\right)$$

where, y_t is the output in period t and $x_{i,t}$ is the i^{th} input used in period t . The symbol β_i denotes the factor share of the input $x_{i,t}$ such that $\beta_i = \frac{w_i x_i}{\sum w_i x_i}$, where w_i is the income of factor x_i . There is a constant return to scale given by $\sum \beta_i = 1$. This ensures that output elasticity is equal to input shares in the total cost TFP_t represents total factor productivity (TFP) or that part of output growth which cannot be explained by growth in any of the inputs.

Tornqvist index: Among index number methods, Tornqvist-Theil Index, which is an approximation to Divisia Index, is to commonly used for constructing the aggregate output index and aggregate input index. The Tornqvist output, input and TFP index in logarithm form can be expressed as follows:

$$TFP \text{ growth} = \ln\left(\frac{TFP_t}{TFP_{t-1}}\right) = \ln\frac{y_t}{y_{t-1}} - \sum_i \frac{1}{2}(S_{it} - S_{it-1})\left(\frac{x_{i,t}}{x_{i,t-1}}\right)$$

where, y_t and y_{t-1} represent the output in periods t and $t-1$, respectively; S_{it} , factor share of input $x_{i,t}$ in period t , and S_{it-1} is the factor share of input $x_{i,t-1}$ in period $t-1$.

The TFP index is an approximation of technological progress, assuming that producers behave competitively, that the production technology is input-output separable, and that there is no technical inefficiency (Antle and Capalbo, 1988). Solow index method and Törnqvist-Theil index methods impose certain theoretical restrictions. These approaches compute TFP based on the assumption that factor income shares are equal to output elasticities (factors are paid their marginal product) and that there is prompt adjustment to altered market conditions. In these approaches that

observed output is equivalent to frontier output, and that growth in TFP comprised only technological progress, that is, shifts in the frontier. As a matter of fact, these TFP estimates may be biased as the firms may be technically inefficient in the use of inputs. These approaches also ignore the fact that fertilizers and irrigation are highly subsidized and output prices are also distorted in the form of minimum support price by the government in India. Under such conditions, there would be biased productivity estimates as input prices are not being determined competitively in the factor market. Further, these methods were also criticized as these assume certain production functions and it may lead to specification biased results. Being non-frontier approaches, i.e. it assumed to behave optimally and therefore, they always operate on the frontier, thus, any change in the TFP index will only reflect shifts in the production frontier. Therefore, change in the technical efficiency cannot be measured. Moreover, most importantly, in these methods residuals may reflect factors other than technical change, e.g. changes in capacity utilization between periods due to demand-side or supply-side fluctuations in the availability of electricity and other inputs.

FRONTIER APPROACHES

Frontier approaches assume the existence of a production function corresponding to the set of maximum attainable output levels for a given input combinations. The advantage of this approach is that it decomposes the changes in TFP into technological progress and technical efficiency changes; the former associated with changes in the best-practice production frontier, and the latter with other productivity changes, such as learning by doing, improved managerial practices, and changes in the efficiency with which a known technology is applied (Kathuria *et al.*, 2013). The two main approaches in the estimation of TFP growth using frontier methods are the Malmquist (nonparametric approach) and the stochastic frontier methods (parametric approaches).

Parametric approach

The stochastic frontier method (Aigner *et al.*, 1977) estimation used standard production function methodology on cross-sectional data of N observed firms. It assumes that a firm (i) uses inputs X_i ($i = 1, \dots, N$) to produce an output Y_i , and the function can be written as follows:

$$Y_i = f(X_i\beta) + (v_i - u_i)$$

The particularity of this model is that the error term is divided into two main components. These are the usual random noise component (v_i) and the inefficiency component (u_i). The noise component is measuring measurement errors and other random errors which are beyond the farmer's capacity. The error term v_i representing pure randomness makes the production frontier stochastic and thus allows the frontier to vary over time for the same firm (farm). This error term $v \sim N(0, \sigma_v^2)$ is a two-sided error term symmetrically

distributed ($-\infty < v_i < \infty$) and it captures the effects of random shocks outside the firm (farm) control, observation and measurement error on independent variables, and usual statistical ‘noise’ generally found in an empirical relationship.

Independent error component (u_i) is assumed to be non-negative and represents technical efficiency. This error term is one-sided and is a truncation of the $u \sim N(0, \sigma_u^2)$ distribution (i.e; half normal distribution or having exponential distribution). Thus, the value of u measures the firm inefficiency level which is also expressing how far a firm’s given output is from its potential output compared to other firms of the sample.

Non-Parametric Approaches

Data Envelopment Analysis

This approach is similar to the stochastic frontier approach with the unique difference of non-requirement for parameters estimation for the farmers’ production technology description. Instead, the technology of the best performing farmer(s) (frontier(s)) is (are) considered as benchmark, and the efficiency of the rest of farmers in the sample will be measured accordingly from the frontier. Use of DEA approach aims to provide measures of the efficiency and productivity of farmers. For the DEA approach, similar to SFA approach, input-output matrix is needed in order to estimate farmers’ TFP and technical efficiency. Unlike the parametric estimation, the deterministic estimation has a single one-sided error component where u is greater than 0 represents technical inefficiency. The advantages of DEA approach are: this can be used with small dataset, without variation in prices, chooses one or more frontier(s) from existing firms and does not requires specification of distribution and functional.

The advantages of DEA approach: this can be used with small dataset, without variation in prices, chooses one or more frontier(s) from existing firms and does not requires specification of distribution and functional.

Malmquist productivity index

The Malmquist productivity index was first introduced by Caves *et al.* (1982). The non-parametric estimation of this index was initiated by Färe *et al.* (1994). The defined the TFP index using Malmquist input and output distance functions (Appendix I). They showed that comparing each firm to the best practice frontier firm provides a measure of its efficiency and a measure of shift in the frontier (from one period to another) which is also similar to the technological progress.

There $x^t = (x_1^t, \dots, x_N^t)$ inputs at period $t = 1, \dots, T$ that are used to produce outputs $y^t = (y_1^t, \dots, y_M^t)$ then the technology at t consists of all feasible (x^t, y^t) i.e.

$$S^t = \{(x^t, y^t): x^t \text{ can produce } y^t\}, x^t \in \mathbb{R}_+^N \text{ and } y^t \in \mathbb{R}_+^M$$

The output distance function as per Ronald Shephard (1970) and is defined relative to the technology S^t as

$$D_0^t(x^t, y^t) = \min \left\{ \theta : \left(x^t, \frac{y^t}{\theta} \right) \in S^t \right\}, x^t \in \mathbb{R}_+^N, t = 1, 2, \dots, T$$

Given x^t , the distance function increases y as much as possible while remaining in S^t . This function is a complete characterization of the technology. We note that $D_0^t(x^t, y^t) \in S^t$ if and only if $D_0^t(x^t, y^t) \leq 1$.

The Malmquist productivity change index computed here is based on the simple idea illustrated above, but it allows comparisons between two periods. Distance functions are used to provide a measure of deviations from maximum average product. The Malmquist TFP index was first introduced by Caves *et al.* (1982). They defined the TFP index using Malmquist input and output distance functions (Appendix I), and thus the resulting index came to be known as the Malmquist TFP index. The period t Malmquist productivity index is given by

$$M^t = \frac{D_0^t(X^{t+1}, Y^{t+1})}{D_0^t(X^t, Y^t)} \dots \quad (1)$$

i.e., they define their productivity index as the ratio of two output distance functions taking technology at time t as the reference technology. Instead of using period t 's technology as the reference technology it is possible to construct output distance functions based on period $(t+1)$'s technology and thus another Malmquist productivity index can be laid down as:

$$M^{t+1} = \frac{D_0^{t+1}(X^{t+1}, Y^{t+1})}{D_0^{t+1}(X^t, Y^t)} \dots \quad (2)$$

The MPI consists of four output distance functions to avoid choosing arbitrary base period and the geometric mean of two output based technical efficiency indices is taken to form:

$$MPI(x^t, x^{t+1}, y^t, y^{t+1}) = \left[\frac{D_0^{t+1}(X^{t+1}, Y^{t+1})}{D_0^{t+1}(X^t, Y^t)} * \frac{D_0^t(X^{t+1}, Y^{t+1})}{D_0^t(X^t, Y^t)} \right]^{0.5} \dots \quad (3)$$

The Malmquist productivity index (MPI) yields a convenient way of decomposing TFP change into technical change (TECH change) and overall technical efficiency change (OTE Change). The MPI can be decomposed as:

$$MPI = \underbrace{\frac{D_0^{t+1}(X^{t+1}, Y^{t+1})}{D_0^t(X^t, Y^t)}}_{\text{OTE Change}} \underbrace{\left[\frac{D_0^t(X^t, Y^t)}{D_0^{t+1}(X^t, Y^t)} * \frac{D_0^t(X^{t+1}, Y^{t+1})}{D_0^{t+1}(X^{t+1}, Y^{t+1})} \right]^{0.5}}_{\text{TECH change}} \dots \quad (4)$$

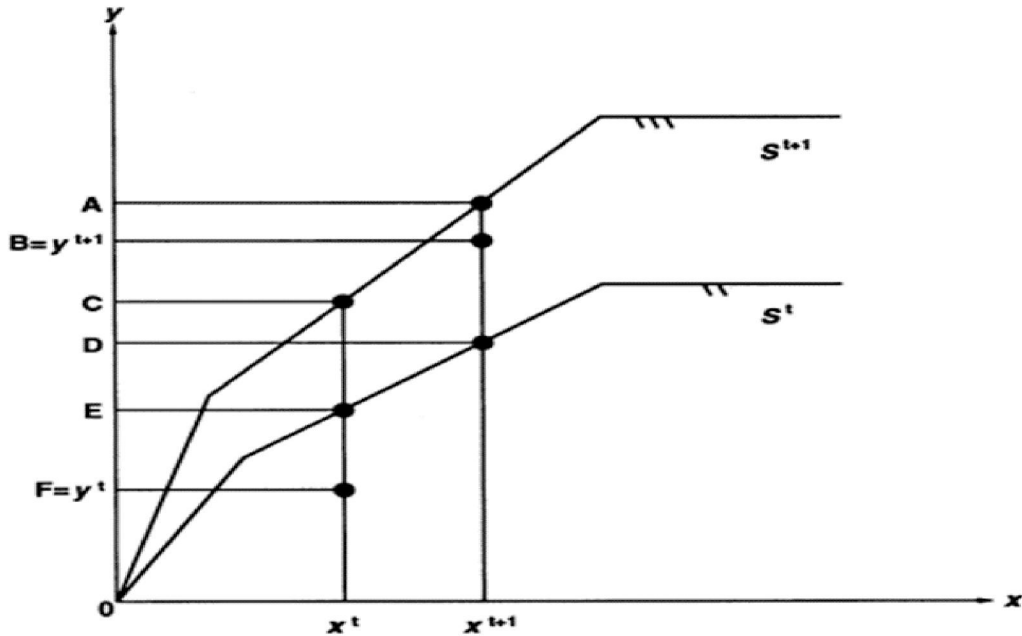


Fig 2: Illustration of the Malmquist output-based productivity index (Färe et al., 1994)

The Fig 2 shows the graphical decomposition of TFP into its components. Two ratios in the square bracket can be thought of as measures of technical progress as measured by shifts in the frontier measured at period $t+1$ (OA/OD) and period t (OC/OE) and then averaged geometrically. The terms outside the bracket represent the changes in efficiency between the two periods (OB/OA and OF/OE). This defines the changes in OTE from period t to $t+1$, i.e., moving closer to the isoquant or ‘catching up’. The second term, i.e., the geometric mean in parenthesis, represents changes in technology, i.e., a shift in the frontier from period t to period $t+1$.

We know that $OTE = PTE \times SE$, therefore, OTE change can be further decomposed into pure technical efficiency change (PTE change) and input scale efficiency change (SE change). Where $PTE \text{ change} = \frac{PTE^{t+1}}{PTE^t}$ and $SE \text{ change} = \frac{SEC^{t+1}}{SEC^t}$.

The MPI can be written as:

$$MPI = PTE \text{ change} * SE \text{ change} * TECH \text{ change} \dots \quad (5)$$

In the output-oriented case all the indices can be interpreted as progress, no change, and regress, when their values are greater than one, equal to one, and less than one, respectively. Further, Technological change can be decomposed into a product of output-biased technological change (OBTC), input-biased technological change (IBTC) and the magnitude of technological change (MTC).

Where,

$$\text{TECH Change} = \text{OBTC} * \text{IBTC} * \text{MTC} \dots \quad (6)$$

$$\text{OBTC} = \left[\frac{D_0^t(X^{t+1}, Y^{t+1})}{D_0^{t+1}(X^{t+1}, Y^{t+1})} * \frac{D_0^{t+1}(X^{t+1}, Y^t)}{D_0^t(X^{t+1}, Y^t)} \right]^{0.5} \dots \quad (7)$$

$$\text{IBTC} = \left[\frac{D_0^{t+1}(X^t, Y^t)}{D_0^t(X^t, Y^t)} * \frac{D_0^t(X^{t+1}, Y^t)}{D_0^{t+1}(X^{t+1}, Y^t)} \right]^{0.5} \dots \quad (8)$$

$$\text{MTC} = \frac{D_0^t(X^t, Y^t)}{D_0^{t+1}(X^t, Y^t)} \dots \quad (9)$$

Since, we are considering only one output i.e. sorghum yields in the present example (given below), there will be no output-biased technological change, i.e., OBTC=1, and equation (6) reduces to

$$\text{TECH Change} = \text{IBTC} * \text{MTC} \dots \quad (10)$$

ILLUSTRATION

Estimation of output oriented Malmquist TPF Index of sorghum in India.

We have compiled data on sorghum yields and input usage from various issues of cost of cultivation survey (aggregate level data). The data were compiled for the time period 2004-2012. Data were compiled for six major sorghum producing states. In analysis the per ha output variable used was yield and five input variables, namely, fertilizer applied (kg/ha), seeds (kg/ha), manure applied (q/ha), human labor (hours/ha) and animal labor (hours) were used. The analysis was undertaken using DEAP 2.1 software. TFP was estimated using Malmquist Productivity Index as given in appendix II.

The average TFP indices of sorghum and their components are reported in Table 1 for selected states. Values of indices greater than one, less than one and equal to one indicate improvements, declines and no change in the performance. Mean TFP growth in sorghum was impressive (6.4%) and this growth is contributed by advancement in the technology. State-wise results showed that Andhra Pradesh, Karnataka, Madhya Pradesh, Maharashtra, Rajasthan and Tamil Nadu registered total factor productivity growth of 4.2, 6.1, 7.3, 5.9, 3.8 and 11.5 per cent, respectively (Table 1). Technological change was greater than one for all the states, suggesting that all states benefited from production enhancing techniques. In Andhra Pradesh and Maharashtra, the growth in TFP has been driven entirely by technological change with no corresponding changes in efficiency. For Karnataka and Madhya Pradesh, mean technical efficiency change exceeds one, indicating there is greater output from given inputs. However, Rajasthan and Tamil Nadu records diminished efficiency, but overall productivity growth was due to advancement in technological capacity.

Decomposing efficiency changes we find that changes have been driven solely by changes in scale efficiency with pure efficiency remaining unchanged. For the country as a whole total factor productivity growth stood at 6.4% and is driven by changes in technology. However, the decline in efficiency for the country as a whole is worrisome.

Table 1: Malmquist productivity index for sorghum and its decomposition (state-wise)

State	Pure efficiency change	Scale efficiency change	Overall Efficiency change	Technological change	TFP change
Andhra Pradesh	1.000	1.000	1.000	1.042	1.042
Karnataka	1.000	1.009	1.009	1.052	1.061
Madhya Pradesh	1.000	1.005	1.005	1.067	1.073
Maharashtra	1.000	1.000	1.000	1.059	1.059
Rajasthan	1.000	0.973	0.973	1.067	1.038
Tamil Nadu	1.000	0.937	0.937	1.189	1.115
Mean	1.000	0.987	0.987	1.078	1.064

Source: Author's own estimation.

REFERENCES

- Aigner, D., C.A.K. Lovell and P. Schmidt (1977), Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6 (1977): 21-37.
- Antle, J. M. and S. M. Capalbo (1988), An introduction to recent developments in production theory and productivity measurement. *Agricultural productivity: Measurement and explanation*, 17-95.
- Caves, D. W., L. R. Christensen and W. E. Diewert (1982), The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity. *Econometrica*, 50(6): 1393–1414.
- Coelli, T. J., D. S. Prasada Rao, C. J. O'Donnell and G. E. Battese (2005), An Introduction to Productivity and Efficiency Analysis, Second Edition. USA: Springer.
- Färe, R., S. Grosskopf and C. A. K. Lovell (1994), Production frontiers; Cambridge University Press.
- Färe, R., S. Grosskopf, B. Lindgren and P. Roos (1994), Productivity developments in Swedish hospitals: a Malmquist output index approach. In Data envelopment analysis: theory, methodology, and applications, 253-272. Springer, Dordrecht.
- Kathuria, V., R. S. N. Raj and K. Sen (2013), Productivity measurement in Indian manufacturing: A comparison of alternative methods. *Journal of Quantitative Economics*, 11: (1&2) (Combined).
- Shephard, R. W. (1970), The Theory of Cost and Production Functions, Princeton University Press, Princeton.

APPENDIX I

To calculate technical change and efficiency change we compute distance functions for the following type: for each state, $k = 1, \dots, K$ and period $t = 1, \dots, T$,

$$[D(x^{k',t}, y^{k',t}) = \max_{(\theta, z)} \theta$$

$$s. t. \theta y_m^{k',t} \leq \sum_{k=1}^K z^{k',t} y_m^{k',t}, m = 1, \dots, M$$

$$x_n^{k',t} \geq \sum_{k=1}^K z^{k',t} x_n^{k',t}, n = 1, \dots, N$$

$$z^{k',t} \geq 0, k = 1, \dots, K$$

where y is output (in our case a scalar, i.e. $M=1$), and x_n is the vector of nonspillover inputs.

The z 's and the u are variables for which we solve. The z 's serve the purpose of constructing the reference technology as convex combinations of the data. The inequalities allow for the usual assumption of strong (or free) disposability of outputs and inputs. The other three components are calculated similarly, substituting the appropriate period data (i.e. t or $t + 1$)

APPENDIX II

Instruction file for estimation of TFP using DEAP 2.1 software.

```
jowar.txt    DATA FILE NAME
jouarout.txt OUTPUT FILE NAME
6           NUMBER OF FIRMS
12          NUMBER OF TIME PERIODS
1           NUMBER OF OUTPUTS
5           NUMBER OF INPUTS
1           0=INPUT AND 1=OUTPUT ORIENTATED
0           0=CRS AND 1=VRS
2           0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA, 3=DEA(1-STAGE),
            4=DEA(2-STAGE)
```

Chapter 26

TOTAL FACTOR PRODUCTIVITY USING STOCHASTIC FRONTIER PRODUCTION FUNCTION

Shiv Kumar, Abdulla and Deepak Singh

INTRODUCTION

This chapter provides total factor productivity analysis using the stochastic frontier approach. Basically the frontier approach can be classified into two methods i.e. parametric and non-parametric approach. The parametric approach being more powerful to non-parametric is always preferred if the assumptions are fulfilled. There are two main frontier approaches to measure the productivity of any firm or industry; first one is Data Envelopment analysis which is a non-parametric approach and second one is stochastic frontier approach which is parametric in nature. The parametric type of frontier has advantages with respect to other alternatives, for example the deterministic frontiers are based on the assumption that the only type of explanation for the deviation between the observed output and its frontier output is due to its own inefficiency. This idea is difficult to maintain at the empirical level due to it ignores the possibility that the observed output can differ from the potential because of two other factors: stochastic shocks and measurement error in the variables. Further, the mathematical programming methods have two disadvantages with respect to specifying a statistical relationship between the outputs and the inputs. On the one hand, the frontier estimation is made over a subsample of the whole and then these methods are extremely sensitive to the existence of outliers. On the other hand, the estimated coefficients lack statistical properties, so it is not possible to make any statistical inference or establish hypothesis contrasts from them. The stochastic frontier production function for unbalanced panel data given by (Battese and Coelli, 1992).

STOCHASTIC FRONTIER MODEL

There are two frontier approach to measure the efficiency and productivity of any firm or industry; first one is Data Envelopment analysis which is a non-parametric approach (having no assumptions) and second one is stochastic frontier approach which is parametric (having assumption that the distribution of the population is known) in nature. The parametric approach being more powerful have advantages if its assumptions are fulfilled over non- parametric approach. The parametric type of frontier and the computation method present advantages with respect other alternatives, for example the deterministic frontiers .The stochastic frontier production function for panel data given by (Battese and Coelli, 1992).

The stochastic frontier production function proposed, has firm effects that are assumed to be distributed as truncated normal random variables and, also, are permitted to vary systematically with time. The model may be expressed as:

$$z_{it} = f(m_{it}, \beta) \exp(-U_{it}) \dots \quad (1)$$

Where, $i=1 \dots N$, and $t=1 \dots T$, Z_{it} represents the production of the i^{th} firm in the t^{th} time period m_{it} represents the input quantities i^{th} inputs; β_j stands for the output elasticity with respect to the j^{th} input; the V_{it} is a random variable which is assumed to be iid $N \sim (0, \sigma_{v2}^2)$, and distributed independently of the U_{it} which has the specification:

$$U_{it} = U_i \eta_{it} = U_i \exp(-\eta(t - T_i)) \dots \quad (2)$$

Where the distribution of U_i is taken to be the non-negative truncation of the normal distribution, $N \sim (0, \sigma_u^2)$, and η is a parameter that represents the rate of change in technical inefficiency.

Battese and Broca, (1997) who replace σ_v^2 and σ_u^2 with $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\gamma = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2}$. The values of the parameter γ must lie between zero to one. If η greater than zero, then the level of inefficiency decays towards. If η less than zero, then the level of inefficiency increases, and if η equal to zero, then the level of inefficiency remains constant.

The estimation of a time-varying Stochastic Frontier model given by Kumbhakar (1990) specified as

$$U_{it} = U_i [1 + \exp(\gamma t + \delta t^2)]^{-1} \dots \quad (3)$$

This model also contains only two extra parameters to be estimated, γ and δ and the hypothesis of time varying technical efficiency can be tested by $\gamma = \delta = 0$.

There are an extensive literature of time-varying inefficiency models including the (Battese and Coelli, 1992, 1995; Coelli, Prasada Rao, O'Donnell, and Battese, 2005; Cornwell, Schmidt, and Sickles, 1990; S. C. Kumbhakar, 1990). The true fixed effect (TFE) and true random effects (TRE) models developed by (Greene, a2005; Greene, b2005).

In this study we have utilized the stochastic frontier production function for unbalanced panel data given by (Battese & Coelli, 1992).

Decomposition of TFP

The stochastic frontier production function is defined above in equation (1) where Z_{it} is the output of the i^{th} firm in the t^{th} time period. Now taking total differentials m_t with respect to time to time we get

$$\frac{d \ln f(m, t)}{dt} = \frac{\partial \ln f(m, t)}{\partial t} + \sum_j \frac{\partial \ln f(m, t)}{\partial m_j} \frac{dm_j}{dt} \dots \quad (4)$$

In the above equation (4) the first and second terms on the right-hand side are the output elasticity of frontier output with respect to time, TP defined as Technological

Progress, the second term measures the input growth weighted by output elasticities with respect to inputs can be express as $\sum_j \varepsilon_j \dot{m}_j$. The rate of changes defined by placing a dot above a variable.

$$\frac{d \ln f(m,t)}{dt} = TP + \sum_j \varepsilon_j \dot{m}_j \dots \quad (5)$$

Now the total differentiating the equation (1) with respect to time and by using equation (4) the equation become

$$\dot{z} = \frac{\partial \ln f(m,t)}{\partial t} - \frac{du}{dt} = TP + \sum_j \varepsilon_j \dot{m}_j - \frac{du}{dt} \dots \quad (6)$$

We know that TP not only affect the overall productivity but also technical inefficiency change. If TP positive the production frontier shifts upward and vice-versa. The rate at which inefficient catch-up to the production frontier expressed as $-\frac{du}{dt}$.

TFP growth can be defined in terms of technical change (TP) and technical efficiency change.

$$T\dot{F}P = \dot{z} - \sum_j S_j \dot{m}_j \dots \quad (7)$$

S_j explains as input J's Share in production costs.

By substituting equation (6) in to equation (7), equation (7) can be re written as

$$\begin{aligned} T\dot{F}P &= TP - \frac{du}{dt} + \sum_j (\varepsilon_j - S_j) \dot{m}_j \\ &= TP - \frac{du}{dt} + (RTS - 1) \sum_j \lambda_j \dot{m}_j + \sum_j (\lambda_j - S_j) \dot{m}_j \dots \end{aligned} \quad (8)$$

Where, $RTS = \sum_j \varepsilon_j$ is the returns to scale and $\lambda_j = \frac{f_j m_j}{\sum_l f_l m_l} = \frac{\varepsilon_j}{\sum_l \varepsilon_l} = \frac{\varepsilon_j}{RTS}$ the part in the above equation measures inefficiency in the allocation of resources that is the deviation of input prices from their marginal product value. This decomposition formula drawn from the (Kumbhakar, 1990, 1991; Kumbhakar, Wang and Horncastle, 2015a). Similarly used in the (Abdulla, 2018; Kim and Han, 2001; Minh *et al.*, 2012).

If technical inefficiency change does not exist or is time-invariant, the above decomposition suggests that technical inefficiency does not any impact on TFP growth. Then it will become as the Solow residual approach.

ILLUSTRATION

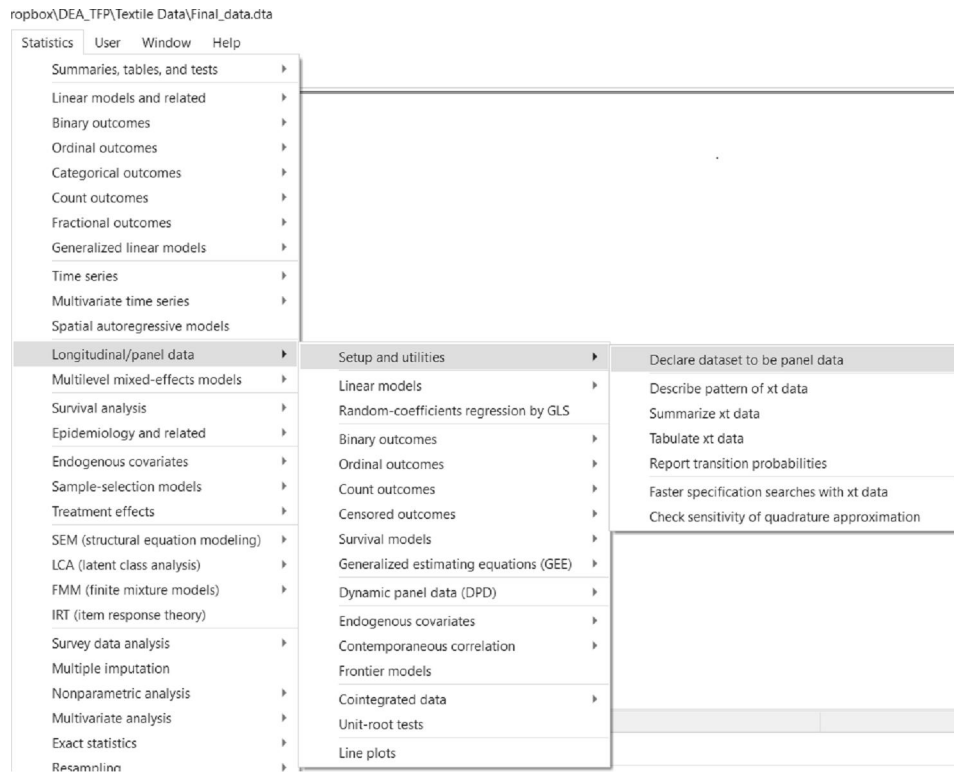
Steps in Calculate the TFP using STATA Software

The steps calculating the TFP using Stata software following the work of (Abatania *et al.*, 2012; Belotti and Ilardi, 2014; Coelli *et al.*, 2005; Kumbhakar *et al.*, 2015b). The data used in the study assessed from the STATA example data (xtfrontier1). In this example, firms produce a product called an output and two inputs viz. labour and capital, assuming that scale technology is constant for all firms. For the estimation of TFP following steps should be followed:

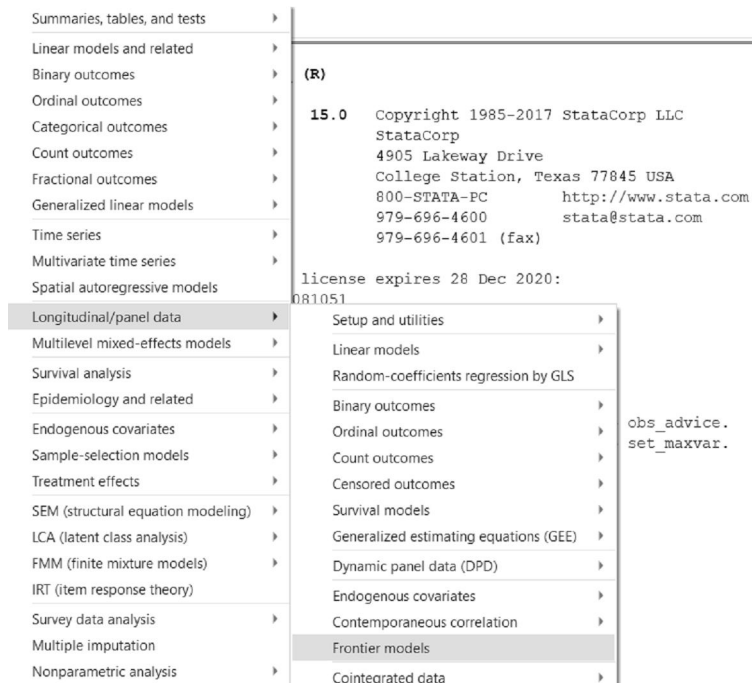
Total Factor Productivity Using Stochastic Frontier Production Function

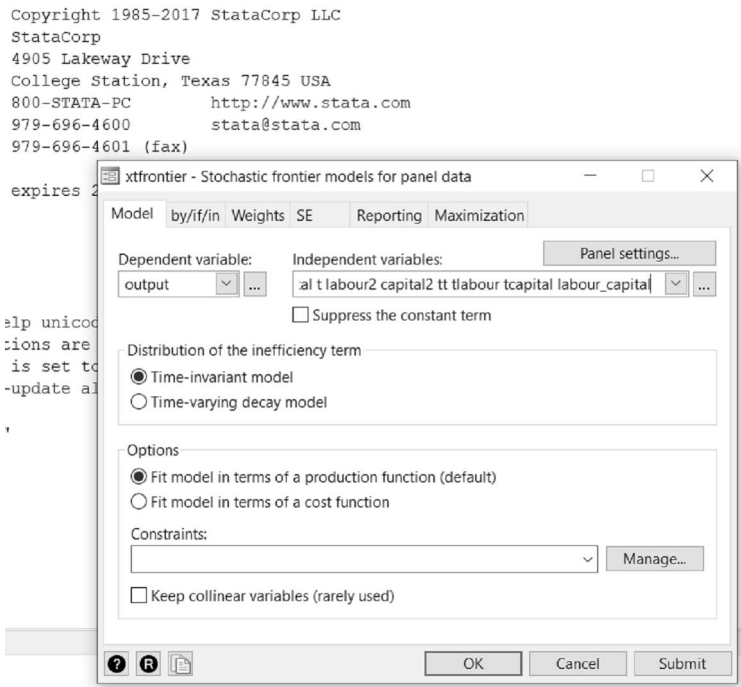
For the estimation of TFP following steps are followed:

Step 1 Declare your data set as Panel data



Step 2 Statistics → Panel data → Frontier models



Step 3 Enter dependent and independent variables

The next step would be selection of time invariant model, then run to the time varying model with the given input output variables followed by LR test to know the appropriate model. For our analysis, we have used the production function approach, however for cost function approach one need to select option fit model in terms of cost function for future analysis. Following figure shows the details of output results.

Step 4 Output Results of the Frontier model shows

```
. xtfrontier output labour capital t labour2 capital2 tt labour tcapital labour_capital, ti
```

```
Iteration 0:  log likelihood = -1469.9058
Iteration 1:  log likelihood = -1469.1431
Iteration 2:  log likelihood = -1468.5717
Iteration 3:  log likelihood = -1468.5647
Iteration 4:  log likelihood = -1468.5647
```

```
Time-invariant inefficiency model
Group variable: id
```

```
Number of obs   =    948
Number of groups =    91
```

```
Obs per group:
    min =     6
    avg =   10.4
    max =    14
```

```
Log likelihood = -1468.5647
```

```
Wald chi2(9)    =   676.51
Prob > chi2     =   0.0000
```

output	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
labour	.2774262	.0394098	7.04	0.000	.2001843 .354668
capital	.1993876	.0426116	4.68	0.000	.1158704 .2829049
t	.0351539	.0371561	0.95	0.344	-.0376706 .1079784
labour2	-.0014112	.0035302	-0.40	0.689	-.0083302 .0055078
capital2	-.0074052	.0052769	-1.40	0.161	-.0177478 .0029374
tt	-.0022016	.002676	-0.82	0.411	-.0074466 .0030433
tlabour	.0017893	.0042241	0.42	0.672	-.0064898 .0100684
tcapital	.0089199	.0046062	1.94	0.053	-.000108 .0179478
labour_capital	.0033683	.0076185	0.44	0.658	-.0115637 .0183002
_cons	2.922886	.1837903	15.90	0.000	2.562664 3.283109

The coefficients of the variables labour and capital are positive and significant shows that there is a direct relationship in output and inputs.

Step 5 Calculate efficiency Statistics → postestimation → predict

The screenshot shows the Stata software interface. On the left, the 'Statistics' menu is open, and 'Postestimation' is selected. On the right, the output window displays the results of the postestimation commands. Below the menu, the 'Postestimation Selector' dialog box is open, showing the 'Predictions' section. The 'Linear predictions and their SEs, technical efficiency' option is selected. The 'predict - Prediction after estimation' dialog box is also open, showing the 'Main' tab. The 'New variable name' is 'Efficiency' and the 'New variable type' is 'float'. The 'Produce' section has four options, with 'Technical efficiency or cost efficiency when -cost- is specified' selected.

Statistics User Window Help

- Summaries, tables, and tests
- Linear models and related
- Binary outcomes
- Ordinal outcomes
- Categorical outcomes
- Count outcomes
- Fractional outcomes
- Generalized linear models
- Time series
- Multivariate time series
- Spatial autoregressive models
- Longitudinal/panel data
- Multilevel mixed-effects models
- Survival analysis
- Epidemiology and related
- Endogenous covariates
- Sample-selection models
- Treatment effects
- SEM (structural equation modeling)
- LCA (latent class analysis)
- FMM (finite mixture models)
- IRT (item response theory)
- Survey data analysis
- Multiple imputation
- Nonparametric analysis
- Multivariate analysis
- Exact statistics
- Resampling
- Power and sample size
- Bayesian analysis
- Postestimation
- Other

Model = -167.73906

Number of obs = 152
Number of groups = 7
Obs per group: min = 5, avg = 21.7, max = 37
Wald chi2(2) = 344.06
Prob > chi2 = 0.0000

Ed.	Err.	Z	P> z	[95% Conf. Interval]
0527086	4.62	0.000	.140385	.3469989
0564333	11.36	0.000	.5304926	.7517071
8117781	5.61	0.000	2.959696	6.141808
.543718	-0.00	0.998	-3.030287	3.020975
6662851	-0.53	0.594	-1.661282	.9505073
.318086	-0.39	0.693	-5.458093	3.628636
4670004			.1898953	2.587022
4733954			.0042436	.9741344
4639091			-.7087639	1.109726
0584592			.3858426	.6149984

Postestimation Selector

Postestimation commands:

- Marginal analysis
- Tests, contrasts, and comparisons of parameter estimates
- Specification, diagnostic, and goodness-of-fit analysis
- Predictions
 - Linear predictions and their SEs, technical efficiency
 - Nonlinear predictions of other predictions, parameters, dynamic forecasts and simulations
- Other reports
- Manage estimation results

predict - Prediction after estimation

Main if/in

New variable name: Efficiency

New variable type: float

Produce:

- ☐ Linear prediction (a + B*x[i, t])
- ☐ Standard error of the linear prediction
- ☐ Technical inefficiency component via E(u[i] | e[i])
- ☐ Technical inefficiency component via M(u[i] | e[i])
- ☒ Technical efficiency or cost efficiency when -cost- is specified

OK Cancel Submit

The predicted technical efficiency will be added in your dataset with a name of efficiency. For predicting technical inefficiency, we can select the option technical inefficiency component via E [u_i|e_i].

Step 6 Calculations of various components of total factor productivity

Calculate Efficiency

predict efficiency, te

Computation of several components of TFP

sort id t

egen tag = tag(id)

* Generate Share of Inputs

generate S1 = labour/total_cost

generate S2 = capital/total_cost

Generate dot of inputs

by id: generate dot_labour = (labour-labour[_n-1])/.5*(labour+labour[_n-1]))

by id: generate dot_capital = (capital-capital[_n-1])/.5*(capital+capital[_n-1]))

*Scale Change

generate eta_labour = _b[labour] + _b[labour2]*labour + _b[labour_capital]*capital
+ _b[tlabour]*t

generate eta_capital = _b[capital] + _b[capital2]*capital + _b[labour_capital]*labour
+ _b[tcapital]*t

*Return to Scale (RTS)

generate RTS = eta_labour + eta_capital

*Calculation of lambda

generate lambda_labour = eta_labour / RTS

generate lambda_capital = eta_capital / RTS

generate sum_lambda_dotx = (lambda_labour)*dot_labour + (lambda_capital)*dot_capital

generate scale = (RTS - 1) * sum_lambda_dotx

*technical change

generate TC = _b[t]+ _b[tt]*t + _b[tlabour]*labour + _b[tcapital]*capital if tag ~= 1

*price change

generate price_effect = (lambda_labour - S1)*dot_labour + (lambda_capital - S2)*dot_capital

*Total Factor Productivity Change

generate TFP = scale + TC + price_effect

After calculating the various components of total factor productivity. The variable with name of scale, TC, price_effect and TFP added in the Stata dataset. The average technical change, price change, scale change and total factor productivity shown is given below.

. sum scale TC price_effect TFP					
Variable	Obs	Mean	Std. Dev.	Min	Max
scale	857	.2115176	20.13811	-239.5815	515.7734
TC	857	.0075795	.0207351	-.102485	.0470682
price_effect	857	-2.958861	58.81452	-791.8989	980.85
TFP	857	-2.739764	59.63594	-978.1621	1061.302

The total factor productivity calculated by using stochastic frontier approach. The unbalanced panel data of 99 firms used in the analysis over the period of 14 years. The total factor productivity growth is the sum of scale change, Technical change and allocative change as shown in equation (8). The average scale change is 21 percent, technical change is about 1 percent and the price change and total factor productivity growth is negative (-295) and (-273) percent respectively.

Commands

```
*Declare data set into panel data
xtset id t, yearly
*Run frontier model
xtfrontier output labour capital t tt labour2 capital2 labour_capital tlabour tcapital, ti
OR
xtfrontier output labour capital t tt labour2 capital2 labour_capital tlabour tcapital,
tv

*Predict technical efficiency
predict efficiency, te
* Compute several components
sort id t
egen tag = tag(id)
* Generate S
generate S1 = labour/total_cost
generate S2 = capital/total_cost
* Generate dotx
by id: generate dot_labour = (labour-labour[_n-1])/.5*(labour+labour[_n-1])
by id: generate dot_capital = (capital-capital[_n-1])/.5*(capital+capital[_n-1])
*Scale Change
generate eta_labour = _b[labour] + _b[labour2]*labour + _b[labour_capital]*capital
+ _b[tlabour]*t
generate eta_capital = _b[capital] + _b[capital2]*capital + _b[labour_capital]*labour
+ _b[tcapital]*t
*RTS
generate RTS = eta_labour + eta_capital
*lambda
generate lambda_labour = eta_labour / RTS
generate lambda_capital = eta_capital / RTS
generate sum_lambda_dotx = (lambda_labour)*dot_labour + (lambda_capital)*dot_
capital
generate scale = (RTS - 1) * sum_lambda_dotx
*Technical Change
generate TC = _b[t]+ _b[tt]*t + _b[tlabour]*labour + _b[tcapital]*capital if tag ~= 1
```

*Price Change

generate price_change = (lambda_labour - S1)*dot_labour + (lambda_capital - S2)*dot_capital

*TFP

generate TFP = scale + TC + price_effect

If technical inefficiency change exist, then

* TEC

generate TEC = coefficient of η *efficiency

TFP

generate TFP = scale + TC + price_change +TEC

REFERENCES

- Abatania, L., A. Hailu, A. Mugera, D. Aigner, C. A. K. Lovell, P. Schmidt, Y. Zhou (2012), Stochastic frontier analysis using Stata. Food Policy. <https://doi.org/10.5539/jas.v4n1p154>
- Abdulla (2018), Performance of Sugar Mills in India: A Comparative Study of Maharastra and Uttar Pradesh. Aligarh Muslum university, Aligarh.
- Battase, G. and S. Broca (1997), Functional forms of stochastic frontier production functions and models for technical inefficiency effects : A Comparative Study for Wheat Farmers in. *Journal of Productivity Analysis*, 8: 395–414. <https://doi.org/10.1023/A:1007736025686>
- Battese, G. E., and T. J. Coelli (1995), A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20(2); 325–332. <https://doi.org/10.1007/BF01205442>
- Belotti, F. and G. Ilardi (2014), Consistent estimation of the fixed-effects SF model sftfe: A Stata command for fixed-effects stochastic frontier models estimation.
- Coelli, T. J., D. S. Prasada Rao, C. J. O'Donnell and G. E. Battese (2005), An introduction to efficiency and productivity analysis. In *An Introduction to Efficiency and Productivity Analysis*. <https://doi.org/10.1007/b136381>
- Cornwell, C., P. Schmidt and R. C. Sickles (1990), Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics*, 46(1–2); 185–200. [https://doi.org/10.1016/0304-4076\(90\)90054-W](https://doi.org/10.1016/0304-4076(90)90054-W)
- Greene, Willam (2005), Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis*, 23(1); 7–32. <https://doi.org/10.1007/s11123-004-8545-1>
- Greene, William (2005), Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, 126(2); 269–303. <https://doi.org/10.1016/j.jeconom.2004.05.003>
- Kim, S. and G. Han (2001), A decomposition of total factor productivity growth in korean manufacturing industries: A Stochastic Frontier Approach.

- Journal of Productivity Analysis*, 16(3): 269–281. <https://doi.org/10.1023/A:1012566812232>
- Kumbhakar, S. C. (1990), Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics*, 46(1–2); 201–211. [https://doi.org/10.1016/0304-4076\(90\)90055-X](https://doi.org/10.1016/0304-4076(90)90055-X)
- Kumbhakar, S. C. (1991), Estimation of technical inefficiency in panel data models with firm- and time-specific effects. *Economics Letters*, 36(1): 43–48. [https://doi.org/10.1016/0165-1765\(91\)90053-N](https://doi.org/10.1016/0165-1765(91)90053-N)
- Kumbhakar, S., H. J. Wang and A. Horncastle (2015a), A Practitioner’s Guide to Stochastic Frontier Analysis Using Stata. <https://doi.org/10.1017/CBO9781139342070>
- Kumbhakar, S., H. J. Wang and A. Horncastle (2015b), A Practitioner’s Guide to Stochastic Frontier Analysis Using Stata. Cambridge University Press.
- Minh, N. K., P. Van Khanh, N. T. Minh and N. T. P. Anh (2012), Productivity Growth, Technological Progress, and Efficiency Change in Vietnamese Manufacturing Industries: A Stochastic Frontier Approach. *Open Journal of Statistics*, 2(02): 224. <https://doi.org/10.4236/ojs.2012.22028>

PART VI

OTHER METHODS

Chapter 27

LINEAR PROGRAMMING: CONCEPT AND ITS APPLICATION IN AGRICULTURE

Harish Kumar H. V. , Rajesh T., Shivaswamy G. P. and Anuja A. R.

INTRODUCTION

Linear programming (LP) is a mathematical modeling technique designed to optimize (maximize or minimize) the usage of limited resources (Hadley, 1997). To define “LP is a mathematical technique of studying wherein we consider maximization (or minimization) of a linear expression (called the objective function) subjected to a number of linear equalities and inequalities (called linear restrictions)”.

History and Application of LP

In 1939, during World War II, a Soviet economist Leonid Kantorovich used LP to plan expenditures and returns in order to reduce costs of the army and to increase losses incurred to the enemy.

LP has wide applications in various fields like military, industry, agriculture, transportation, health system, economics and behavioral sciences etc., and is also utilized for some engineering problems. Transportation, energy, telecommunications, and manufacturing are the major industries that use linear programming models. LP has proven useful in modeling diverse types of problems in planning, routing, scheduling, assignment, and design.

Assumptions of Linear Programming

1. The LP models are “*deterministic*” in nature: Assumes everything is certain and equation is mathematical in nature.
2. The LP models are “*proportional*” in nature: This condition follows directly from linearity assumptions for objective function and constraints. This means that the objective function and constraints expand and contract proportionately to the level of each activity. This condition represents constant returns to scale rather than economies or diseconomies of scale.
3. The LP models are “*additive*” in nature: That is the Left Hand Side (LHS) should be equal to Right Hand Side (RHS). The assumption of proportionality guarantees linearity if and only if the joint effects or interactions are non-existent. That means the total contribution of all activities is identical to sum of the constraints per each activity individually.
4. The decision variables are “*divisible*”: That is the fractional levels for decision variables are permissible, the objective function and constraints are continuous function.

5. Non-negativity: The value of variables must be zero or positive but not negative.

So LP is a special case of mathematical programming to achieve the best outcome (such as maximum profit or minimum cost) in a mathematical model whose requirements are represented by linear relationships (Dorfman, 1996).

Example

Reddy Mikks (R-M) company produce both interior and exterior paints from two raw materials M_1 and M_2 (Taha, 2007). The following Table 1 provides the basic data of the problem.

Table 1: Data for the illustration

Particulars	Tonnes of raw material required per tonne of		Maximum availability with R-M (tonnes)
	Exterior paint	Interior paint	
Raw material M_1	6	4	24
Raw material M_2	1	2	6
Profit per tonne (\$ 000's)	5	4	-

- The market survey restricts maximum daily demand of interior paints to 2 tonnes.
- Additionally the daily demand for interior paint cannot exceed that of exterior paint more than 1 tonne.
- The R-M company wants to determine the optimum product mix of interior and exterior paints that maximizes total daily profit.

Let us formulate the problem

LP model includes three basic elements

- 1) **Decision variables** that we seek to determine
 X_1 = Production of exterior paints (in tonnes)
 X_2 = Production of interior paints (in tonnes)
 - 2) **Objective function** that we aim to optimize
 Main objective is to maximize total daily profit
 Let Z represents total daily profit (in \$ 000's)

$$\text{Max } Z = 5X_1 + 4X_2$$
 - 3) **The constraints** that we need to satisfy
 - a) Restriction on raw material usage: The usage of raw material for production of both paints should not exceed raw material availability.
 - Usage of raw material M_1 : $6X_1 + 4X_2 \leq 24$
 - Usage of raw material M_2 : $X_1 + 2X_2 \leq 6$
 - b) Demand restrictions
- **Maximum daily demand of interior paint is limited to 2 tonnes: $X_2 \leq 2$**

- **Excess of daily production** (daily demand for interior paint cannot exceed that of exterior paint more than 1 tonne): $X_2 - X_1 \leq 1$

So the LP model for above optimization problem looks like below

$$\text{Max } Z = 5X_1 + 4X_2$$

Subject to,

$$6X_1 + 4X_2 \leq 24$$

$$X_1 + 2X_2 \leq 6$$

$$X_2 \leq 2$$

$$X_2 - X_1 \leq 1$$

$$X_1 \text{ \& } X_2 \geq 0$$

STANDARD FORM OF LP MODEL

To solve LP problem manually it must be put in a common form which we call as standard form.

Properties of standard form

- **All the constraints should be expressed as equations by adding slack or surplus and or artificial variables.**

A constraint of the type $\leq (\geq)$ can be converted to an equation by adding **slack** variable to (subtracting **surplus** variable from) the left side of the constraint.

Ex 1: $3X_1 + 2X_2 \leq 6$

$3X_1 + 2X_2 + S_1 = 6$, where S_1 is a slack variable represents the unused amount of resources

Ex 2: $2X_1 + X_2 \geq 6$

$2X_1 + X_2 - S_2 = 6$, where S_2 is a surplus variable represents the excess amount of resources

Note: The introduction of slack and surplus variables alters neither the nature of the constraint nor the objective function. Accordingly such variables are incorporated into objective function with zero coefficient.

- **The right hand side of each constraint should be made non-negative (if not).**

The RHS of the equation can always be made non-negative by multiplying both the sides by -1.

Ex: $2X_1 + 3X_2 - 7X_3 = -5$ can be written as $-2X_1 - 3X_2 + 7X_3 = 5$

$2X_1 - X_2 \leq -5$ can be written as $-2X_1 + X_2 \geq 5$ (the direction of inequality is reversed when both sides are multiplied by -1).

- **The objective function must be maximization type**

Solving of maximization problem is easier than solving of minimization problem. So we can convert minimization form to maximization form for easy calculation and later we can interpret it as minimization solution. The maximization of a function is equivalent to minimization of a negative of the same function and vice-versa.

For a given set of constraints,

Max $Z=5X_1+2X_2+3X_3$ is mathematically **equivalent** to Min $(-Z) = -5X_1-2X_2-3X_3$.

Equivalence means that for the same set of constraints the optimal values of X_1 , X_2 and X_3 are the same in both cases. The only difference is that the values of the objective function, although equal numerically, will appear with opposite signs

Table 2: General and standard form of LP model involving only less than or equal constraints (\leq)

General form	Standard form
$\text{Max } Z = 5X_1+4X_2$ subject to; $6X_1+4X_2 \leq 24$ $X_1+2X_2 \leq 6$ $X_2 \leq 2$ $X_2 - X_1 \leq 1$ $X_1 \& X_2 \geq 0$	$\text{Max } Z = 5X_1+4X_2+0S_1+0S_2+0S_3+0S_4$ subject to; $6X_1+4X_2 + S_1 = 24$ $X_1+2X_2 + S_2 = 6$ $X_2 + S_3 = 2$ $X_2 - X_1 + S_4 = 1$ $X_1, X_2, S_1, S_2, S_3 \& S_4 \geq 0$

Artificial Variable (AV)

In case of problems with infeasible solution artificially we introduce a variable into objective function to obtain feasible solution. We use AV only to start solution and subsequently force them to be zero in the solution otherwise the resulting solution will be infeasible. To guarantee such assignments in the optimal solution, AVs are incorporated into objective function with very large positive co-efficient in minimization problem or very large negative co-efficient in maximization problem.

AVs do change the nature of constraint since they are added only to one side of inequality. That is if the original constraint is an equation ($=$) or of the type greater than or equal to (\geq), then we have no longer basic starting feasible solution.

Table 3: General and standard form of LP model involving all kind of constraints ($\leq, =, \geq$)

General form	Standard form
$\text{Max } Z = 5X_1+2X_2$ subject to; $6X_1+X_2 \leq 6$ $4X_1+3X_2 \geq 12$ $X_1+X_2 = 1$ $X_1 \& X_2 \geq 0$	$\text{Max } Z = 5X_1+2X_2+0S_1+0S_2-MA_1-MA_2$ subject to; $6X_1+X_2 + S_1 = 6$ $4X_1+3X_2 - S_2 + A_1 = 12$ $X_1+X_2 + A_2 = 1$ $X_1, X_2, S_1, S_2, A_1 \& A_2 \geq 0$

Solution to LP Problem

There are two approaches for solving LP problems.

- 1) Graphical approach and
- 2) Simplex technique

1) Graphical approach: LP problems which involve only two decision variables can be solved graphically. Since it is not possible to display the set of feasible solution for more than two variables in a graph for locating best optimal solution. There are two graphical solution methods namely, extreme point solution method and iso-profit (Cost) function line method. Of these, extreme point solution method is most commonly used method for solving LP problem involving two decision variables.

Extreme point solution method: Extreme point refers to corner of the feasible region i.e. the point lies at the intersection of two constraint equations. In this method, the co-ordinates of all corner or extreme points of the feasible region are determined and then value of the objective function at each of these points is computed and compared. The co-ordinates of an extreme point where the optimal (maximum or minimum) value of the objective function is found represent the solution of the given LP problem.

Example:

$$\text{Max } Z = 5X_1 + 7X_2$$

Subjected to,

$$\begin{aligned} X_1 &\leq 6 \\ 2X_1 + 3X_2 &\leq 19 \\ X_1 + X_2 &\leq 8 \\ X_1, X_2 &\geq 0 \end{aligned}$$

Solution:

Here we are not going to add any slack or surplus variable but we are just putting it into equation.

$$\begin{aligned} X_1 &= 6 \\ 2X_1 + 3X_2 &= 19 \\ X_1 + X_2 &= 8 \end{aligned}$$

$X_1 + X_2 = 8$ Extreme points: a) $X_1=0$ $X_2=8$, Co-ordinates:(0,8) b) $X_2=0$ $X_1=8$, Co-ordinates:(8,0)	$2X_1 + 3X_2 = 19$ Extreme points: a) $X_1=0$ $X_2=6.33$, Co-ordinates: (0,6.33) b) $X_2=0$ $X_1=9.5$, Co-ordinates: (9.5,0)	$X_1=6$ Extreme points: a) $X_1=6$, $X_2=0$, (6,0)
--	---	--

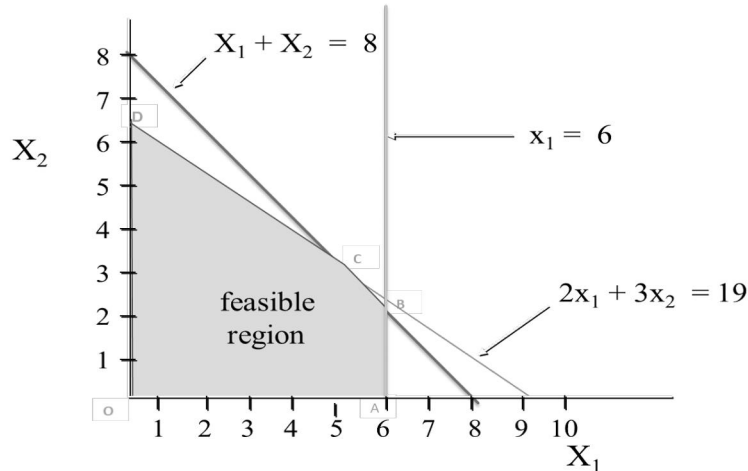


Fig 1: Combined-constraint graph showing feasible region

The shaded zone is called feasible area where all the constraints holds good or this region satisfies all constraints so it is called *feasible region* (Fig 1).

- The corners or vertices of the feasible region are referred to as the extreme points.
- An optimal solution to an LP Maximization problem can be found at an extreme point of the feasible region.
- When looking for the optimal solution, you do not have to evaluate all feasible solution points.
- Consider only the extreme points of the feasible region.

Table 4: Value of objective function at extreme points of feasible region

Extreme Point	Co-Ordinates	Z value ($Z=5X_1+7X_2$)
O	(0,0)	0
A	(6,0)	30
B	(6,2)	44
C	(5,3)	46
D	(0,6.33)	44.31

At point C all constraints are satisfied and the Z value is highest hence it is optimal point.

Solution: At $X_1=5$ and $X_2=3$, Max $Z=46$

2) Simplex method

It is an algorithm adopted to solve LP problem which employs an iterative procedure that starts at a feasible corner point, normally the origin and systematically moves from one feasible point to another point until it reaches optimum point.

Linear programming solvers are now part of many spreadsheet packages, such as Microsoft Excel. The leading commercial package is “LINDO”. We can solve LP problems in packages like “R” and “SAS” also.

Special Cases in Simplex Method of Application

1. Degeneracy

In case of model consisting of at least one redundant (No longer needed or not useful) constraint then the optimum value won't improve upon iterations instead same solution is generated over the iterations.

Example

$$\text{Max } Z = 3X_1 + 9X_2$$

Subject to,

$$X_1 + 4X_2 \leq 8$$

$$X_1 + 2X_2 \leq 4$$

$$X_1 \text{ \& } X_2 \geq 0$$

In above case the first constraint is a redundant constraint.

2. Alternative optima

Alternative optima exists when objective function running parallel to one of the constraints. Then the objective function will assume same optimal value at more than one solution point.

Example

$$\text{Max } Z = 2X_1 + 4X_2$$

Subject to,

$$X_1 + 2X_2 \leq 5$$

$$X_1 + X_2 \leq 4$$

$$X_1 \text{ \& } X_2 \geq 0$$

In above case the objective function runs parallel to first constraint.

3. Unbounded solutions

The solution to a maximization LP problem is unbounded if the value of the solution may be made indefinitely large without violating any of the constraints. Sometimes feasible solution for the given LP problem exists and this has infinite values for the objective function. For real problems, this is the result of improper formulation.

4. Infeasible or non-existent solutions

No unique solution to the LP problem satisfies all the constraints, including the non-negativity conditions. Graphically, this means a feasible region does not exist. Causes includes formulation error, too high expectations by management or too many restrictions have been placed on the problem (i.e. the problem is over-constrained).

INTEGER PROGRAMMING

In case of linear programming, the decision variables considered are supposed to take any real value. However in practical situations it makes no sense in assigning a real value to a variable where it has meaning only when it takes only integer values (Rao, 2007). To be clear let us consider a practical problem like optimum size of herd in a dairy project, it makes no sense if our optimal value from LP solution is 5.8.

In such situations, we naturally tend to round-off the optimal value to the nearest integer value say “6” in above example. However, the round-off may have following fundamental problems,

- a) The round-off solution may not be feasible.
- b) The objective function value given by the rounded-off solutions (even if some are feasible) may not be the optimal one.
- c) Even if some of the rounded-off solutions are optimal, checking all the rounded-off solutions is computationally expensive.

So integer programming deals with the solution of mathematical programming problems in which some or all the variables can assume non-negative integer values only.

Types of integer programming problems

- 1) Pure integer programming problem: An integer programming problem in which all variables are required to be integers.
- 2) Mixed integer programming problem: If some variables are restricted to be integer and some are not restricted i.e. can be continuous or fractional.
- 3) Binary integer programming problem/ 0-1 programming problems: If some or all variables are restricted to be either “0” or “1”. It can be pure or mixed.

The general form of integer programme is as below:

$$\text{Max } Z = 7X_1 + 9X_2$$

subject to,

$$-X_1 + 3X_2 \leq 6$$

$$7X_1 + X_2 \leq 35$$

X_1 & X_2 are non-negative integers.

Applications of Linear Programming in Agriculture

Case-1: Naidu Dairy farm uses at least 800 Kg of *Special feed* daily. The *Special feed* is a mixture of corn silage and soybean meal with the following composition.

Table 5: Constituents of special feed and their unit cost

Feed stuff	In terms of kg per every kg of feed stuff		
	Protein	Fiber	Cost (Rs./Kg)
Corn silage	0.09	0.02	20
Soybean meal	0.60	0.06	62

The dietary requirements of *Special feed* must have at least 30 per cent protein and at most 5 per cent fiber. Now the Naidu Dairy farm wishes to determine the daily minimum cost of feed mix?

Solution:

Decision variables:

X_1 = Quantity of corn silage to be used in feed mix (Kg)

X_2 = Quantity of soybean meal to be used in feed mix (Kg)

Objective function

$$\text{Min } Z = 20X_1 + 62X_2$$

Constraints

Demand constraint (Daily requirement): $X_1 + X_2 \geq 800$

Protein constraint: $0.09X_1 + 0.60X_2 \geq 0.30(X_1 + X_2)$ on simplification $-0.21 X_1 + 0.30 X_2 \geq 0$

Fiber constraint: $0.02X_1 + 0.06X_2 \leq 0.05(X_1 + X_2)$ on simplification $-0.03 X_1 + 0.01 X_2 \leq 0$

Overall the LP model looks like

$$\text{Min } Z = 20X_1 + 62X_2$$

Subject to,

$$X_1 + X_2 \geq 800$$

$$-0.21 X_1 + 0.30 X_2 \geq 0$$

$$-0.03 X_1 + 0.01 X_2 \leq 0$$

$$X_1 \& X_2 \geq 0$$

R code for the above LP problem

```
library(lpSolve)
```

```
obj=c(20,62)
```

```
mat=matrix(c(1,1,-0.21,0.3,-0.03,0.01), nrow=3, byrow=TRUE)
```

```
rhs=c(800,0,0)
```

```

dir=c(">=", ">=", "<=")
prod.sol= lp("min", obj, mat, dir, rhs, compute.sens = TRUE)
prod.sol$status
prod.sol$objval
prod.sol$solution
prod.sol$duals
prod.sol$duals.from
prod.sol$duals.to
prod.sol$sens.coef.from
prod.sol$sens.coef.to

```

A) Optimal solution

Z	29835.29
X1	470.58
X2	329.41

The daily minimum cost of feed mix by using 470.58 Kg of corn silage and 329.41 Kg of soybean meal is Rs. 29835.29.

B) Sensitivity analysis**a) Maximum change in resource availability (RHS of binding constraints)**

Binding Constraint	Shadow price	RHS	Sensitivity (Range)
Special feed	37.29	800	0 to 1×10^{30}
Protein	82.35	0	-168 to 138

b) Maximum change in marginal cost (Co-efficients of DV's in objective function)

Variable	Value of DV's	Unit price	Sensitivity (Range)
Corn silage (X1)	470.58	20	-43.40 to 62.00
Soybean meal (X2)	329.41	62	20 to 1×10^{30}

Case 2: Venkatesh, a Crop+Dairy farming system based farmer wishes to maximize the total revenue with the available resources. The below table provides the information on the resource availability and the information on resource requirement for the enterprises from his past experience. Ragi being the regular diet of Venkatesh's family he needs minimum 1 acre of his land to be under the same which also serves the fodder security of his dairy. Since the dairy is earning him the regular income for family maintenance he insists at least one cross breed (CB) cow in his farming system.

Resources	Availability	Per unit requirement			
		Tomato	Cabbage	Ragi	CB Cow
Land (Acres)	4	-	-	-	-
Labour (Man days)	350	180	65	32	38
Capital (Rs.)	250000	125000	65000	12500	33000
Water (acre inches)	100	24.5	17.8	9.4	0.5
Returns (Rs.)	-	280000	135000	19000	65000

Solution:

Decision variables:

- X_1 = Area under Tomato crop to be taken (Acres)
- X_2 = Area under Cabbage crop to be taken (Acres)
- X_3 = Area under Ragi crop to be taken (Acres)
- X_4 = Number of cross breed cows to be considered in his farming system

Objective function

$$\text{Max } Z = 280000X_1 + 135000X_2 + 19000X_3 + 65000X_4$$

Constraints

- Land constraint (Overall): $X_1 + X_2 + X_3 \leq 4$
- Labour constraint: $180X_1 + 65X_2 + 32X_3 + 38X_4 \leq 350$
- Capital constraint: $125000X_1 + 65000X_2 + 12500X_3 + 33000X_4 \leq 250000$
- Water constraint: $24.5X_1 + 17.8X_2 + 9.4X_3 + 0.5X_4 \leq 100$
- Constraint for Ragi mandate: $X_3 \geq 1$
- Constraint for Dairy mandate: $X_4 \geq 1$

Overall the LP model looks like

$$\text{Max } Z = 280000X_1 + 135000X_2 + 19000X_3 + 65000X_4$$

Subject to,

- $X_1 + X_2 + X_3 \leq 4$
- $180X_1 + 65X_2 + 32X_3 + 38X_4 \leq 350$
- $125000X_1 + 65000X_2 + 12500X_3 + 33000X_4 \leq 250000$
- $24.5X_1 + 17.8X_2 + 9.4X_3 + 0.5X_4 \leq 100$
- $X_3 \geq 1$
- $X_4 \geq 1$
- $X_1 + X_2 + X_3 \geq 0$ & X_4 is a non-negative integer

R code for the above LP problem

```
library(lpSolve)
obj=c(280000,135000,19000,65000)
mat=matrix(c(1,1,1,0,180,65,32,38,125000,65000,12500,33000,24.5,17.8,9.4,
0.5,0,0,1,0,0,0,0,1), nrow=6, byrow=TRUE)
rhs=c(4,350,250000,100,1,1)
dir=c("<=", "<=", "<=", "<=", ">=", ">=")
```

```
prod.sol= lp("max", obj, mat, dir, rhs, int.vec=4, compute.sens = TRUE)
prod.sol$status
prod.sol$objval
prod.sol$solution
prod.sol$duals
prod.sol$duals.from
prod.sol$duals.to
prod.sol$sens.coef.from
prod.sol$sens.coef.to
```

A) Optimal solution

Z	536713.28
X1	1.37
X2	0.50
X3	1
X4	1

The maximum total revenue that the farmer can achieve is Rs. 5,36,713.3/- by cultivating Tomato, Cabbage and Ragi in 1.37, 0.50 and 1 acre respectively, along with 1 CB cow.

B) Sensitivity analysis

a) Maximum change in resource availability (RHS of binding constraints)

Binding Constraint	Shadow price	RHS	Sensitivity (Range)
Labour	370.62	350	283.20 to 364.48
Capital	1.70	250000	239944.4 to 284847.8
Ragi	-14188.81	1	0 to 1.98
CB cow	-5391.60	1	0 to 2.52

b) Maximum change in marginal profit (Co-efficients of DV's in objective function)

Variable	Value of DV's	Unit price	Sensitivity (Range)
Tomato (X1)	1.37	280000	259615.4 to 373846.15
Cabbage (X2)	0.50	135000	118802.5 to 145600
Ragi (X3)	1	19000	-1*10 ³⁰ to 33188.81
CB cow (X4)	1	65000	-1*10 ³⁰ to 70391.61

REFERENCES

- Dorfman, R. (1996), Linear Programming and Economic Annalysis. McGraw-Hill. New York.
- Hadley, G. (1997), Linear programming. Narosa publishing house. New Delhi.
- Rao, S. S. (2007), Engineering Optimization: Theory and Practice. New Age International Publishers. New Delhi.
- Taha, H. A. (2007), Operation Research: In Introduction. Seventh edition. Prentice Hall India. New Delhi. <https://nptel.ac.in/courses/105108127/>

Chapter 28

MULTI-OBJECTIVE PROGRAMMING

Chandra Sen

INTRODUCTION

Multi-objective programming (MOP) is the field of applied mathematics that studies problems of achieving multiple conflicting objectives over a domain of mathematical relations. Multi-objective programming is also termed as multi-objective optimization, vector optimization, multi-criteria optimization, multi-attribute optimization or Pareto optimization. The MOP methods are increasingly used for making decisions in achieving multiple conflicting objectives simultaneously. Many a times, the decision maker fails to make good decisions under a complex situation. A natural explanation for such a situation is that the applied mathematics is not sufficiently realistic to incorporate all the criteria and constraints. Several MOP methods have been developed to help in such a complex situation. These methods are getting more and more importance in applications in engineering, agriculture and management.

Multi-objective programming methods

\mathcal{E} -Constraint method

The method was introduced by Haimes *et al.* (1971). In this method the one objective is selected for max./min. keeping remaining objectives within user specified values. The problem is formulated as:

$$\begin{aligned} & \text{Max./Min. } Z_r(x) \\ \text{Subject to,} \\ & Z_m(x) \leq / \geq / = \mathcal{E}_m \quad n=1, \dots, N \quad r \neq n \\ & a_i x_i \leq / \geq / = b_i \quad i=1, \dots, K \end{aligned}$$

The choice of the objective to be optimized and the values of the remaining objectives kept as constraints are subjective. Hence, the solution may not be more acceptable.

Lexicographic method

In Lexicographic method, the objective functions are arranged in order of importance and each objective is optimized as per its priority. The optimal value of first objective is put as constraint for achieving the second objective. The process ends in optimizing the last objective. The method can be explained as:

$$\text{Max./Min. } Z_r(x)$$

Subject to,

$$\begin{aligned} Z_m(x) &\leq / \geq / = \xi_m \quad n=1, \dots, N \quad r \neq n \\ a_i x_i &\leq / \geq / = b_i \quad i=1, \dots, K \end{aligned}$$

The solution obtained using this method is also not much preferable because of the subjective prioritization the objectives.

Weighted-sum method

The weighted sum method was suggested by Gass and Satty (1955). This is the most commonly used in solving multi-objective optimization problems. The method is described as:

$$\text{Max./ Min. } \sum_{j=1}^k w_j Z_j$$

Subject to,

$$AX = b \quad \text{and } X \geq 0$$

Where w_j is the weight assigned to j^{th} objective function.

It is very easy to formulate and optimize the combined objective function. However the method cannot generate efficient solution always due to following limitations:

- (i) The weights (w_j) assigned to different objectives are subjective.
- (ii) In most of the MOP applications, the objectives are non commensurable. Hence the addition of the values of different dimensions is not logical.
- (iii) The method becomes biased to the dominating objective/s.

Scalarizing method

Realizing the above mentioned problems the scalarizing method for solving MOP problems was proposed by Sen (1982). The method is described as below:

$$\text{Maximize } Z = \frac{\sum_{j=1}^r Z_j}{|\Theta_r|} - \frac{\sum_{j=r+1}^s Z_j}{|\Theta_s|}$$

Subject to,

$$AX = b \quad \text{and } X \geq 0$$

$$\Theta_r \text{ \& } \Theta_s \neq 0$$

Where,

Θ_r and Θ_j is the optimal value of r^{th} and j^{th} objective function and there are 'r' maximization and 's-r' are minimization objective functions.

The method has been successfully applied in resource use planning in agriculture (Sen, 1982; Sen, 1983; Sen and Dubey, 1994; Singh, 2005; Kumar, 2012; Gautam, 2013; Kumari *et al.*, 2017; Maurya, 2018; Gwandi *et al.*, 2019 etc.) for increasing income and employment with lesser use of fertilizer, irrigation, environment pollution etc. The application of the method can be understood with

following example:

Example 1

$$\text{Max. } Z_1 = 7000X_1 + 10000X_2 + 4000X_3 + 5000X_4 + 9000X_5$$

$$\text{Max. } Z_2 = 89X_1 + 50X_2 + 100X_3 + 79X_4 + 95X_5$$

$$\text{Min. } Z_3 = 10X_1 + 15X_2 + 14X_3 + 13X_4 + 12X_5$$

$$\text{Min. } Z_4 = 150X_1 + 120X_2 + 140X_3 + 100X_4 + 130X_5$$

Subject to,

$$X_1 + X_2 + X_3 + X_4 + X_5 = 7 \text{ (Land restriction)}$$

$$X_1 \geq 0.7 \text{ (Home requirement)}$$

$$X_2 \geq 0.3 \text{ (Home requirement)}$$

$$X_3 \geq 0.4 \text{ (Home requirement)}$$

$$X_5 \geq 0.6 \text{ (Home requirement)}$$

where X_1, \dots, X_5 are crops, Z_1 = Income (Rs.), Z_2 = Employment (man days), Z_3 = Plant Protection chemicals (litre) and Z_4 = Fertilizer use (kg.).

Solution

All the objectives have been optimized individually to find the optimal values of each objective. The multi-objective function was formulated as explained above. Finally the scalarized multi-objective function was optimized with the same common constraints. The results are presented in Table 1.

Table 1: Individual and multi-objective optimization

Item	Individual Optimization				Sen's MOP
	Max. Z_1	Max. Z_2	Min. Z_3	Min. Z_4	
X_1	X1=0.7, X2=5.3, X3=0.4, X4 =0, X5=0.6	X1=0.7, X2=0.3, X3=5.4, X4 =0, X5=0.6	X1=5.7, X2=0.3, X3=0.4, X4 =0, X5=0.6	X1=0.7, X2=0.3, X3=0.4, X4 =5.0 X5=0.6	X1=0.7, X2=0.3, X3=0.4, X4 =0, X5=5.6
Z_1	64900	34900	49900	39900	59900
Z_2	424.3	674.3	619.3	569.3	649
Z_3	99.3	94.3	74.3	89.3	84.3
Z_4	875	975	1025	775	925

All the individual optimizations have generated different solutions. This reveals the presence of conflicts among objectives. The maximization of income increased the income Rs. 64900 which is highest in comparison to other individual optimal solutions. The similar pattern have been observed in all the individual optimizations. However the Sen's MOP method has generated a compromising solution achieving all the objective simultaneously. The income can be increased up to Rs. 59900 which is not as high as in individual optimization but better as remaining individual optimizations. The achievements of remaining objective are also superior over individual optimizations.

Averaging method

Several averaging methods for solving MOP problems have been proposed (Sulaiman *et al.*, 2013; Nejamudin *et al.*, 2016; Akhtar *et al.*, 2017; Samsun and Abdul, 2017) during past two decades. The combined objective function was formulated by scalarizing the objective functions by various averages. The averaging methods are formulated as:

$$\begin{aligned} \text{Maximize } Z &= \frac{\sum_{j=1}^r Z_j}{|\Theta_1|} - \frac{\sum_{j=r+1}^s Z_j}{|\Theta_2|} \\ \text{Subject to,} \\ AX &= b \quad \text{and } X \geq 0 \\ \Theta_j &\neq 0 \quad \text{for } j=1, 2, \dots, s. \end{aligned}$$

where

$|\Theta_1|$ = Mean, Geometric Mean and Harmonic Mean of optimal values of maximization objective functions.

$|\Theta_2|$ = Mean, Geometric Mean and Harmonic Mean of optimal values of Minimization Objective functions.

The averaging methods for solving MOP problems are not efficient (Chandra 2018a, Chandra 2018b, Chandra 2018c, Chandra 2019) in providing the appropriate solutions due the limitations with these methods similar to weighted-sum methods. These methods have been applied to solve the MOP problem of example 2.

Example 2

$$\text{Max. } Z_1 = 12500X_1 + 25100X_2 + 16700X_3 + 23300X_4 + 20200X_5$$

$$\text{Max. } Z_2 = 21X_1 + 15X_2 + 13X_3 + 17X_4 + 11X_5$$

$$\text{Min. } Z_3 = 370X_1 + 280X_2 + 350X_3 + 270X_4 + 240X_5$$

$$\text{Min. } Z_4 = 1930X_1 + 1790X_2 + 1520X_3 + 1690X_4 + 1720X_5$$

Subject to,

$$X_1 + X_2 + X_3 + X_4 + X_5 = 4.5$$

$$2X_1 \geq 1.0$$

$$3X_4 \geq 1.5$$

All the objectives of the above example have been optimized individually to examine the presence conflicts among the objectives. The solution of individual optimization is given in Table 2.

Table 2: Individual optimization matrix

Item	Individual Optimization			
	Max. Z_1	Max. Z_2	Min. Z_3	Min. Z_4
X_1	0.5, 3.5, 0, 0.5, 0	4, 0, 0, 0.5, 0	0.5, 0, 0, 0.5, 3.5	0.5, 0, 3.5, 0.5, 0
Z_1	105750	61650	88600	76350
Z_2	71.5	92.5	57.5	64.5
Z_3	1300	1615	1160	1545
Z_4	8075	8565	7830	7130

It is clear from Table 2 that all the four objective have different and conflicting solutions. The above example has been solved using averaging methods. The Mean, Geometric mean and Harmonic mean have been used for scalarizing the objective functions. The results are presented in Table 3.

Table 3: Multi-objective optimization

Item	Existing Averaging Techniques			Improved Averaging Techniques		
	Mean	Harmonic Mean	Geometric Mean	Mean	Harmonic Mean	Geometric Mean
Xi	0.5, 3.5, 0, 0.5, 0	0.5, 3.5, 0, 0.5, 0	0.5, 3.5, 0, 0.5, 0	0.5, 0, 0, 4, 0	0.5, 0, 0, 4, 0	0.5, 0, 0, 4, 0
Z ₁	105750	105750	105750	99450	99450	99450
Z ₂	71.5	71.5	71.5	78.5	78.5	78.5
Z ₃	1300	1300	1300	1265	1265	1265
Z ₄	8075	8075	8075	7725	7725	7725

The results in Table 3 revealed that all the averaging methods have achieved the first objective only and ignored the remaining three objectives. None of these methods optimizes all the objectives simultaneously.

Improved Averaging Method

The problems with existing averaging methods have been realized and improved averaging methods for solving MOP problems have been proposed (Chandra, 2019). The improved averaging method is explained as:

$$\text{Maximize } Z = \frac{\sum_{j=1}^r Z_j}{|\Theta_i|} - \frac{\sum_{j=r+1}^s Z_j}{|\Theta_j|}$$

Subject to,

$$AX = b \quad \text{and } X \geq 0$$

$$\Theta_j \neq 0 \quad \text{for } j=1, 2, \dots, s.$$

where

$|\Theta_i|$ = Mean, Geometric Mean and Harmonic Mean of optimal and suboptimal values of i^{th} Maximization Objective function.

$|\Theta_j|$ = Mean, Geometric Mean and Harmonic Mean of optimal and suboptimal values of J^{th} Minimization Objective function.

The improved averaging methods have also been applied to solve example 2 and the results are presented in Table 3. The improved averaging methods have achieved all the objectives simultaneously. The values of all the objectives are closer to their individual optima. Hence, the improved averaging methods are efficient in solving the MOP problems.

Several methods for solving MOP problems have been discussed in the chapter. There is continuous improvements in the existing methods for obtaining acceptable solutions

for achieving multiple objectives at a time. It can be concluded that improved averaging methods are superior over other available methods for solving MOP problems.

REFERENCES

- Akhtar, H., G. Modi and S. Duraphe (2017), Transforming and optimizing multi-objective quadratic fractional programming problem, *International Journal of Statistics and Applied Mathematics*, 2 (1): 01-05.
- Chandra Sen (2018a), Sen's Multi-Objective Programming method and its comparison with other techniques. *American Journal of Operational Research*, 8 (1): 10-13.
- Chandra Sen (2018b), Correlation technique for solving Multi -Objective Programming (MOP) problems-an evaluation. *International Journal of Mathematics Trends and Technology (IJMTT)*, 60 (3): 187-190.
- Chandra Sen (2018c), Multi objective optimization techniques: misconceptions and clarifications. *International Journal of Scientific and Innovative Mathematical Research*, 6 (6): 29-33.
- Chandra Sen (2019), Improved Scalarizing Techniques for Solving Multi-Objective Optimization Problems. *American Journal of Operational Research*, 9 (1): 8-11.
- Gass, S. and T. Saaty (1955), The computational algorithm for the parametric *objective* function. *Naval Research Logistics Quarterly*, 2: 39.
- Gautam, Kusumakar (2013), Natural and Human Resource Use Planning for Vindhyan Region of Eastern Uttar Pradesh, PhD thesis, Department of Agricultural Economics, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India.
- Gwandi, O., V. Kamalvanshi, M. K. Maurya and Saket Kushuwaha (2019), Farmers' livelihood strategies by optimizing resource use in farming of district Varanasi, Uttar Pradesh: application of Sen's Multi-Objective Programming approach. *Research Journal of Agricultural Sciences*, 10 (1): 116-119.
- Haimes, Y. Y., L. S. Lasdon and D. A. Wismer (1971), On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man and Cybernetics*, 1: 296-297.
- Kumar, Hraday (2012), Economic Analysis of Fresh Water Aquaculture in Maharajganj district of Eastern Uttar Pradesh. Ph.D. thesis. Department of Agricultural Economics, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India.
- Kumari, M., O. P. Singh and D. C. Meena (2017), Optimizing cropping pattern in eastern Uttar Pradesh using sen's Multi-Objective Programming Approach. *Agricultural Economics Research Review*, 30 (2), 285-291.

- Maurya, Mukesh Kumar (2018), Livelihood Security of Farmers in Eastern Uttar Pradesh An Economic Analysis. Ph.D. thesis, Department of Agricultural Economics, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India.
- Nejmaddin, A. S., R. M. Abdullah and Snur O. Abdull (2016), Using optimal geometric average technique to solve extreme point multi-objective quadratic programming problems, *Journal of Zankoy Sulaimani*, 63-72.
- Samsun Nahar and Md. Abdul Alim (2017), A New Statistical Averaging Method to Solve Multi-Objective Linear Programming Problem. *International Journal of Science and Research*, 6(8): 623-629.
- Sen, Chandra and P. P. Dubey (1994), Resource use planning in Agriculture with single and Multi-objective programming approaches (A comparative study). *Journal of Scientific Research*, 44: 75-81.
- Sen, Chandra (1982), Integrated multi period rural development plan for Dwarahat block, Almora (U.P.): A multi-Objective programming Approach. Ph.D. thesis, Department of Agricultural Economics, G.B.P. University of Agriculture and Technology, Pantnagar, Uttrakhand.
- Sen, Chandra (1983), A new approach for multi-objective rural development planning, *The Indian Economic Journal*, 30(4): 91-96.
- Singh, P. K. (2005), Multilevel multi-objective planning for Agriculture for district Mau, U.P. Ph.D. thesis, Dept. of Agricultural Economics, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India.
- Sulaiman, N. A. and B. K. Abdulrahim (2013), Using Transformation Technique to Solve Multi- Objective Linear Fractional Programming Problem. *International Journal of Recent Research and Applied Science*, 14 (3): 559-567.

Chapter 29

STRUCTURAL EQUATION MODELING

P. Sethuraman Sivakumar, N. Sivaramane and P. Adhiguru

INTRODUCTION

One of the important goals of social science research is to accurately explain and predict individual or group behaviour in a specific context. The context may be utilisation of improved varieties by farmers or their choice of using specific pest or disease management strategies, or their decision to choose a specific marketing channel for selling the farm produce. In all these cases, the social researchers need to explain and predict the farmers' decisions through network of factors and their relationships. When these social factors are in both manifest and latent form, which regulate the decision process through direct and indirect effects on a network, specialised estimation techniques are required to explain this process. Structural equation modeling (SEM) is one of the statistical approaches which help the researchers to identify the factors of a specific phenomenon and assess their relationships in a way to predict the outcomes in an objective and valid manner. The SEM is a versatile technique used by social, behavioral, and educational scientists as well as biologists, economists, marketing, and medical researchers to explain a complex phenomenon.

Structural equation modelling is a multivariate method, which combines factor analysis and multiple regressions that enable researcher to simultaneously examine a series of interrelated dependence relationships among the measured variables and latent constructs (Hair Jr *et al.*, 2006). It is also known as 'causal modelling' or 'analysis of covariance structures'.

Simultaneous Equation Model (SIM) and Structural Equation Model (SEM)

The Simultaneous Equation Model (SIM) is composed of a set of linear simultaneous equations which represent a set of relationships among variables or describe joint dependence of variables. The simultaneous equation models have more than one endogenous variable (Y_1, Y_2) which are included in the estimation.

For example, the market for MBA (Agri-Business) professionals is described in terms of demand behaviour (demand from graduates specialised in Agri-business management), supply behaviour (numbers graduated in an academic year), and equilibrium levels of employment and wages. In this process, wages are included in both demand and supply estimations, thereby creating simultaneous or joint determination of the equilibrium quantities.

To estimate the market for the MBA (Agri-business) graduates, the numbers employed (p),

Salary (s) and enrolment (a) are included. An opportunity cost component is included through the median income (m) of Agricultural Officers.

The demand equation (Eq 1) indicates that the demand for Agri-business management graduates is determined by college enrolments and by the salary for Agri-business management graduates.

$$p_t = \beta_{11} + \beta_{12}a_t + \beta_{13}s_t + e_{1t} \dots \quad (1)$$

While the supply equation (Eq 2) indicates that the supply of Agri-business management graduates is determined by the salary (s), the median income of Agricultural Officer (m - opportunity cost for students enrolling in MBA –Agribusiness), and lagged quantity supplied (p-the pool of available Agribusiness graduates that adjusts slowly to market innovations).

$$p_t = \beta_{21} + \beta_{22}m_t + \beta_{23}s_t + \beta_{24}p_{t-1} + e_{2t} \dots \quad (2)$$

The Structural Equation Model is basically simultaneous equation model, which differs from its predecessor in following ways.

1. Structural equation models are complete Simultaneous Equations Models i.e. total number of endogenous variables is equal to the number of equations in the model.
2. The primary variables used in Structural Equation Modelling are usually latent variables, while simultaneous equation models use observed or manifest variables

Econometricians and statisticians mostly prefer SIM than SEM, while marketers, psychologists, sociologists and educationists mostly use SEM.

Constructs in SEM

The Structural Equation Modeling uses latent factors or constructs and interrelationships among them to represent the processes in a social phenomenon. The constructs are special types of concepts, which are deliberately and consciously invented or adopted for a scientific purpose (Kerlinger and Lee, 1999). These constructs are latent or unobservable factors that are abstract and indirectly observable through configuration of multiple observable variables. Few examples of constructs or latent factors in economics include capitalism, food security, impact of technologies, while empowerment, scientific orientation, innovativeness are few popular constructs used in extension. The indirect measurement of constructs involves examining the consistency among multiple measured variables or indicators which are gathered through various data collection methods.

The measured or manifest or observed variables are measured directly and used as indicators for defining the construct. For example, food security is a construct which can't be measured directly. However, the food security can be measured using indicator variables such as local food production and monthly income.

In SEM, the observed variables can be categorical, discrete or continuous, but latent variables must always be continuous variable (Kline, 1999).

The benefits of using constructs or latent factors

Representing theoretical concepts

In general, the theories in behavioural sciences like extension are dealing with concepts which are latent in nature and interrelated with observable concepts in a complex way. While using these unobservable constructs, the researcher often faces the problem of accurately defining them in a way to design appropriate research questions. The SEM helps the researcher to develop questions which accurately define the constructs, and enable the research respondents to provide proper response in a clear way, thereby reducing the measurement error.

Improving statistical estimation

The SEM provides robust methods to reduce the measurement errors associated with the observed or indicator variables. It helps to measure reliability i.e. the degree to which a set of indicators measure the same latent factor. High reliability indicates low or no measurement error.

Estimating direct and indirect effects

The SEM enables researchers to develop and test complex multivariable models of a specific social phenomenon. While modelling social phenomenon, SEM provides estimates of direct and indirect effects of variables on others. While direct effects are effects of a specific viable that go directly to another variable, the indirect effect indicates the effect of one variable transmitted to another, through an intermediary variable. For example, a farmers knowledge about an Integrated Pest Management package can directly improve the its adoption (direct effect), it can also develop a positive attitude towards IPM which enables its adoption. The combination of direct and indirect effects makes up the total effect of farmers knowledge of IPM on its adoption.

Types of Structural Equation Models

1. Path analysis models

The path analysis or structural models are pictorial representation of a theory of variable relations among observed variables. The path models in SEM are termed

as structural models, which test a series of hypotheses about the relationship among observed variables. Using path analysis, the researcher can employ many regression models simultaneously, and indirect and direct effects of the latent factors on others can be measured at the same time.

2. Confirmatory factor analysis models

The Structural Equation Model, two types of factor analyses are employed—exploratory and confirmatory. In exploratory factor analysis, the constructs or latent factors are extracted based on the relationships among the observed or indicator variables as explained by theory. In confirmatory factor analysis, the theoretically defined factor structure identified through exploratory factor analysis or through other means is confirmed. The confirmatory factor analysis models are also termed as measurement models, as indicate how well the latent factors are represented by observed variables or indicators.

3. Structural regression models

The structural regression models represent the combination of measurement and structural models, which are often, used for testing social science theories. By combining the measurement model and the structural models, this approach allows the inclusion of measurement errors in the analysis to produce accurate results.

4. Latent Change Models

The Latent change or latent growth curve models (Bollen and Curran, 2006), enable researchers to examine long-term changes of a specific phenomenon (both intra-individual temporal development and inter-individual similarities and differences) by examining the patterns of growth, decline, or both in longitudinal data.

APPLICATIONS OF STRUCTURAL EQUATION MODELS

(a) Construct validation

The SEM models are used for validating the constructs or latent factors of the phenomenon under study. They evaluate the extent to which a research instrument measures a latent variable which it is supposed to assess. The SEM is used for assessing the psychometric properties of a measurement device i.e. reliability and validity (Raykov and Marcoulides, 2006).

(b) Theory development

A theory is a simplified representation of a limited part of the reality (Pawar, 2009). It explains various aspects of the limited part of reality. A theory can be divided into two parts: one that specifies relationships between theoretical constructs; and another that describes relationships between constructs and measures (Bagozzi and Phillips, 1982).

SEM techniques are useful for identifying and establishing relationships between constructs (Bollen, 1989).

In SEM, the constructs or latent variables identified through Exploratory Factor Analysis are tested using Confirmatory Factor Analysis (CFA) to confirm their measurement properties like reliability and validity. Later the theoretical relationships are tested through multiple hypotheses specified in the Structural Model. Theory development through SEM is referred as exploratory SEM as it involves both development and confirmation of theories.

(c) Scale development

Scale development is an integral part of any empirical social research. Scales are basically a way of measuring socio-psychological phenomenon, by assigning a set of symbols or numerals to the individuals or their behaviours following specific rules, and the scores indicate the magnitude or degree of possession of the behaviour by individuals or groups under study (Kerlinger and Lee, 2000). When the scales are designed to measure latent socio-psychological variables, SEM provides a simple and accurate approach for their development. The SEM approach helps to identify and describe the latent constructs and test the relationships among them besides ensuring the good measurement quality.

STEPS IN STRUCTURAL EQUATION MODELING

Specifying a conceptual model

At the beginning of the research, the researcher specifies a conceptual model derived from theory. The researcher may choose an existing model for confirmation/ refinement or develop a new model depending on the research questions. The conceptual model specifies relationships among variable and test them through hypotheses. This conceptual model is essentially a path model which identifies the variables of a phenomenon and depict relationships between them. For example, a researcher wishes to know the utilisation of a Agricultural Information Kiosk by the farmers.

Two stage modelling in SEM

This conceptual model is testing using the SEM approach through two stages i.e. measurement model and structural model. In the first stage, the measurement model is tested while the second stage follows a structural model.

The measurement model specified the latent variables or constructs along with their indicator variables and measures how well these indicators together represent the latent variables. The measurement model involves confirmatory factor analysis (CFA) which estimates the construct validity of scales.

The second stage is the structural model showing how latent variables or constructs are related with each other by specifying multiple dependence relationships among

them. The structural model is a path diagram showing the relationship among latent variables.

Basic elements in SEM

A. Variables

Latent or unobserved constructs / variables or factors

The latent variables can't be measured directly, but indirectly measured through indicators. The latent variables are depicted as circle or ellipse in SEM.

e.g. Food security, empowerment.

Measured or observed or manifest variable

The measured variables are observed value of an item or question, which are indicators of latent variables. In SEM, they are depicted as square or rectangle.

e.g. An attitude statement in a scale

Exogenous Variables

The exogenous variables are independent variable that cause fluctuations in another latent variable. The examples include age, education, annual income etc.

Endogenous variables

The endogenous variables are multi-item equivalent of dependent variables, which are influenced by exogenous variables in the model.

B. Error terms in SEM

Measurement error

Measurement error is the degree to which the variables we can measure do not perfectly describe the respective latent construct. It is often associated with observed or indicator variables.

Residual error

It is the error in prediction of endogenous variables from exogenous variables. The residual error is associated with endogenous variables in the model.

Measurement model

The measurement model in SEM is developed through Confirmatory factor analysis (CFA). It is a statistical technique used to verify the factor structure of a set of observed variables. It helps to identify items suitable for each factor and also estimate reliability and construct validity of the scale.

The purpose of CFA is twofold – (i) it confirms a hypothesized factor structure, and (ii) Used as a validity procedure in the measurement model

Difference between EFA and CFA

In Exploratory Factor Analysis, the data determines the factor structure with a statistical objective to extract variance. In CFA, a theoretical factor structure is specified and tested for its fit with the observed covariances among the items in the factors. The statistical objective of CFA is to reproduce covariance matrix.

Steps in CFA

Step 1: Testing Assumptions of CFA

- Detection and management of Univariate outliers using Box plot, skewness and kurtosis
- Checking univariate normality – Normal probability plot
- Checking multivariate normality – Mardia multivariate kurtosis (Mardia, 1970)
- Sample size

The CFA is sensitive to sample size. The minimum sample sizes estimated from the number of constructs and commonalities are given in Table 1.

Table 1: Minimum sample size requirements

Number of constructs	No. of items / construct	Minimum sample size
Five or less	> 3	100
Seven or less	> 3	150
Large number of constructs (>7)	< 3	500

Source: Hair Jr *et al.* (2006)

Step 2: Model specification - Creating a visual diagram of the measurement model

This model is a hypothesized structure of latent (factor) and indicator (items) variables derived from EFA. The CFA helps us to test and validate this hypothesized structure. The SEM-CFA model can specify latent variables, indicators, directional (factor loadings of indicators on latent variables) and non-directional (correlation among latent factors). Various elements and types of relationships among factors and variables are explained in Table 2. An example of CFA is given in Fig 1.

Directionality

The model specification requires the researcher to describe the pattern of directional and non-directional relationships among the variables. The classical test theory assumes that all models are reflective - All the directions of arrows should originate from latent variable and move towards the items. Directional effects are represented

by regression coefficients. All directional effects (between latent variables, latent to indicator variables, residual in indicator variable) are considered as parameters. Non-directional arrows are specified between latent factors and are essentially covariances (correlation between latent factors) (Hair Jr *et al.*, 2006). In a reflective model, minimum of three items/ indicators are required for each factor or latent variable. It indicates that the researcher needs at least three statements for each dimension of the construct or variable.

Model identification

Model identification is an important aspect of CFA Modelling (Hair Jr *et al.*, 2006). For each free parameter, it is necessary that at least one algebraic solution is possible expressing that parameter as a function of the observed variances and covariances. If at least one solution is available, the parameter is identified. If not, the parameter is unidentified.

To correct under-identified model, the model must be changed or the parameter changed to a fixed value. To do this, the researcher adds a regression weight of one for every directional path of the first indicator variable of the latent factor. However, IBM SPSS AMOS creates this fixed value by default. To obtain accurate estimates, an over-identified model is required.

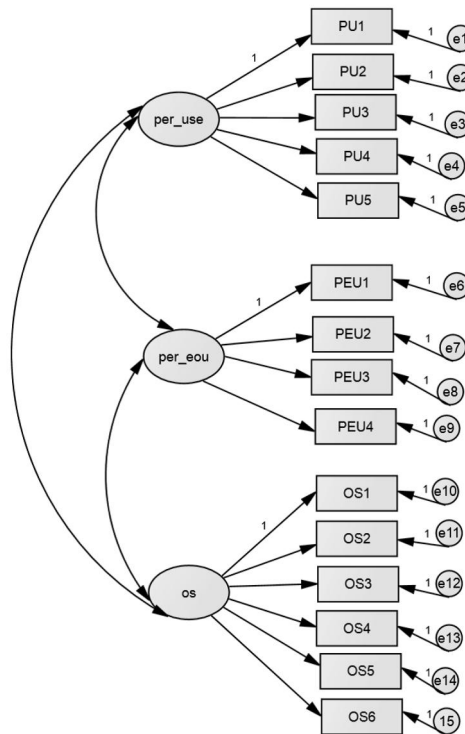




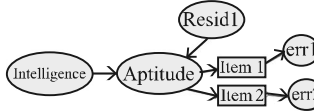
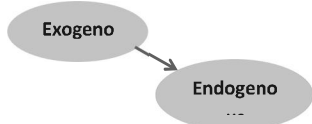
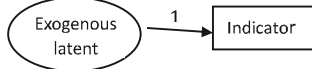


Fig 1: Visual diagram of the computer use behaviour constructs or latent factors and their items

Table 2: Elements and relationships of factors and variables in CFA

Terminology	Alternate name	Meaning	Symbol
Latent variable	Factor, construct	Unobserved hypothetical variable	
Indicator	Measured or manifest variable	Observed variable	
Factor loading	Path loading	Correlation between latent variable and indicator	
Non-directional association	Covariance, correlation	Correlation between two latent variables	
Indicator error	Predictor error, measurement error	Error in indicator that is not accounted for by latent variable. Indicator error is also considered a latent variable (err1, err2)	
Latent variable error	Residual error	Error in prediction of endogenous variables from exogenous variable (Resid1)	
Explained variance		Percentage of variance in dependent latent variable accounted for by predictor(s)	
Independent variable	Exogenous variable, predictor	Cause fluctuations in another latent variable	
Dependent variable	Endogenous variable criterion	Variable that is predicted by other latent variables or indicators	
Set parameter	Constrained parameter; Fixed path	Parameter that is set at a constant and not estimated. Parameters fixed at 1.0 reflect an expected 1:1 association between variables. Parameters set at 0 reflect the assumption that no relationship exists	

Source: Hair Jr *et al.* (2006)

Step 3: Working with software

Software available for analyzing SEM/CFA

- IBM SPSS AMOS (Analysis of Moment Structures)
- LISREL (Linear Structural Relationships)
- SAS – PROC CALIS Procedure (Covariance Analysis and Linear Structural Equations)
- EQS (Equations)

Among them, IBM SPSS AMOs is widely used as it is easy to use and helps to quickly specify, view, and modify the model graphically using simple drawing tools. It prints a high quality image of final model. All visual diagrams are created using AMOS GRAPHICS.

Step 4. Choosing the estimation procedure and outputs

Estimation is the mathematical algorithm that will be used to identify estimates for each free parameter (Hair Jr *et al.*, 2006). Various estimation methods are as follows.

Maximum likelihood method- Provides “most likely” parameter values to achieve best fit of the model, but sensitive to multivariate normality. However, MLM is the widely used method in CFA

Other estimation methods include weighted Least squares, generalised least squares, and Asymptotically distribution free (ADF)

Step 5. Goodness-of-fit and parameter estimates

AMOS output provided normality testing coefficients and two sets of estimates for assessing the model.

1. Goodness-of-fit indices – The test if the proposed model is same as the theoretical model
2. Reliability and validity of the scale.
 - (a) Tests of Multivariate normality

Mardia’s multivariate Kurtosis

 - The Mardia’s (1970) coefficient measures the multivariate kurtosis which indicate the presence or absence of multivariate normality.
 - Interpretation - Very small multivariate kurtosis values (e.g., less than 1.00) are considered negligible while values ranging from one to ten often indicate moderate non-normality. Values that exceed ten indicate severe non-normality (Mardia, 1970)

(b) Goodness-of-Fit indices

(i) Absolute fit indices

These indices determine how well a *a priori* model fits the sample data (McDonald and Ho, 2002). They indicate how well the proposed theory fits the data. Commonly used absolute fit indices include Chi-Squared test, RMSEA, GFI, AGFI, the RMR and the SRMR. (Table 3)

(ii) Incremental fit indices

Incremental fit indices, also known as comparative (Miles and Shevlin, 2007) or relative fit indices (McDonald and Ho, 2002), are a group of indices that do not use the chi-square in its raw form but compare the chi-square value to a baseline model. For these models the null hypothesis is that all variables are uncorrelated (McDonald and Ho, 2002). The incremental indices include TLI, NFI, IFI, CFI, RFI and AGFI.

(iii) Parsimonious fit measures

They diagnose whether model fit is due to over fitting the data with too many coefficients. They include Normed Chi-square, PGFI, PNFI

The cut-off values for the Goodness-of-fit coefficients are displayed in the Table 3. As a thumb rule, if three or more goodness-of-fit measures are in the acceptable range, the model is accepted.

Table 3: Criteria for model fit assessment, item validity and reliability

Test		Guideline	Reference
Absolute fit measures:			
Likelihood ratio Chi-square (χ^2)	H0: $\Sigma = \Sigma(\theta)$ HA: $\Sigma = \Sigma\alpha$	Small, Insignificant χ^2 ($p > 0.05$)	Joreskog and Sorbum (1996)
Root Mean Square Residual (RMR)	Average residuals between observed and estimated input matrices.	RMR < 0.07; <0.03 – excellent fit	Steiger (2007)
Goodness-of-Fit Index (GFI)	Overall degree of fit of the squared residuals from prediction compared with the actual data). Less influenced by sample size and normality	GFI > 0.95	Tabachnick and Fidell (2007)
Root mean square residual (RMSEA) and standardised root mean square residual (SRMR)	The RMR and the SRMR are the square root of the difference between the residuals of the sample covariance matrix and the hypothesised covariance model.	RMSEA \leq 0.06 or SRMR \leq 0.09	Tabachnick and Fidell (2007), Hu and Bentler (1999)

Structural Equation Modeling

Incremental or Comparative fit measures			
Tucker-Lewis Index (TLI) or NNFI	A comparative index between the proposed and the null model	TLI > 0.90	Hair Jr <i>et al.</i> (2006)
Normed Fit Index (NFI)	A relative comparison of the proposed model to the null model. $[\chi^2_{\text{null}} - \chi^2_{\text{proposed}}] / \chi^2_{\text{null}}$	NFI > 0.90	Bentler and Bonnet (1980)
Incremental Fit Index (IFI)	$(\chi^2_{\text{indep}} - \chi^2_{\text{model}}) / (\chi^2_{\text{indep}} - df_{\text{model}})$	IFI = Higher values close to 1	Kelloway (1998)
Comparative Fit Index (CFI)	Estimated based on non-central χ^2 distribution. Calculated by $1 - [(\chi^2_{\text{model}} - df_{\text{model}}) / (\chi^2_{\text{indep}} - df_{\text{indep}})]$	CFI > 0.90	Kelloway (1998)
Relative Fit Index (RFI)	$(\chi^2_{\text{indep}} - \chi^2_{\text{model}}) - [df_{\text{indep}} - (df_{\text{model}}/n)] / [\chi^2_{\text{indep}} - (df_{\text{indep}}/n)]$	RFI > 0.90	Kelloway (1998)
Adjusted Goodness-of-Fit Index (AGFI)	Goodness-of-fit adjusted by degrees of freedom (<i>df</i>).	AGFI > 0.90	Hair Jr <i>et al.</i> (2006)
Parsimonious fit measures			
Normed chi-square	χ^2 / df Alternate to Chi-square Used when Chi-square of the model is significant	2.0 to 5.0	Wheaton <i>et al.</i> (1977), Tabachnick and Fidell (2007)
Parsimonious Goodness-of-Fit Index (PGFI)	$1 - (P/N) \times GFI$. P = No of estimated parameters in the model, N = Number of data points. Adjusts GFI for the degrees of freedom in the model.	PGFI = Higher values close to 1	Kelloway (1998)
Parsimonious Normed Fit Index (PNFI)	$(df_{\text{model}}/df_{\text{indep}}) \times NFI$. Adjusts NFI for model parsimony	PNFI = Higher values close to 1	Kelloway (1998)

(c) Reliability and validity

Convergent validity

- Assessed through standardized loadings (λ), composite reliabilities (squared multiple correlations) and variance extracted by each latent factor.

Cutoff values

- Standardised regression coefficient for each indicator variable : $\lambda > 0.6$ (Bollen, 1989)
- Composite reliability – squared multiple correlation for each indicator variable : $R^2 \geq 0.50$ (Bagozzi and Yi, 1988)

- Variance extracted by each latent factor – ≥ 0.50 (Fornell and Larcker, 1981)

Composite Reliability

- Squared multiple correlations (R^2) are used as a measure of reliability of each Indicator variable
- Used to assess the amount of variation in latent variables explained by predictors.
- $R^2 > 0.5$ (Bagozzi and Yi, 1988)

REFERENCES

- Bagozzi, R. P. and L. W. Phillips (1982), Representing and testing organizational theories: A holistic construal. *Administrative Science Quarterly*, 27:459–489.
- Bagozzi, R. P. and Y. Yi (1988), On the Evaluation of Structural Equation Models. *Journal of the Academy of Marketing Science*, 16 (1) :74-94.
- Bentler, P. M. and D. C. Bonnet (1980), Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88 (3), 588-606.
- Bollen, K. A. and P. J. Curran (2006), Latent Curve Models: A structural equation approach. New York: Wiley.
- Bollen, K. A. (1989), Structural equations with latent variables. John Wiley and Sons, New York, New York, USA
- Fornell C. R. and D. F. Larcker (1981), Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research* 18 (1): 39-50.
- Hair, Jr. J. F., W. C. Black, B. J. Babin, R. E. Anderson and R. L. Tatham (2006), Multivariate Data Analysis (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
- Hair, J. F., W. C. Black, B. J. Babin and R. E. Anderson (2010), Multivariate Data Analysis. 7th Edition, New Delhi Pearson
- Hu, L. T. and P. M. Bentler (1999), Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives, *Structural Equation Modeling*, 6 (1): 1-55.
- Joreskog K. and D. Sorbom (1996), LISREL 8: User's Reference Guide. Chicago, IL: Scientific Software International, Inc. pp.271–274.
- Kelloway, K. E. (1998), Using LISREL for structural equation modeling: A researchers guide. Thousand Oaks. Sage.
- Kerlinger and Lee (2000), Foundations of Behavioral Research. Orlando, FL: Harcourt College Publishers.
- Kline, R. (1999), Principles and practice of structural equation modeling (3rd Ed. b.). New York: Guilford.

- Mardia, K. V. (1970), Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57: 519–530.
- McDonald, R. P. and Ho, M.-H.R. (2002), Principles and Practice in Reporting Statistical Equation Analyses. *Psychological Methods*, 7 (1): 64-82.
- Miles, J. and M. Shevlin (2007), A time and a place for incremental fit indices, *Personality and Individual Differences*, 42 (5): 869-74.
- Pawar, B. (2009), Theory building for hypothesis specification in organizational studies. Thousand Oaks, CA: Sage.
- Raykov, T. and G. A. Marcoulides (2006), A first course in structural equation modeling (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Steiger, J. H. (2007), Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42 (5): 893-98.
- Tabachnick, B. G. and L. S. Fidell (2007), Using Multivariate Statistics (5th ed.). New York: Allyn and Bacon.
- Wheaton, B., B. Muthen D. F. Alwin and G. Summers (1977), Assessing reliability and stability in panel models. *Sociological Methodology*, 8 (1): 84-136.

Chapter 30

PARTIAL EQUILIBRIUM MODEL

Shinoj Parappurathu

INTRODUCTION

A partial equilibrium is a type of economic equilibrium, wherein the clearance on the market of some specific goods is obtained independently from prices and quantities demanded and supplied in other markets. In other words, in such equilibrium, the prices of all substitutes and complements, as well as income levels of consumers are constant. Here, the dynamic process is such that prices adjust until supply equals demand. It is a powerfully simple technique that allows one to study equilibrium, efficiency and comparative statics. The stringency of the simplifying assumptions inherent in this approach make the model considerably more tractable, but may produce results which, while seemingly precise, do not effectively model real world economic phenomena. In partial equilibrium analysis, the effects of policy actions are examined only in the markets that are directly affected. Supply and demand curves are used to depict the price effects of policies. Producer and consumer surplus is used to measure the welfare effects on participants in the market. A partial equilibrium analysis ignores effects on other industries in the economy or assumes that the sector in question is very, very small and therefore has little if any, impact on other sectors of the economy.

Partial equilibrium *versus* general equilibrium

While partial equilibrium analysis considers only a particular market or sector and the underlying demand and supply dynamics, general equilibrium seeks to explain the behaviour of supply, demand and prices in a whole economy with several or many interacting markets, by seeking to prove that a set of prices exists that will result in an overall equilibrium. As with all models, this is an abstraction from a real economy; it is proposed as being a useful model, both by considering equilibrium prices as long-term prices and by considering actual prices as deviations from equilibrium. General equilibrium theory enables one to study economies using the model of equilibrium pricing and seeks to determine in which circumstances the assumptions of general equilibrium will hold. The theory dates to the 1870s, particularly the work of French economist Léon Walras. The distinction between partial and general equilibrium models can be made in terms of (a) *ceteris paribus* assumptions and (b) the variables of interest that are endogenous. At one extreme is the typical model of a commodity market that takes the price and quantity of that commodity as endogenous treating all of other goods as constant and exogenous to the analysis. On the other extreme, are the detailed economy-wide models in which all prices and quantities are endogenous to, and measured in the analysis, so that the extreme *mutatis mutandis* (everything

allowed to change) replaces extreme *ceteris paribus*. Most economic analyses fall somewhere in between these two extremes. Another way of distinguishing is in terms of techniques of analysis. For instance, when Marshallian supply-and-demand models are used, the analysis is typically regarded as being a partial equilibrium analysis, whereas when a social accounting matrix (SAM) is involved, it is regarded as general equilibrium analysis (Alston *et al.*, 1998).

Modeling approaches to partial equilibrium analysis

A partial equilibrium modeling problem can be approached from different angles based on the kind of modeling framework used for solving it. A simple partial equilibrium model could consist of linear equations of demand and supply specified by an equilibrium condition as illustrated below;

Building the model

- Demand equation (behavioural equation): $QD = a - bP$ ($a, b > 0$)
 - Supply equation (behavioural equation): $QS = -c + dP$ ($c, d > 0$)
 - Equilibrium condition: $QD = QS$
- Method of finding equilibrium
- Solution by eliminating variables

Non-linear models with quadratic terms or higher degree polynomial terms can be used to replace simple linear models when the situation demands detailed examination of the underlying market situation. Still, we have the specific condition for determining the existence of equilibrium with economic implications. In addition to algebraic approach, linear models can be solved using graphical approach also.

Partial equilibrium models and agricultural policy analysis

Partial equilibrium models are widely used in sector specific policy analyses and have found innumerable applications in the context of economic policy analysis in agriculture (Goletti and Rich, 1998; Minot, 2009; Roningen, 1997; Sadoulet and de Janvry, 1995). Such models are used under various contexts; agricultural commodity market outlook models like IMPACT model of International Food Policy Research Institute (IFPRI), Rice outlook model of International Rice Research Institute (IRRI), World food model of Food and Agricultural Organization (FAO) are typical partial equilibrium models with the primary objective of generating short and medium term outlook of major food commodities. IMPACT model is a multipurpose model with the capability of simulating important policy variables under various alternative scenarios. Some other models like the World Agricultural Trade Simulation model (WATSim) of Food and Agricultural Policy Research Institute (FAPRI) are specifically designed to model international agricultural trade and related policy simulations. These models could be of different dimensions; some are multi-commodity multi-region spatial models while some others are single commodity national models.

Structure of a typical agricultural policy model

A typical agricultural outlook model under partial equilibrium framework consists of the following sub-components.

1. Producer core system
 - (i) Area equation
 - (ii) Yield equation
 - (iii) Production equation
 - (iv) Supply equation
2. Consumer core system
 - (i) Food demand equation
 - (ii) Feed demand equation
 - (iii) Other uses demand equation
 - (iv) Total demand equation
3. Trade core system
 - (i) Export equation
 - (ii) Import equation
4. Price linkage equation
5. Model closure

The producer core system depicts the supply side of the commodity under question whereas the consumer core system depicts the demand side. Various demographic and conditional variables like research investment, irrigation, weather parameters like rainfall, temperature, other qualitative variables that determine the choice of consumers etc. can also be incorporated into both producer and consumer core systems as exogenous variables. The trade core system is inserted in the case of an open economy where the goods are traded outside the economy. The price linkage equations link the demand and supply sides with equilibrium conditions. In addition to this, a number of policy variables like tariffs, subsidies, and support prices can also be incorporated into these models exogenously to capture the effects of policy changes. The technical parameters of the various equations have to be estimated based on real data either time series, cross section or pooled. The accuracy of the model output would depend a great deal on how realistic these estimates are. With this basic structure, the model could have various sub-sectoral dimensions which can include crop sector, livestock sector, dairy sector, input sector and spatial dimensions that may vary from regional dimension, national dimension and global dimension depending upon the spatial coverage the modeler intends to incorporate. Both linear and non-linear programming approaches can be applied to derive optimum feasible solutions and various algorithms are available for solving such models. Software packages like SAS, GAMS, Microsoft spread sheet, etc. are enabled with features to construct commodity outlook models and solve them using alternative iterative procedures. A simple partial equilibrium model designed for agricultural policy analysis under GAMS is presented in Annexure I.

ILLUSTRATION

Commodity outlook for Indian agriculture: The case of cereal outlook model

The cereal outlook model was developed by the authors for generating commodity outlooks for rice, wheat and maize in India. A detailed account of this model is available in Parappurathu, *et al.* (2014a, 2014b). The model was constructed under a dynamic as well as spatial partial equilibrium modeling framework that incorporate a system of simultaneous equations for effectively depicting the linkages between various economic variables in the balance sheet of major cereals in India. The model takes cognizance of the key economic variables such as production, demand, stocks, trade, prices and policy variables related to the primary commodities. It has sought to generate medium-and long-term projections, given the past trends in behaviour of the variables in question as well as magnitude of technical coefficients which govern their behavior. Technically, the model derives equilibrium values of the variables based on the econometric linkages established through a set of equations that cuts across commodity as well as spatial dimensions. It is an open model as it takes into account the trade flows of the commodities with respect to the rest of the world and the endogenous prices are attached to the world market prices. The Model is dynamic in the sense that the current prices and quantities are related to the past prices and quantities and the equilibrium is attained through a dynamic recursive iterative process that continuously adjusts the quantities and prices across time periods till the overall model converges to an equilibrium state. Spatial dimensions have been incorporated by specifying supply side equations separately for six regions in the country.

Model structure

The Cereal Outlook Model is a typical agricultural-related model that incorporates the major demand and supply side variables, output and input prices, as well as other exogenous variables like income and population; and policy variables like support prices, tariffs, etc. A schematic representation of the linkages in the model is shown in Fig 1.

Broadly, the Model comprises of the following integral components: (i) a producer core system that integrates the linkages between area, yield, production, stock changes and supply of the individual grains; (ii) a consumer core system that includes the equations for food and other uses demand, feed demand and total demand; (iii) a trade core system that incorporates the export and import equations; (iv) a set of price linkages equations for depicting the relationships between producer prices with consumer prices and national and regional prices; and (v) a model closure equation that links the various cores of the model with certain closure conditions.

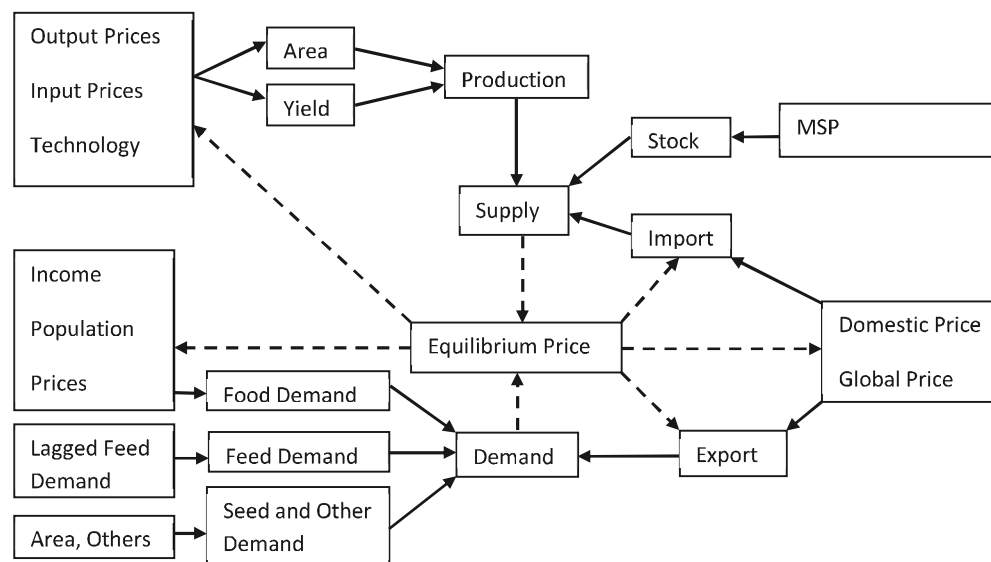


Fig 1: Modeling framework of cereal outlook model: an illustration

Source: Parappurathu *et al.* (2014a)

However, all such partial equilibrium models would be essentially bound by some basic economic assumptions like perfect competition, constant returns to scale etc. which can be relaxed to some extent depending upon circumstances. The utility of such models varies depending upon the way each of the sub-systems are modeled based on the discretion of the modeler and the purpose for which it is built; whether for forecasting, policy simulations or baseline situation assessment.

REFERENCES

- Alston, J. M., G. W. Norton and P. G. Pardey (1998), *Science under scarcity*. CAB International.
- Goletti, F. and K. Rich (1998), *Policy Simulation for Agricultural Diversification*. Report prepared for the UNDP project on Strengthening Capacity Building for Rural Development in Viet Nam, Washington, D.C.: IFPRI.
- Minot, N. (2009), *Using GAMS for Agricultural Policy Analysis*, Technical Guide, International Food Policy Research Institute, Washington, D.C.
- Parappurathu, S., A. Kumar, S. Kumar, and R. Jain (2014a), A partial equilibrium model for future outlooks on major cereals in India, *Margin-The Journal of Applied Economic Research*, 8 (2): 155-192.
- Parappurathu, S., A. Kumar, S. Kumar, and R. Jain (2014b). *Commodity outlook on major cereals in India*, Policy Paper No. 28, National Centre for Agricultural Economics and Policy Research, New Delhi.
- Roningen, V. O. (1997), *Multi-Market, Multi-Region Partial Equilibrium Modeling*. In *Applied Methods for Trade Policy Analysis: A Handbook*, J.F. Francois and K.A. Reinert, eds. (Cambridge: Cambridge University Press), 231-257.
- Sadoulet, E. and A. deJanvry (1995), *Quantitative Development Policy Analysis* (Baltimore: Johns Hopkins University Press), chapter 11.

ANNEXURE I

Example of an Agricultural Sector Partial Equilibrium Model in GAMS

Source: Minot (2009)

* FILE: SDP4

* GAMS PROGRAM TO SIMULATE SUPPLY AND DEMAND FOR FOUR GOODS

* IN SIX REGIONS WITH INTERNAL TRADE, IMPORTS AND EXPORTS

* WITH TRADE TAXES AND QUOTAS AND REGIONAL TRADE RESTRICTIONS

* Note: LC refers to local currency units. OPTION LIMCOL = 0;

OPTION LIMROW = 0;

\$OFFSYMLIST;

\$OFFSYMREF;

SET

C Crops /Rice Maize Mustard Citrus /

RW Region including world

/WEST CENTRAL EAST S_WEST S_CENT S_EAST WORLD / R(RW) Region

/WEST CENTRAL EAST S_WEST S_CENT S_EAST /;

ALIAS (R,RR), (RW, RRW) ;

TABLE P0(C,R) Original price (LC per kg)

	WESTCENTRAL		EAST	S_WEST	S_CENT	S_EAST	RICE
	2987	2982	2756	2636	2354	2368	
MAIZE	2112	1988	1882	1439	1321	1694	
MUSTARD	1553	1372	1003	1245	731	731	
Citrus	535	473	499	584	486	486	;

TABLE WP(C,*) World price (US\$ per ton)

	X	M
Rice	270	320
Maize	126	141
Mustard	0	150
Citrus	32	100 ;

TABLE TAX(C,*) Trade tax (fraction)

	X	M
Rice	.01	.00
Maize	.00	.00
Mustard	.00	.00
Citrus	.00	.00 ;

TABLE QUOTA(C,*) Trade quota (1000 tons)

	X	M
Rice	3000	9999
Maize	9999	9999
Mustard	9999	9999
Citrus	9999	9999;

TABLE DPE(C,R) Price elasticity of demand

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	Rice
	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Maize	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Mustard	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
Citrus	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0;

TABLE DYE(C,R) Income elasticity of demand

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	Rice
	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Maize	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Mustard	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Citrus	1.0	1.0	1.0	1.0	1.0	1.0	1.0 ;

TABLE SPE (C,R) Price elasticity of supply

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	Rice
	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Maize	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Mustard	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Citrus	1.0	1.0	1.0	1.0	1.0	1.0	1.0 ;

TABLE D0(C,R) Original demand (1000 tons)

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	RICE
	2445	1522	1121	497	1189	2593	
MAIZE	187	130	100	54	51	128	
MUSTARD	315	367	132	56	75	225	
CITRUS	319	248	188	77	82	217 ;	

TABLE S0(C,R) Original supply (1000 tons)

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	RICE
	2571	1190	973	238	520	7129	
MAIZE	183	110	50	102	198	75	
MUSTARD	398	450	144	71	10	115	
CITRUS	45	210	331	258	501	70 ;	

TABLE TCOST(RW,RRW) Cost of transportation (LC per kg)

Partial Equilibrium Model

	WEST	CENTRAL	EAST	S_WEST	S_CENT	S_EAST	WORLD	WEST
	0	149	9999	9999	230	233	67	
CENTRAL	149	0	325	9999	9999	9999	189	
EAST	9999	325	0	173	294	364	294	
S_WEST	9999	9999	173	0	164	9999	164	
S_CENT	230	9999	294	164	0	47	35	
S_EAST	233	9999	364	9999	4	70	35	
WORLD	67	189	294	164	35	35	0 ;	

TABLE ITX (C,R,RR) Implicit tax on internal trade (LC per kg)

RICE	WEST	WEST	EAST	S_WEST	S_CENT	S_EAST
		400	400			

RICE EAST 200

RICE .S_WEST 100 100

RICE .S_CENT 400 200 100

RICE .S_EAST 400 100

MAIZE .WEST

MAIZE .S_CENT 300 300 300

MAIZE .S_EAST 300

MUSTARD .WEST 300 300

MUSTARD .S_CENT 300

MUSTARD .S_EAST 300

CITRUS.WEST 300 300

CITRUS.S_CENT 300

CITRUS.S_EAST 300 PARAMETERS

DA Intercept of demand equation

DB Price coefficient of demand equation

DC Income coefficient of demand equation SA Intercept of supply equation

SB Price coefficient of supply equation NER Nominal exchange rate (LC per US\$)

PX(C) Export price (LC per kg)

PM(C) Import price (LC per kg)

Y0(R) Expenditure in 1995 (m LC per capita) Y92(R) Expenditure in 1992-93 (m LC per capita)

/ WEST 1.102

CENTRAL 0.871

EAST 1.267

S_WEST 1.481

S_CENT 1.840

S_EAST 1.469 /;

Y0(R) = 1.6*Y92(R) ;

$DB(C,R) = DPE(C,R)*D0(C,R)/P0(C,R) ;$
 $DC(C,R) = DYE(C,R)*Y0(R)/P0(C,R) ;$
 $DA(C,R) = D0(C,R) - DB(C,R)*P0(C,R) - DC(C,R)*Y0(R) ; SB(C,R) =$
 $SPE(C,R)*S0(C,R)/P0(C,R) ;$
 $SA(C,R) = S0(C,R) - SB(C,R)*P0(C,R) ; NER = 10;$
 $PX(C) = NER*WP(C,'X')*(1-TAX(C,'X')) ;$
 $PM(C) = NER*WP(C,'M')*(1+TAX(C,'M')) ;$ VARIABLES
P(C,R) Equilibrium price (LC per kg) D(C,R) Quantity demanded (thousand
tons) S(C,R) Quantity supplied (thousand tons) ; POSITIVE VARIABLES
TQ(C,R,RR) Transported quantity (thousand tons) IXT(C) Implicit export tax
(LC per kg)
IMT(C) Implicit import tax (LC per kg) X(C,R) Exports (thousand tons)
M(C,R) Imports (thousand tons);
EQUATIONS
DEMAND Demand equation SUPPLY Supply equation
IN_OUT Shipments into and out of region DOM_TRADE Domestic trade price
relationships EXPORTS Export price relationships
IMPORTS Import price relations
XQUOTA Export quota
MQUOTA Import quota;
DEMAND(C,R)..
 $D(C,R) = E = DA(C,R) + DB(C,R)*P(C,R) + DC(C,R)*Y0(R) ;$
SUPPLY(C,R)..
 $S(C,R) = E = SA(C,R) + SB(C,R)*P(C,R) ; IN_OUT(C,R)..$
 $S(C,R) + SUM(RR,TQ(C,RR,R)) - SUM(RR,TQ(C,R,RR)) - X(C,R) + M(C,R)$
 $= E = D(C,R) ; DOM_TRADE(C,R,RR)..$
 $P(C,R) + TCOST(R,RR) + ITX(C,R,RR) = G = P(C,RR) ; EXPORTS(C,R)..$
 $P(C,R) + IXT(C) + TCOST(R,'WORLD') = G = PX(C) ; IMPORTS(C,R)..$
 $PM(C) + IMT(C) + TCOST('WORLD',R) = G = P(C,R) ; XQUOTA(C)..$
 $QUOTA(C,'X') = G = SUM(R,X(C,R)) ; MQUOTA(C)..$
 $QUOTA(C,'M') = G = SUM(R,M(C,R)) ; TQ.FX(C,R,R) = 0 ;$
MODEL MARKET / DEMAND SUPPLY
IN_OUT DOM_TRADE.TQ EXPORTS.X IMPORTS.M XQUOTA.IXT MQUOTA.
IMT /;
SOLVE MARKET USING MCP;

Chapter 31

PRODUCTION FUNCTION ANALYSIS

Suresh Kumar, Dharam Raj Singh and Girish Kumar Jha

INTRODUCTION

Production function defines the technical relationship that transforms inputs (resources) into outputs (commodities). Mathematically, a production function relates the maximum amount of output that can be obtained from a given levels of inputs. More precisely, it describes a boundary or frontier representing the limit of output that can be obtained from given level and their combination of inputs. Alternatively, a production function $y = f(x_1, x_2, \dots, x_n)$ represents the maximum output y^* that can be achieved using input vector $x = (x_1, x_2, \dots, x_n)$. All values of x greater than or equal to zero constitute the domain of this function. The range of the function consists of each output level (y) that results from each level of input (x) being used.

A general form of agricultural production function can be written as

$$y = f(x_1, x_2, \dots, x_n; x_{n+1}, x_{n+2}, \dots, x_m; x_{m+1}, x_{m+2}, \dots, x_l)$$

where,

y = output, x_1, x_2, \dots, x_n are of decision variables, and the quantities/ levels of these variable is under the control of decision makers, for example seed rate, application of fertilizer; $x_{n+1}, x_{n+2}, \dots, x_m$ is a subset of predetermined variables whose values are known to decision maker in advance at the time of decision-making process, for instance, fertility level of field; and $x_{m+1}, x_{m+2}, \dots, x_l$ indicate the subset of uncertain variables, the level of these variables neither known to decision makers nor can be controlled during the production process, for example rainfall during the production. Fundamentally, production function shows a physical relationship between output and inputs, however, sometime, output and input are expressed in monetary terms if the either data on physical terms are not available or some time inputs cannot be aggregated in physical units.

Assumptions of production function analysis

1. The production function is defined only for the non-negative values of inputs and outputs ($y \geq 0$ and $x_i \geq 0; i = 1, 2, \dots, l$).
2. The production function presupposes technical efficiency. This means that every possible combination of inputs is assumed to result in maximum level of output. In other words, there is no technical inefficiency in production process.
3. The input- output relationship or the production function is single valued and continuous. It means the first ($\frac{\partial y}{\partial x_i}$) and second ($\frac{\partial^2 y}{\partial x_i^2}$) order partial derivatives of the y w.r.t. each input variables are non-vanishing.

4. The production function in input-output, input-input and output-input plane is characterized by

(a) decreasing marginal product for all factor product combinations i.e.

$$\frac{\partial^2 y}{\partial x_i^2} < 0, \quad i = 1, 2, \dots, l.$$

(b) decreasing rate of technical substitution between any two factors, it means

$$\frac{\partial^2 x_i}{\partial x_j^2} < 0, \quad i, j = 1, 2, \dots, l. (i \neq j)$$

(c) an increasing rate of product transformation between any two products. It can be expressed as:

$$\frac{\partial^2 y_i}{\partial y_j^2} > 0, i, j = 1, 2, \dots, m (i \neq j)$$

5. The returns to scale are assumed to be decreasing. It can be stated as

$$\sum \left(\frac{x_i}{y} \right) \left(\frac{\partial y}{\partial x_i} \right) < 1, \quad i = 1, 2, \dots, l$$

6. All the factors of production and products are perfectly divisible.
 7. The parameters determining the firm's production function do not change over the time period considered. Also, these parameters are not allowed to be random variables.
 8. The exact nature of any production function is assumed to be determined by a set of technical decisions taken by the producer.

Homothetic production function

Most of the production functions used in empirical analyses are homothetic functions. It can be defined in terms of its isoquants; the isoquants are radial projections of each other or are “radial extensions” of the unit isoquant. The slope of the isoquants is constant along every ray through the origin, that is, the isoquant slope at A equals the slope at B. And that at C equals that at D (Fig 1).

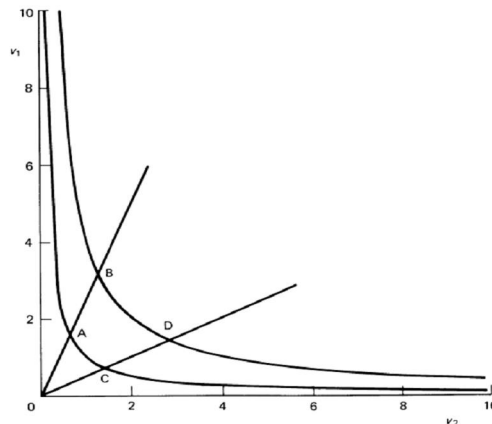


Fig 1: Homothetic production function

Homogeneous production functions

$$f(kx_1, kx_2) = k^\lambda f(x_1, x_2)$$

if each input is increased by a factor k , then the total output increased by k^λ times, where λ is the degree of homogeneity. If λ takes value of unity, then function is homogenous function of degree one. Homogeneous production functions are homothetic, exhibit constant elasticity of substitution and fixed returns to scale. Cobb-Douglas, Leontief and CES production function are homogenous production function. On the other hand, non-homogeneous function is characterized by varying elasticity of substitutions, and varying economies of scale, and its example is translog production function.

Most Commonly used Production Functions in Agriculture

Key characteristics of some important production functions used in agricultural production and resource economics are given Table 1. Important features, advantages and disadvantages of these functions have been discussed here.

Linear Production Function

It is very simple but not commonly used in production economics as it violates the fundamental assumption of production function analysis. For instance, its second order derivative is not less than zero, which means marginal product is not decreasing, and it exhibits constant return to scale. Isoquants are straight lines showing that there is a perfect substitution, which is rare.

Quadratic production function

Quadratic function describes a parabolic function commonly used by the biologists particularly for studying the effect of fertilizer response on crop yields. A well-behaved quadratic function ($b_{ii} < 0$) is concave to input axis. with isoquants and product transformation curve are convex and concave, respectively. Isoclines are linear and converge to a point on the highest isoquant on the isoquant map representing the maximum attainable physical output and this point is called “Von Leibig point”.

Cobb-Douglas production function

It was proposed by labor economist Paul H. Douglas and mathematician Charles W. Cobb in an effort to fit Douglas’s empirical results for production, employment, and capital stock in the U.S. manufacturing into a simple function (Cobb and Douglas, 1928). In the economic theories, economists often use Cobb-Douglas (CD) function in their studies since the parameters of this function can be easily identified, analyzed and interpreted, and its seemingly good empirical fit across many data sets (Miller, 2008). This also satisfies the neo-classical criteria i.e. marginal products decrease as inputs increase. Most importantly, the partial elasticities of production, which measures the responsiveness of output to unit increase of input, are identical to the

production coefficients (b_i 's). Further, the sum of partial elasticities of production ($\sum b_i$) can be interpreted as a measure of economies of scale, i.e. the percentage change of output relative to the percentage change in all inputs. It is a homogeneous function that provides a scale factor enabling one to measure the returns to scale and interprets the elasticity coefficient with relative ease. It is also characterized by continuous production iso-quants so that for realizing a given level of output there is a scope of varying input prices, and therefore, inferences can be drawn about the possibility and the rate of substitution among different inputs.

This function allows either constant, increasing or/decreasing marginal productivity, however, it does not allow an input-output curve embracing all three simultaneously. It can be easily demonstrated that CD function ($y = AK^\alpha L^{1-\alpha}$) exhibits constant returns to scale and elasticity of substitution equals to unity. Further, in a competitive market factors are paid as per their marginal products, then α and $1-\alpha$ are equal to capital and labour's share of output, respectively. This proves the validity of Euler's theorem. Elasticity of substitution equal to unity implies that factor shares will remain constant for any capital-labour ratio because any changes in factor proportions will be exactly offset by changes in the marginal productivities of the inputs. Thus, the observed income shares will be constant through time. For instance, any change in K/L (capital-labour ratio) will be matched by a proportional change in w/r (wage-rent ratio), through exactly matched by a percentage increase in the marginal rate of technical substitution (MRTS).

Therefore, the relative income shares of capital and labor will remain constant. Consequently, shares of output are allocated to capital and labour even though the capital-labour ratio may change over time yet will remain constant. However, this is a major limitation of the CD function, as empirical investigation shows that share of labour in national income was not fixed but instead varied as wage rates varied. Another limitation is that its isoclines are straight lines passing through the origin, and these isoclines are scale lines. This implies that at the different level of outputs (in the isoquant map) a fixed proportion or mix of inputs will be used. That is to say if the elasticity of scale was ' μ ' for one level of output and one factor combination then it will be ' μ ' for all levels of output and all factor combinations. This results in the long run average cost curve (LRAC) which is either continuously rising, a horizontal line or continuously falling. Thus, the LRAC curve cannot take the 'U'-shape so often assumed for it in the theory of the firm (Heathfield, 2016).

Square-root production function

It has the features of CD and quadratic production functions. Unlike the CD function it does not require the fixed mix of inputs for producing different level of output. Further, it overcomes the limitation of linear isoclines of quadratic production function. Further, the rate of decline in marginal product is slower than the quadratic function leading

to no mirror effect as depicted in the case of quadratic production function. Further, elasticity of substitution is not constant like CD function, and varies with the level of inputs.

Table 1: Key characteristics of some important production functions

Function	Functional Form	Marginal product*	Elasticity of production/ scale	Elasticity of substitution	Isoclines and Ridge line
Linear	$y = b_0 + \sum_{i=1}^n b_i x_i$	$MP_1 = b_1$	1	∞	undefined
Quadratic	$y = b_0 + \sum_{i=1}^n b_i x_i + \sum_{i=1}^n b_{ii} x_i^2 + \sum_{i=1}^n \sum_{j=1}^n b_{ij} x_i x_j$, $i < j$ or $y = a + b_1 x_1 + b_{11} x_1^2$	$MP_1 = b_1 + 2b_{11}x_1$	$E_{p1} = \frac{b_1 + 2b_{11}x_1}{b_1 + b_{11}x_1}$		Straight lines
Cobb-Douglas	$y = b_0 \prod x_i^{b_i}$	$MP_1 = \frac{y}{x_1} b_1$	$E_{p1} = b_1$	1	Straight lines and all are passing through origin
Transcendental	$y = b_0 \prod x_i^{b_i} e^{\beta_i x_i}$	$MP_1 = y \left(\frac{b_1}{x_1} + \beta_1 \right)$	$E_{p1} = b_1 + \beta_1 x_1$		Pass through origin and non-linear
Constant elasticity substitution (CES)	$y = A \left(\sum_{i=1}^n (\lambda_i x_i^{-\rho}) \right)^{-\frac{1}{\rho}}$, $A > 0, \rho > 0, \lambda_i \geq 0, \sum \lambda_i = 1, \rho > -1$ or $y = A [\lambda x_1^{-\rho} + (1-\lambda)x_2^{-\rho}]^{-\frac{1}{\rho}}$	$MP_1 = \frac{\lambda}{A^\rho} \left(\frac{y}{x_1} \right)^{1+\rho}$	$E_{p1} = \frac{\lambda}{A^\rho} \left(\frac{y}{x_1} \right)^\rho$	$\frac{1}{1+\rho}$	Pass through origin and are straight lines.
Tranlog	$\ln y = \ln b_0 + \sum_{i=1}^n b_i \ln x_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n b_{ij} \ln x_i \ln x_j$	$\left(b_1 + \frac{1}{2} \sum_{j=1}^n b_{1j} \ln x_j \right) \frac{y}{x_1}$	$\left(b_1 + \frac{1}{2} \sum_{j=1}^n b_{1j} \ln x_j \right)$	Depends on the level of output and on the level of inputs	-
Square-root	$y = b_0 + \sum_{i=1}^n b_i x_i^{0.5} + \sum_{i=1}^n b_{ii} x_i + \sum_{i=1}^n \sum_{j=1}^n b_{ij} x_i^{0.5} x_j^{0.5}$ $i, j = 1, 2, \dots, n (i < j)$	$b_{11} + 0.5b_1 x_1^{-0.5}$	$(b_{11} + 0.5b_1 x_1^{-0.5}) \frac{x_1}{y}$		curvilinear

*Marginal product (MP_i) has been given for x₁.

Constant elasticity of substitution production function

Arrow *et al.* (1961) introduced the constant elasticity of substitution (CES) production function which has the advantage to be a generalization of the three main functions *viz.*, linear function (for perfect substitutes), the Leontief function (for perfect complements) and the CD function, which assume respectively an infinite, a zero and a unit elasticity of substitution (ES) between production factors, respectively. As its name suggests, the CES production function exhibits constant elasticity of substitution, $\varepsilon = \frac{1}{1+\rho}$, between

capital and labour. Leontief, linear and Cobb–Douglas functions are special cases of the CES production function. When $\rho = \infty$, then it reduces to Leontief function or perfect complements production function, and if $\rho = 0$, it reduces to CD function and if the value of $\rho = -1$ it becomes a case of linear or perfect substituted function.

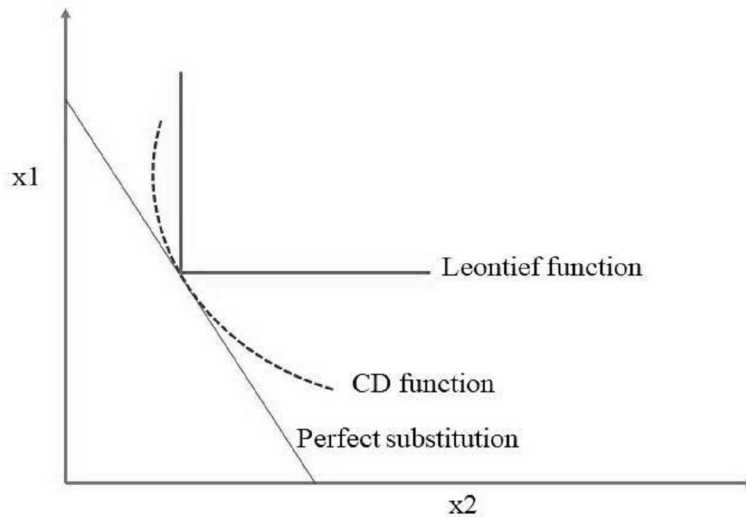


Fig 2: CES reduces to Leontief, CD and linear function

As far as limitation of this function is concerned, generalization of it for more than two factors (or n) is difficult (Diewert, 1971). It assumes that the elasticity of substitution between capital and labour is invariant with respect to relative factor inputs, it means that elasticity of substitution is different from zero and unit but remain constant at all the levels of inputs. Further, it is a non-linear function, therefore, estimation of parameters is cumbersome task.

Transcendental production function

It is a hybrid between the CD and exponential function. The main characteristics of this function are that it allows increasing, decreasing and negative marginal products separately or all three simultaneously. Therefore, it is useful in depicting all the three stages of classical (traditional) production function. Further, its elasticity of production and elasticity of substitution varies over the range of inputs, and its isoclines pass through the origin, are curved and converge at a single point indicating the maximum attainable output.

Translog production function

This is also known as the transcendental logarithmic production, and exhibits flexible elasticity of substitution between inputs and also with no restriction on return to scale. It allows the elasticity of scale to change with output and/or factor proportions. One of the main advantages of the production function is that, unlike in case of Cob Douglas production function, it does not assume rigid premises such as: perfect or smooth substitution between production factors or perfect competition in the product and

factors markets (Klacek, *et al.*, 2007). In a translog production function, the number of parameters practically explodes as the number of production factors increases, and consequently, multicollinearity is a major shortcoming in the estimation of this production function (Pavelescu, 2011).

Steps in production function analysis

1. *Statement of the problem and selection of potentially relevant variables*

The first step in production function analysis is formulation of the problem that is determination of the question(s) to be addressed by the analysis. It helps in selection of relevant variables and in deciding the procedure (sampling design/scheme) to be followed for data collection.

2. *Model specification*

According to Fuss, McFadden, and Mundlak (FMM), “. . . a wide variety of compatible functional forms will usually be available,” and they list five criteria for choosing a single form (Hall, 1998):

- i) parsimony in parameters: excess parameters exacerbate multicollinearity problems and, in small samples, seriously reduce error degrees of freedom,
- ii) ease of interpretation: prefer a form in which parameters have an intrinsic and intuitive economic interpretation and in which functional structure is clear,
- iii) computational ease: although non-linear forms are feasible, linear-in-parameters systems have less expensive computations and more fully developed statistical theory,
- iv) interpolative robustness: the chosen functional form should be consistent with maintained hypotheses in the range of the data and
- v) extrapolative robustness: the chosen functional form should be consistent with maintained hypotheses outside the range of the data.

For empirical analysis, it is necessary to specify the functional form of the production process which meets the economically reasonable restrictions. Any functional form needs to be evaluated on the basis of the criteria of theoretical consistency, domain of applicability, flexibility, factual conformity and computational facility (Lau, 1986). For selecting an appropriate functional form, firstly, there is need to examine the relationship between each predictor variable and the response variable. For this end, scatterplots and correlations can be used. Scatterplot helps in examination of relationship whether it is linear or non-linear. Further, it also helps in identification of deviations from the pattern (outliers). For instance, as given in Fig 3 scatter plot (a) shows that a liner model is suitable. Similarly, for X_2 a quadratic model is suitable as there is first increase and after reaching peak, there is decline in response as shown in (b). Plot in (c) shows that third data set (X_3) has one point that distorts the slope intercept of the fit line (outlier). The plot in (d) shows that the fourth data set (X_4) is not suitable for linear fitting, the fitted line being determined essentially by one extreme observation. In case of multiple

independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables. Further, large correlation coefficients in the correlation matrix of predictor variables indicate the presence of multicollinearity.

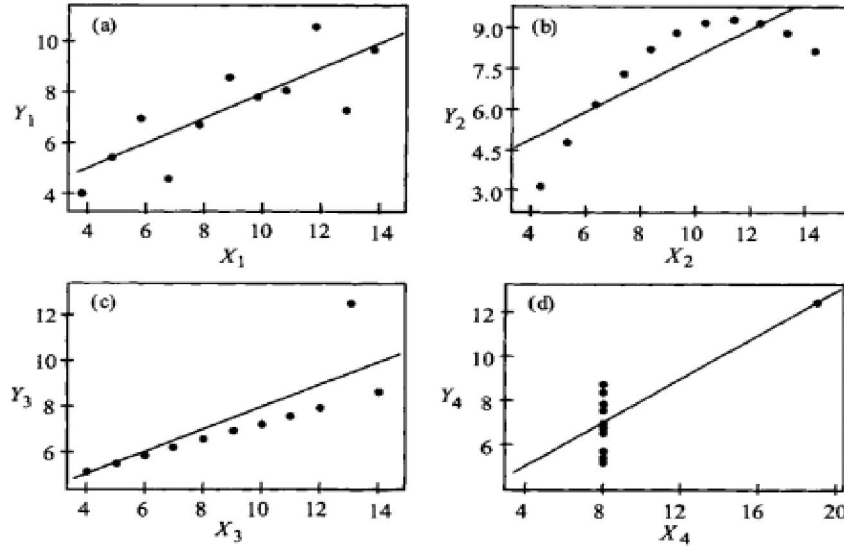


Fig 3: Scatter plots of the data (adopted from Chatterjee and Hadi, 2015)

3. Method of estimation

Depending on the selected form of model (linear or non-linear), a suitable method of parameter estimation can be used. The most commonly used methods are the ordinary least square, GLS (generalized least square) and maximum likelihood method.

4. Selection of appropriate functional form

The objective of model selection is to select a simple model that “best” explains or predicts the data. The most common approach for comparison and selection of models is penalized model selection criteria i.e. Akaike Information Criterion (Akaike, 1974) and Bayesian Information Criterion (Schwarz, 1978), and both are available in most statistical software packages.

$$AIC = -2 \ln \mathcal{L}_r + 2k$$

$$BIC = -2 \ln \mathcal{L}_r + k \ln n$$

Here, \mathcal{L}_r and k are the likelihood estimate and number of parameters to be estimated under model r , respectively. The model with the smallest AIC/BIC value is deemed as the best among the compared. $Adj R^2$ is another useful criterion for model selection. It is used to compare two or more models having the same dependent variable.

$$Adj R^2 = 1 - (1 - R^2) \left(\frac{n-1}{n-k} \right)$$

where

n = number of observations

k = number of parameters to be estimated including intercept.

R^2 = coefficient of determination

5. Residual tests and diagnostic plots

Residual vs fitted plot is used to test the adequacy of the assumed model between the dependent and independent variable. When residuals “bounce randomly” around the 0 line (Fig 4), then it is suggested that the assumption that the relationship is linear is reasonable. When residuals roughly form a “horizontal band” around the 0 line, then it can be said that the error term have zero mean and constant variance (homoscedasticity). If there are patterns and trends in residual, then, there is a need to transform the response variable to obtain the homogeneity in error variance.

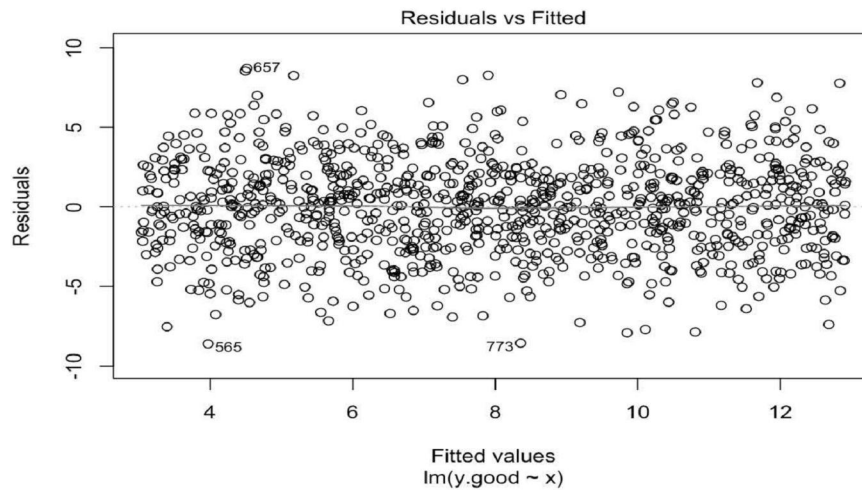


Fig 4: Residuals vs fitted plot showing linear relationship (Chouldechova, 2019)

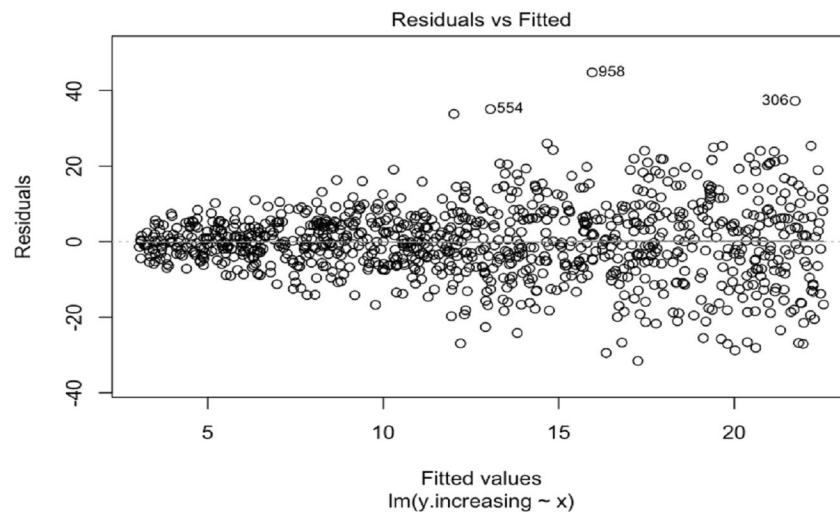


Fig 5: Residuals vs fitted plot showing heteroscedasticity (Chouldechova, 2019)

Further, if no residual stands out from the random pattern of residuals, then the presence of outliers can be ruled out. Moreover, brief description of some challenges likely to encounter during the estimation of production function through regression are given Table 2:

Table 2: Some challenges in the estimation of production function

Assumption	Plot/ statistical test	Probable solution
Multicollinearity	Variance inflation factor (VIF) values are greater than 10 may merit further investigation, Farrar-Glauber test	Principal component analysis, Ridge regression
Heteroscedasticity	White's test, Breusch-Pagan test.	Weighted least square (WLS), Feasible GLS, use Robust errors
Check for outliers	Scatter plot, box-plot	If many, then robust regression
Autocorrelation test	Durbin-Watson test	Suitable transformation
Normality of residuals	Normal QQ plot, Anderson-Darling, Kolmogorov-Smirnov, Shapiro-Wilk	Transformation of response variable following Box-Cox method
Model specification	Regression specification error test (RESET) for omitted variables	-

ILLUSTRATION

Example 1: Steps in production function analysis

Plot level data of maize cultivation were taken for analysis, and variables HL, Fert, SEED and MH stand for human labour, fertilizer, seed rate and machine hours. Yield is a dependent variable represented by y.

Scatter plot of variables

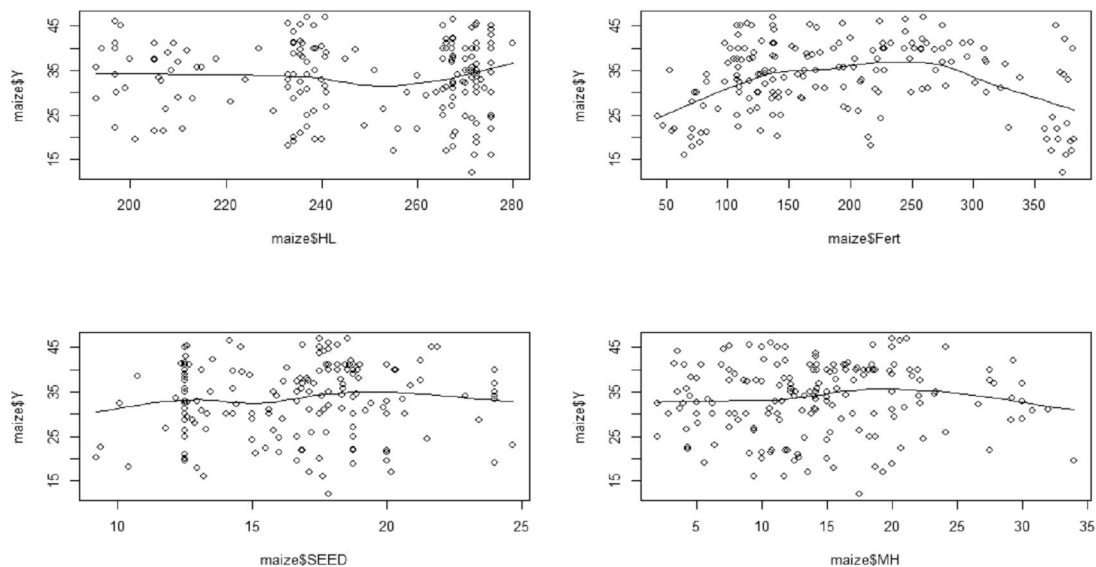


Fig 6: Scatter plots

From the scatter plot (Fig 6), it seems that there is a quadratic relationship between the yield and fertilizer (Fert). Further, variable seed, human labour and machine hours seems to have no relationship, therefore, can be discarded. With the help of scatter plots, it can be stated that a quadratic model seems to be better choice than the simple linear model, and the same is indicated by the lower values of AIC and BIC in case of quadratic function (Table 3).

Table 3: Estimation of linear and quadratic production function

Variables	Linear model				Quadratic model			
	Estimate	Std. Error	t value	Pr(> t)	Estimate	Std. Error	t value	Pr(> t)
Intercept	33589.14	1404.42	23.92	0.000	15260.00	2739.00	5.57	0.000
Fert	-1.024	6.587	-0.155	0.877	206.20	28.30	7.29	0.000
I(Fert^2)					-0.474	0.063	-7.479	0.000
DF	172				171			
F-statistic	0.0242			0.877	27.98			0.000
R-square	0.0001405				0.247			
Adj. R-square	-0.005376				0.2997			
AIC	3624				3577			
BIC	3634				3589			

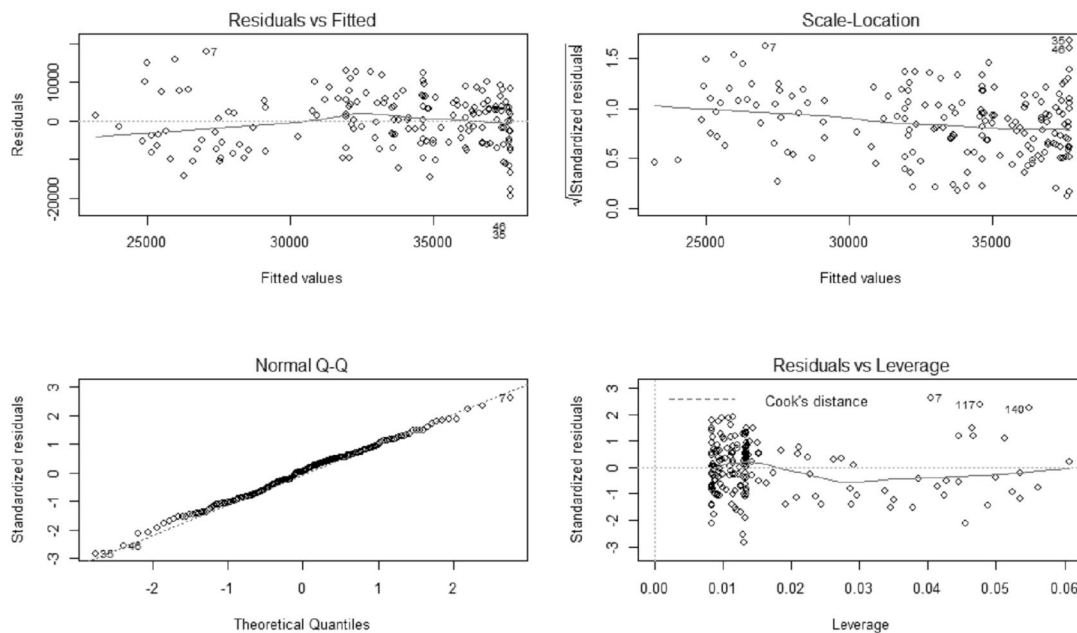


Fig 7: Residual diagnosis (quadratic model)

Residual diagnosis plots are depicted in Figure 7. Residual vs fitted plot is used to examine the random behaviour/pattern of error terms. A horizontal line without distinct patterns is an indication that error term is random with zero mean. Further, in normal Q-Q plot, if points follow the straight dashed line, this implies that residuals

are following the normal distribution. Scale-location is used to check the homogeneity of variance of the residuals i.e. homoscedasticity. Here, a horizontal line with equally spread points is a good indication of homoscedasticity. However, in our case this indicates that there is problem of heteroscedasticity, and the same is also indicated by Breusch-Pagan test (Table 4), and therefore to correct the inflated standard errors, robust standard errors are given. Further, it can be seen that observations 7, 45, and 46 are outlier points. Therefore, these observations have to be discarded for further analysis. Durbin-Watson statistics is near 2, which shows that there is no autocorrelation problem in data set.

Table 4: Estimation of linear and quadratic production function (after discarding influential variables)

Variables	Estimate	Robust Std. Error	t value	Pr(> t)
Intercept	13780.00	2578.100	5.303	0.000
Fert	226.10	27.663	8.386	0.000
I (Fert^2)	-0.523	0.067	-8.639	0.000
DF	168			
F-statistic	37.38			0.000
R-square	0.31			
Adj. R-square	0.30			
AIC	3495			
BIC	3507			
Breusch-Pagan	BP = 7.461, df=2, p-value = 0.02398			
Durbin-Watson	DW = 2.0312, p-value= 0.5762			

Example 2: Estimation of restricted Cobb Douglas production function

Data used for illustration is cross-sectional, covering 50 states for year 2005 (Gujarati, 2015). We estimated Cobb Douglas model given below and results are given in table 5. Here, R^2 is very high (0.964) which is not expected to be so high in cross-sectional data having heterogeneous states (Gujarati, 2015).

$$Y_i = b_1 L_i^{b_2} K_i^{b_3}$$

For estimation, the model was linearized as follows:

$$\ln Y_i = b_1 + b_2 \ln K_i + b_3 \ln L_i + u_i$$

where,

Y_i, K_i, L_i and u_i are output, capital, labour and random error terms with usual assumptions, respectively.

However, the correlation between labour and capital is expected to be very high i.e. 0.94 as the larger states tend to have more capital and hence more labour. Therefore, one can suspect the presence of multicollinearity, and the same is evident from VIF value (12.9 > 10). As evident from the Breusch-Pagan (BP = 4.8641, p-value = 0.088) and Durbin-Watson test (DW = 1.9464, p-value = 0.4361) that there is no issue of heteroscedasticity and autocorrelation.

Table 5: Estimation of Cobb Douglas production function

Variables	Estimate	Std. Error	t value	Pr(> t)
Intercept	3.8876	0.39623	9.812	0.000
lnlabor	0.46833	0.09893	4.734	0.000
lncapital	0.52128	0.09689	5.38	0.000
DF	48			
F-statistic	645.9			0.000
R-square	0.964			
Adj. R-square	0.962			
AIC	14.85			
BIC	22.58			
VIF	12.9			
Breusch-Pagan	BP = 4.8641, df = 2, p-value = 0.087			
Durbin-Watson	DW = 1.9464, p-value = 0.4361			

As we know, for estimating output both capital and labour play an important role and hence cannot be excluded from the production function.

Table 6: Estimation of Restricted (constant return to scale) Cobb-Douglas production function

Variables	Estimate	Std. Error	t value	Pr(> t)
Intercept	3.75624	0.18537	20.264	0.000
Ln(labor/capital)	0.52376	0.09581	5.466	0.000
DF	49			
F-statistic	29.88			
R-square	0.378			
Adj. R-square	0.366			
AIC	13.00			
BIC	18.79			
Breusch-Pagan	BP=2.254, p-value = 0.133			
Durbin-Watson	DW = 1.9368, p-value = 0.4178			

From table 5, it can be seen that sum of output-labor and output-capital elasticities is 0.99, which is about 1. Therefore, it can be said that this is a case of constant return to scale. Keeping this in view, we estimated aforesaid equation in the following manner (restricted equation, $b_2 = 1 - b_3$):

$$\ln\left(\frac{Y_i}{L_i}\right) = b_1 + b_2 \ln\left(\frac{K_i}{L_i}\right) + u_i$$

where,

$\frac{Y_i}{L_i}$ is the output-labour ratio and $\frac{K_i}{L_i}$ is capital-labour ratio. It can be seen and AIC and BIC values have declined indicating the better performance of the model (table 6) and now R-square is come down as expected in case of cross-sectional data.

R codes for analysis (Example 1)

```
attach(maize)
library(olsrr)
library(MASS)
library(lmtest)
library(sandwich)
summary(maize)
# scatter plots
layout(matrix(c(1,2,3,4),2,2))
scatter.smooth(x=maize$HL, y=maize$Y)
scatter.smooth(x=maize$SEED, y=maize$Y)
scatter.smooth(x=maize$Fert, y=maize$Y)
scatter.smooth(x=maize$MH, y=maize$Y)
# fitting linear model
model<-lm(y~Fert, data=maize)
summary(model)
AIC(model)
BIC(model)
# plots for residual diagnosis
plot(model)
## fitting quadratic model
model2<-lm(y~Fert+I(Fert^2), data= maize)
summary(model2)
AIC(model2)
BIC(model2)
#Outlier diagnosis
plot(model2)
#Breusch-Pagan test for heteroscedasticity
bptest(model2)
#Durbin-Watson test for autocorrelations
dwtest(model2)
# for robust standard errors
coefest(model2, vcov = vcovHC(model2, type="HC1"))
#Example 2 Restricted CD function
attach(exam)
```

```
lnoutput
cdmodel<-lm(lnoutput~lnlabor +lncapital, data=exam)
summary(cdmodel)
dwtest(cdmodel)
AIC(cdmodel)
BIC(cdmodel)
bptest(cdmodel)
#restricted model
rcdmodel<-lm(lnoutlab~lncaplab, data=exam)
summary(rcdmodel)
dwtest(rcdmodel)
AIC(rcdmodel)
BIC(rcdmodel)
bptest(rcdmodel)
```

REFERENCES

- Akaike, H. (1974), A new look at the statistical model identification. In Selected Papers of Hirotugu Akaike, 215-222. Springer, New York, NY.
- Arrow, K. J., H. B. Chenery, B. S. Minhas and R. M. Solow (1961), Capital-labor substitution and economic efficiency. *The review of Economics and Statistics*, 43(3): 225-250.
- Chatterjee, S. and A. S. Hadi (2015), Regression analysis by example. John Wiley & Sons.
- Chouldechova (2019), Regression diagnostic plots, accessed 2019, Retrieved from https://www.andrew.cmu.edu/user/achoulde/94842/homework/regression_diagnostics.html
- Cobb, C. W. and P. H. Douglas (1928), A theory of production. In Proceedings of the Fortieth Annual Meeting of the American Economic Association, 18(1): 139-165.
- Diewert, W. E. (1971), An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy*, 79(3): 481-507.
- Griffin, R. C., J. M. Montgomery and M. E. Rister (1987), Selecting functional form in production function analysis. *Western Journal of Agricultural Economics*, 12(1836-2016-150929): 216-227.
- Gujarati, D. N. (2015), Econometrics by example, Palgrave.
- Hackman, S. T. (2007), Production economics: integrating the microeconomic and engineering perspectives. Springer Science & Business Media.
- Hall, H. H. (1998), Choosing an empirical production function: Theory, non-nested hypotheses, costs of specifications (No. 1638-2016-135136).

- Heathfield, D. F. (2016), An introduction to cost and production functions. Macmillan International Higher Education.
- Klacek, J., M. Vošvrda and Š. Schlosser (2007), KLE Translog production function and total factor productivity. *Statistika*, 87(4): 261-274.
- Lau, L. J. (1986), Functional forms in econometric model building. Handbook of econometrics, 3: 1515-1566.
- Miller, E. (2008), An assessment of CES and Cobb-Douglas production functions (pp. 2008-2005). Congressional Budget Office.
- Pavelescu, F. M. (2011), Some aspects of the translog production function estimation. *Romanian Journal of Economics*, 32(1): 41.
- Schwarz, G. (1978), Estimating the dimension of a model. *The annals of statistics*, 6(2): 461-464.

Chapter 32

SOCIAL NETWORK ANALYSIS

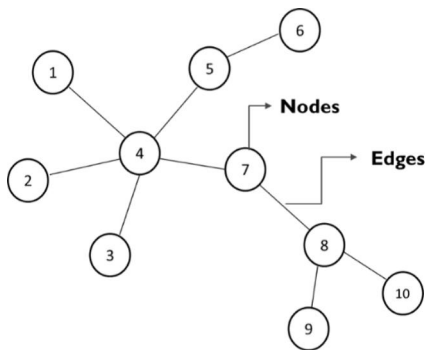
Subash S. P.

INTRODUCTION

Social Network Analysis (SNA) is a methodology to map and quantify actors (nodes) and their relationship in a network (Box 1). SNA with its wide range of utility has evolved with applications in various disciplines. The network perspective is becoming a key approach in social and biological sciences (Borgatti and Li, 2009). Social network analysis is an effective tool in understanding the social relations and interactions of the individuals in the group (Borgatti *et al.*, 2009). It helps to understand the actors and the relationship between them in a specific social context (Clark, 2006a).

Box 1: What is a network?

Network is diagram connecting points by lines. The points and lines are named differently according to the disciplines.



Points	Lines	Discipline
Vertices	Edges, Arcs	Math
Nodes	Links	Computer Science
Sites	Bonds	Physics
Actors	Ties, Relation	Sociology

SNA has been utilized to explore various dimensions in social sciences. Studies used SNA to map and quantify the value chains (Lazzarini, 2001; Borgatti and Li, 2009; Trienekens, 2011; Bellamy and Basole, 2013), monitoring and impact assessment (Ekboir *et al.*, 2011), and to understand adoption of technology (Matuschke, 2008; Magnan *et al.*, 2015) and rural innovations (Spielman *et al.*, 2010). Lazzarini *et al.* (2001) introduced the concept of netchain as a technique for value chain analysis exploring the interrelationship different actors in the value chain. Ekboir *et al.* (2011) used SNA to map collaboration between researchers and organizations. Ramirez (2013) applied social network analysis to analyses the social interaction influencing the decision-making behaviour on adoption of new

technology. Value chain analysis approach has evolved from adding value within the firm to between the firms.

SNA could also be used to understand and identify key farmers or informants (actors) in a village. This could help us to transfer the information or technology effectively and efficiently. Studies by Clark (2006a), Douthwaite *et al.* (2006), Misra *et al.* (2014), Wood *et al.* (2014) gave some directions in this regard. Clark (2006a) and Douthwaite *et al.* (2006) came up with two case studies from Bolivia and Colombia to strengthen agricultural supply chain using SNA. Though the studies just focused on mapping the supply chain it gave linkages between various actors in the supply chain which could be helpful in building farmers capacities for networking. Application of Social network analysis to understand the knowledge networks were argued as highly effective method by Müller-Prothmann (2007) and Krätke (2010) though Helms (2010) had criticised its limitations in understanding the efficiency and effectiveness of knowledge sharing. Wood *et al.* (2014) used SNA to map knowledge sharing of 17 farmers and five scientists with a network of more than 192 people. The results showed that farmer accept the knowledge delivered by particular person irrespective of roles and farming experience. Recently Mittal *et al.* (2017), Vishnu *et al.* (2019) used SNA to map and quantify information networks using SNA. Mittal *et al.* (2015, 2017) studied knowledge and social networks of farmers in India based on their case study on key informants in Bihar. Vishnu *et al.* (2019) used SNA to map important information sources as well as patterns of information access by livestock farmers. Other than mapping and quantifying network Matuschke (2008) had outlined research approach of combining SNA with econometric techniques to evaluate the impact of network in technology adoption.

Genesis

Earlier works on social network analysis dates back to 1930's (Scott, 2000 for history of SNA). Kohler worked on mind and other Gestalt tradition researchers work was the basis for social network theory. Other three Gestalt scientist Moreno (sociogram), Lewin (group behavior) and Heider (balance theory) further developed the theory. Morenos sociogram allowed researchers to visualize relations among different groups. Lewin further employed mathematical techniques such as topology and set theory to explore the social space, but Koenig (1936) graph theory provided the crucial breakthrough in application of mathematical concept in sociometric analysis. On the other hand, anthropologist Radcliffe-Brown developed in a relatively nontechnical form of social network analysis. Other anthropologist and sociologist like Warner and Mayo build on his concepts (Fig 1 depicting lineage of SNA). The current form SNA is developed based on socio-metric technique, graph theory and computer algorithms.

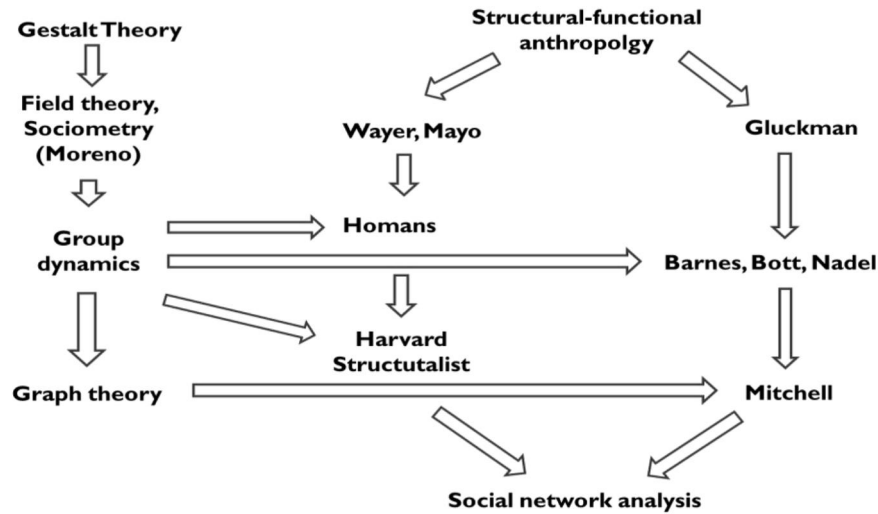


Fig 1: Lineage of social network analysis (Scott, 2000).

MAPPING

The network analysis could be done by visualizing network maps (Freeman, 2000 and Liebowitz, 2005) or using quantitative network measures (Freeman, 1979 and Landherr *et al.*, 2010).

Network maps

The network maps are visual representation of the relationship. Interpretations can be drawn from understanding the major source of information from the large number of actors attached to the nodes. The relationship is represented in lines or arrows. The position of the nodes represents the importance of the source of information. They are positioned in centre, periphery or isolated based on their importance (Bartholomay and Chazdon, 2011). The attributes help us to distinguish the actors based on their characteristics to draw inference regarding the trait which are associated with the selection of a particular source. Colouring or giving shape to nodes/actors based on their attributes help as to group them (Fig 2).

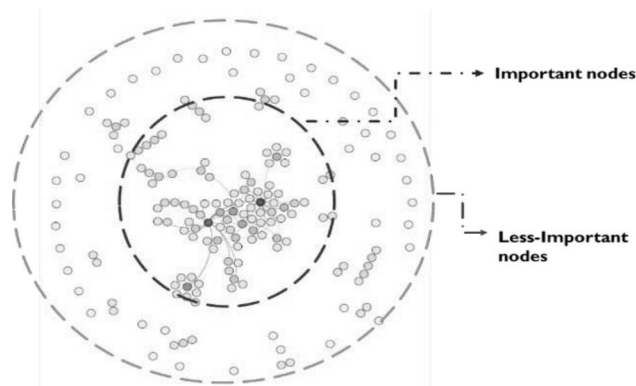


Fig 2: Network maps

Note: The dotted circles are embedded in the image for explanation.

QUANTITATIVE ANALYSIS

Density

Density is the level of connectedness in the graph. It is defined as the number of actual links divided by number of possible links.

Centrality measures

Network centrality is a common measures employed in social network analysis. Calculating centrality in the networks is an important measure quoted in various reviews (Freeman, 1979; Valente *et al.*, 2008). Centrality measures are used to identify the most important (key) player in the network (Scott, 2000; Landerr *et al.*, 2010). There are different approaches/methods to analyse centrality measures mentioned conceptually (Freeman, 1979; Borgatti, 2005) and empirically (Costenbander and Valente, 2003; Borgatti, 2006; Kiss and Bichler, 2008). The commonly employed measures of centrality are degree, betweenness, closeness and eigenvector (Valente *et al.*, 2008; Landherr *et al.*, 2010) (Table 1). The first three measures were reviewed and put forward by Freeman (1978/79), while eigen vector measure of centrality is given by Bonacich (1972). Though these measures are distinct, a certain level of correlation exists between them due to their conceptual relation (Valente *et al.*, 2008).

Table 1: Centrality measures

S. No.	Centrality measures	Description
1	Degree	It denotes the number of ties a node has. More the number of ties the higher the level of centrality and vice versa. It outlines the importance of a particular actor in the network.
2	Closeness	It measures the number of ties between nodes and all other nodes. It is used to measure the length of time it takes for information to pass between a node to other nodes. It identifies the actor which is more productive in spread of the information. The shorter the distance between the nodes the more productive is the information spread and vice versa.
3	Betweenness	It measures the number of times a node falls along the shortest path between two other nodes. It measures the control of network; nodes having higher betweenness have ability to hinder or change information passed along them. It shows how well the actors are connected.
4	Eigen Vector	It is a measure of quality of the nodes connected. It counts and assigns weights according to the centrality. It measures the popularity of a particular actor (Bonaich and Llyod, 2001).

There are several software's to analyse network (Subash and Vishnu, 2017). They could be categorised into packages using graphical use interface (GUI) [UCINET, Gephi]

and those uses script (R. Python). In this chapter, we discuss analysing network data using Netdraw (an open access component of UCINET) and Gephi.

SNA using Netdraw

The example of survey format and data collection was given by Clarke (2006b). The graphics in social network analysis are based on two types of information:

- 1) nodes which represent the people or institutions, and
- 2) ties; the different relationships among the actors or nodes.

The step wise procedure of analyzing data in Netdraw is given below

Step 1: Encoding data in Excel

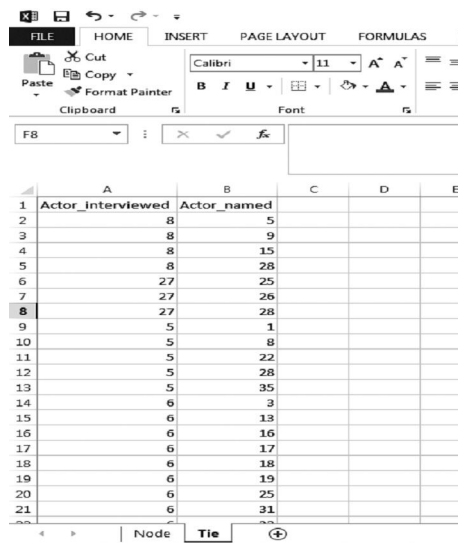
Once the surveys have been completed the data needs to be transferred into an Excel database. Two databases should be created: one with the information on the nodes and their attributes, and the other with the tie information. It is easier to start with the database on the ties as this can then be used to create the database on the nodes.

Creating a tie database

Excel is used to create a two-column database (Fig 3).

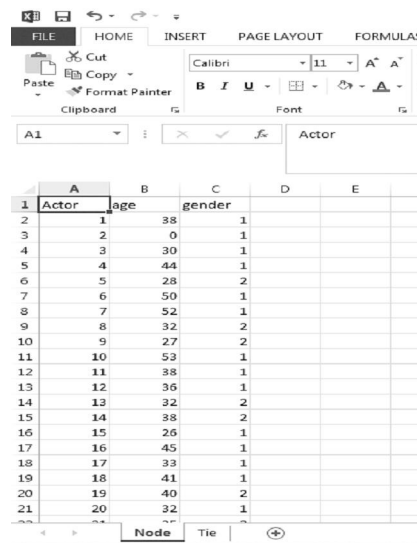
1. Actors interviewed (the name of the person interviewed is repeated for each person they have named),
2. Actors named,

The first two columns follow the same format in all social network studies, while you can add other variables such (type of information, means by which the information was accessed etc.) in other columns depending on the variables chosen for the study. It is possible to have more than one tie between the same two actors demonstrating that more than one type of information flows between them.



	A	B	C	D	E
1	Actor_interviewed	Actor_named			
2		8	5		
3		8	9		
4		8	15		
5		8	28		
6		27	25		
7		27	26		
8		27	28		
9		5	1		
10		5	8		
11		5	22		
12		5	28		
13		5	35		
14		6	3		
15		6	13		
16		6	16		
17		6	17		
18		6	18		
19		6	19		
20		6	25		
21		6	31		

Fig 3



	A	B	C	D	E
1	Actor	age	gender		
2	1	38	1		
3	2	0	1		
4	3	30	1		
5	4	44	1		
6	5	28	2		
7	6	50	1		
8	7	52	1		
9	8	32	2		
10	9	27	2		
11	10	53	1		
12	11	38	1		
13	12	36	1		
14	13	32	2		
15	14	38	2		
16	15	26	1		
17	16	45	1		
18	17	33	1		
19	18	41	1		
20	19	40	2		
21	20	32	1		

Fig 4

Creating a database of the node and attributes

You can create node database with attribute in two ways. First you could create a table with each actors and their attributes, second could use the tie database and remove duplicates using 'remove duplicate' option in excel. Create a row with attributes for all the actors and actors named in the tie database (Fig 4).

Use short names or numbers to make the chart less complicated. Also do not use space between names. You could use “_” if space is required.

Step 2: Creating a notepad (.txt) file

Open a notepad file and in the first line add *Node data. Copy and paste the table created in node database. At the end of the copied table in notepad start a new line *Tie data (Fig 5).

```

SNA IIM FDP data - Notepad
File Edit Format View Help

*Node data
Actor age gender work_exp dorm floor room_no group
1 38 1 10 31 2 22 1
2 0 1 0 31 4 37 4
3 30 1 2 31 3 31 5
4 44 1 20 31 3 29 3
5 28 2 4 31 1 9 2
6 50 1 18 31 2 20 5
7 52 1 28 31 2 18 4
8 32 2 8 31 1 10 3
9 27 2 3 31 1 11 1
10 53 1 33 31 2 17 3
11 38 1 16 31 4 36 6
12 36 1 9 31 2 23 1
13 32 2 6.5 31 1 12 5
14 38 2 15 31 0 1 2
15 26 1 1 31 4 35 3
16 45 1 20 31 2 19 6
17 33 1 6 31 3 27 5
18 41 1 13 31 3 25 4
19 40 2 14 31 0 2 5
20 32 1 9 31 3 26 6
21 35 2 12 31 0 3 4
22 27 1 3 31 4 39 2
23 32 1 10 31 4 34 4
24 30 2 6 31 1 13 5
25 46 2 20 31 0 4 6
26 36 2 6 31 1 14 4
27 47 2 15 31 0 5 1
28 0 2 9 31 1 15 3
29 29 1 4 31 4 33 1
30 43 2 19 31 0 6 3
31 0 1 0 31 3 3 6
32 35 1 10 31 2 24 1
33 50 2 20 31 0 7 6
34 33 1 10 31 3 31 2
35 0 2 7 31 1 16 2

*Tie data
Actor Actor_named
8 5
8 9
8 15
8 28
27 25
27 26
27 28
5 1
5 8
5 22
5 28
5 35

```

Fig 5

Step 3: Importing the notepad (.txt) file in Netdraw

Netdraw reads files in different formats, we could use the “vna” format, as it is the simplest way to convert data from notepad. Once the software is installed, open the programme and go to Open where various options will appear. Click on Vna Text File and then Complete (Fig 6). With the tie and node data combined all of the information necessary is in one file. Once the correct file has been found open it in this window.

A new window with two columns will appear. In the first column where it says File format mark .vna, and in the second, Type of File mark Network with Attributes (Fig 7). Then the first image of the network will appear (Fig 8).

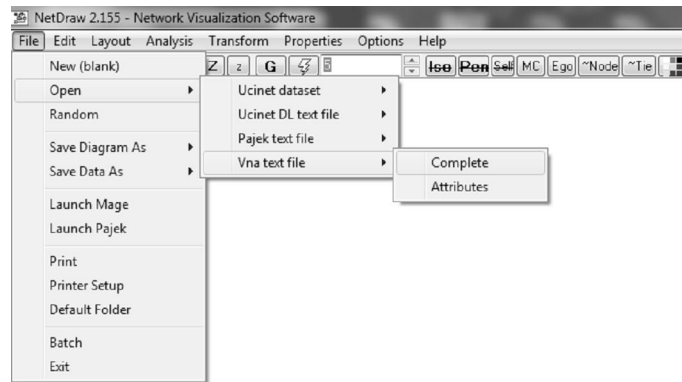


Fig 6

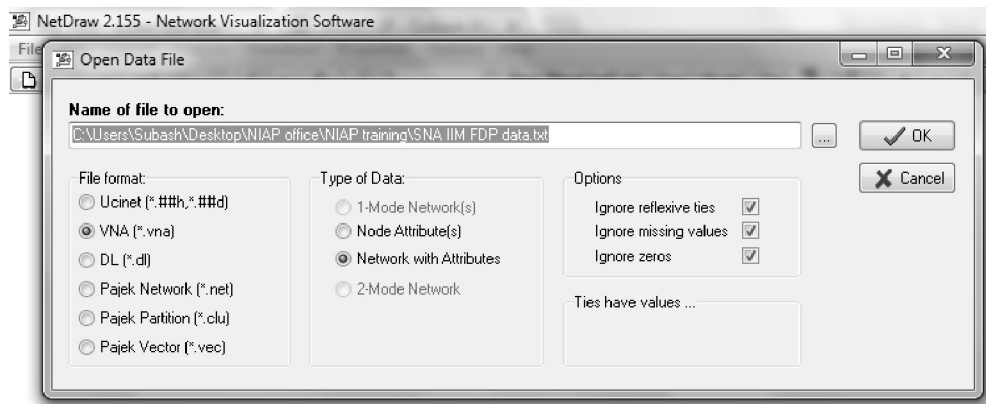


Fig 7

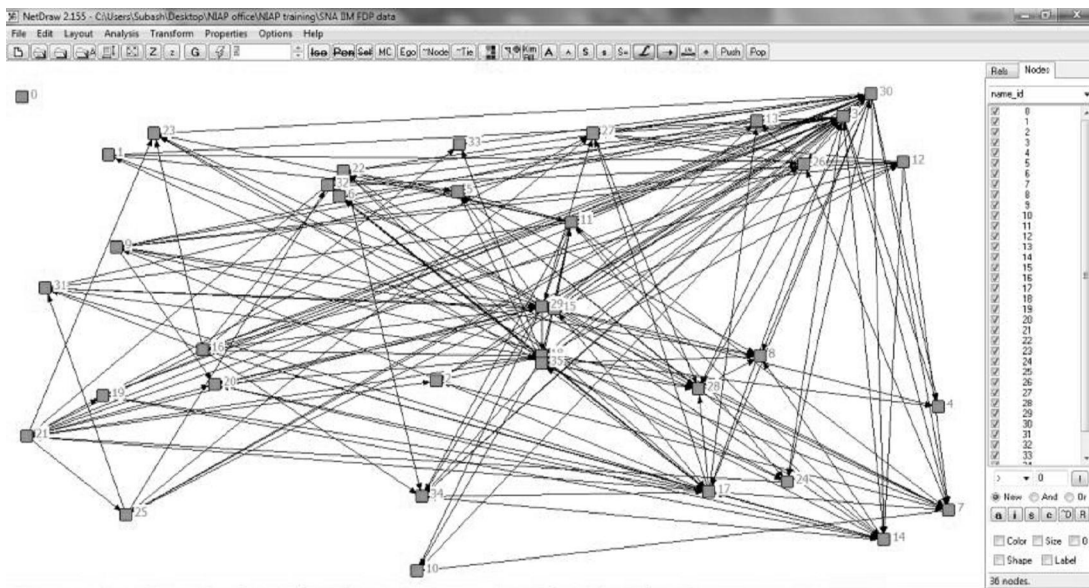


Fig 8

Step 4: Structuring the network

The next step is to give the network some structure. Go to *Layout* → *Graph-theoretic Layout* → *Spring Embedding* (Fig 9a). The window below will appear, click on OK (Fig 9b). Now the network map is easier to read (Fig 10). This step can be repeated until a clearer network image appears. The reader can also experiment with the other layout possibilities to see if they create network structures more visually pleasing to the purposes of other studies. Some loose nodes may appear. This may be due to a mistake in the data (a comma, a space, an extra letter or one missing, etc.). These nodes may also simply not be connected to any other actor in the network. In order to find out why the nodes are isolated it is important to go back to the .txt file and check everything thoroughly. These adjustments are normal and it takes patience to obtain the correct chart.

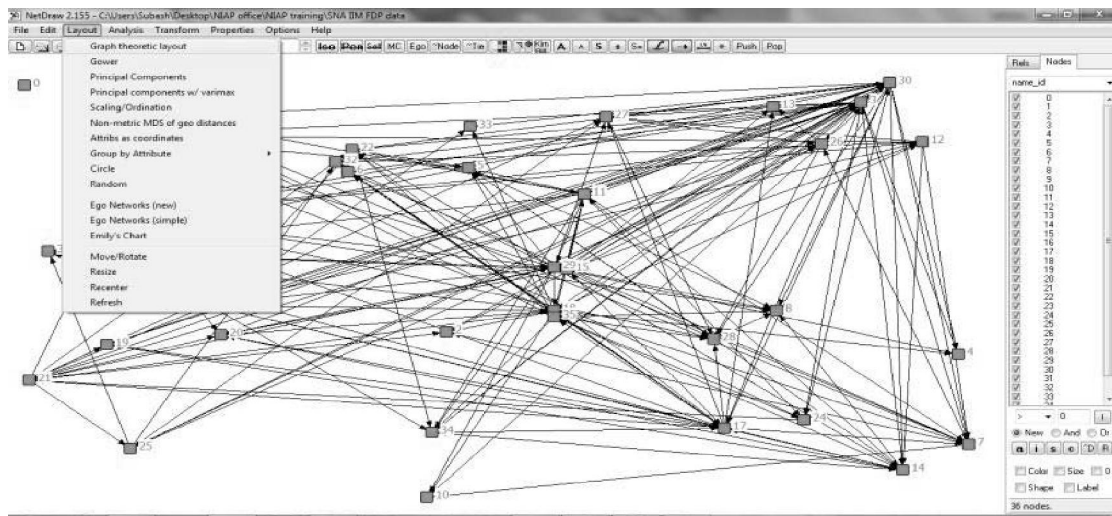


Fig 9a

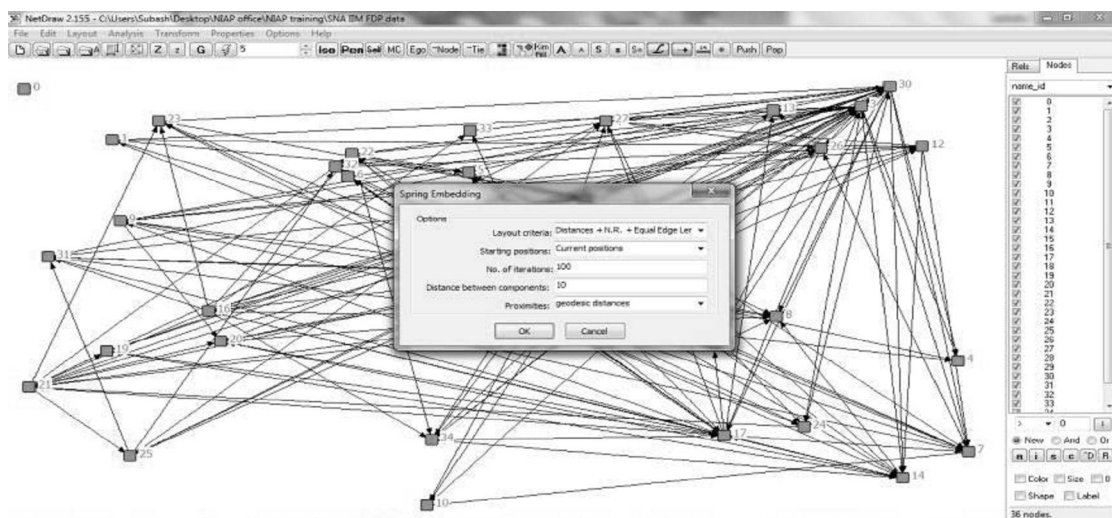


Fig 9b

Social Network Analysis

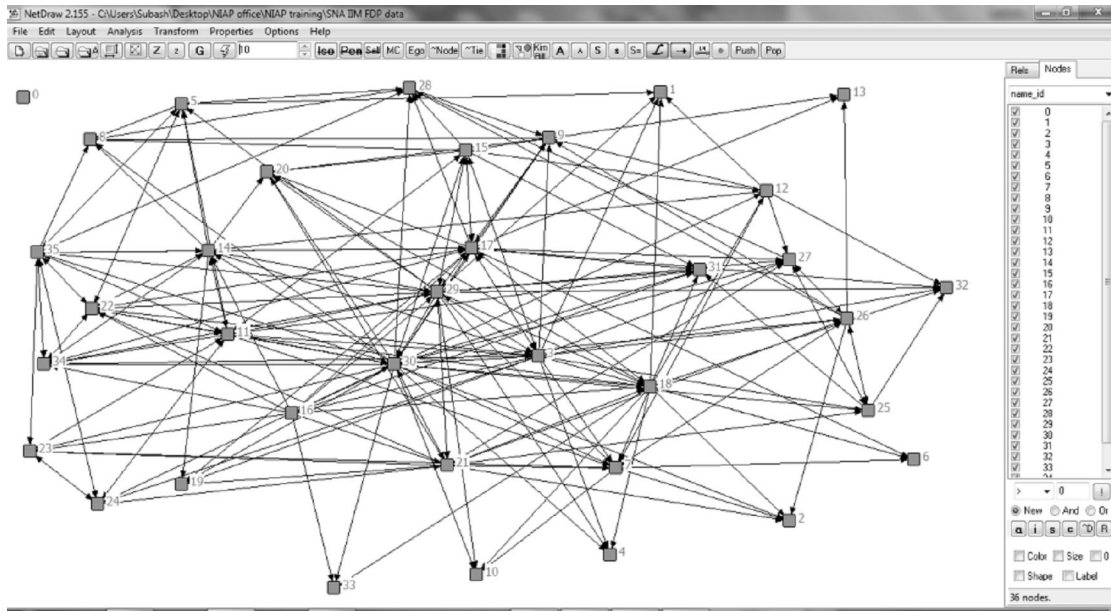


Fig 10

For other operations such as coloring the nodes, ties and changing shape of nodes refer Clarke (2006b).

Step 5: Analyzing data in Netdraw

Once you have a network map, go to analysis → Centrality measures click it (Fig 11). Go back to network diagram. Right click the nodes and click attributes (Fig 12). The centrality measures would be listed in the dialogue box (Fig 13).

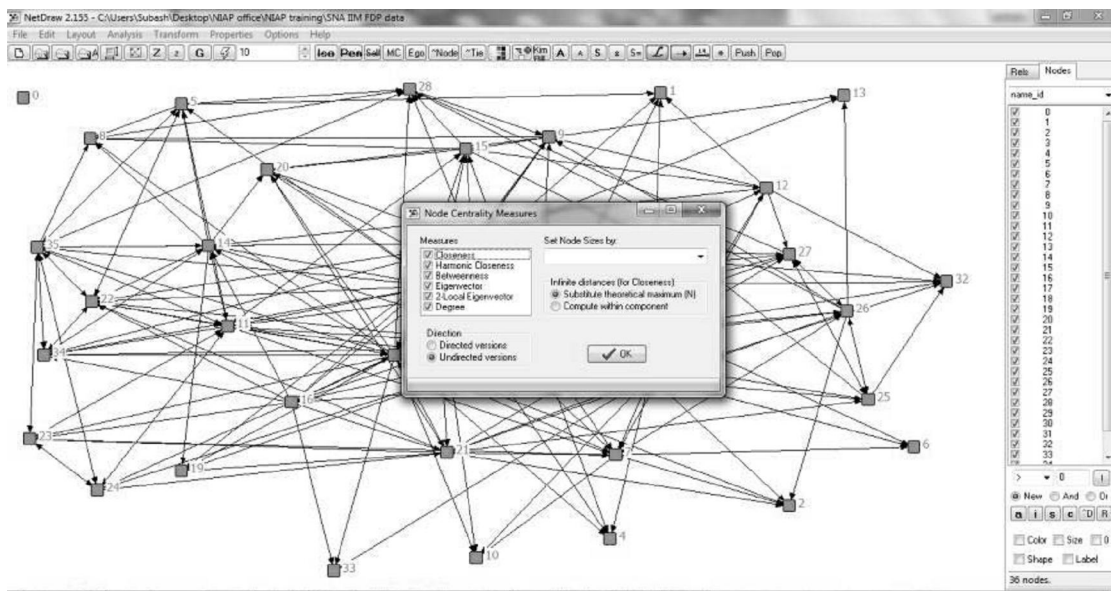


Fig 11

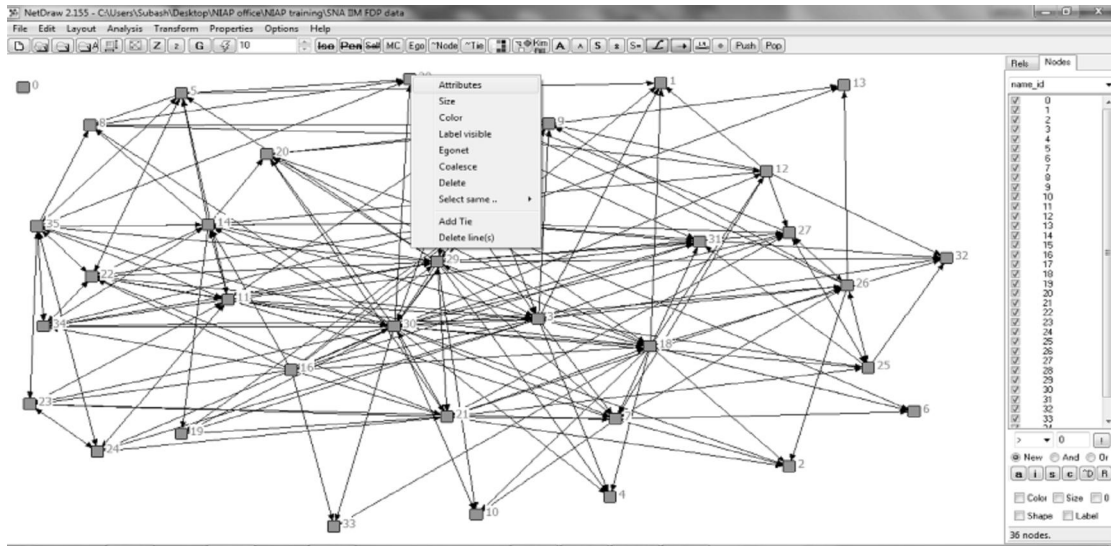


Fig 12

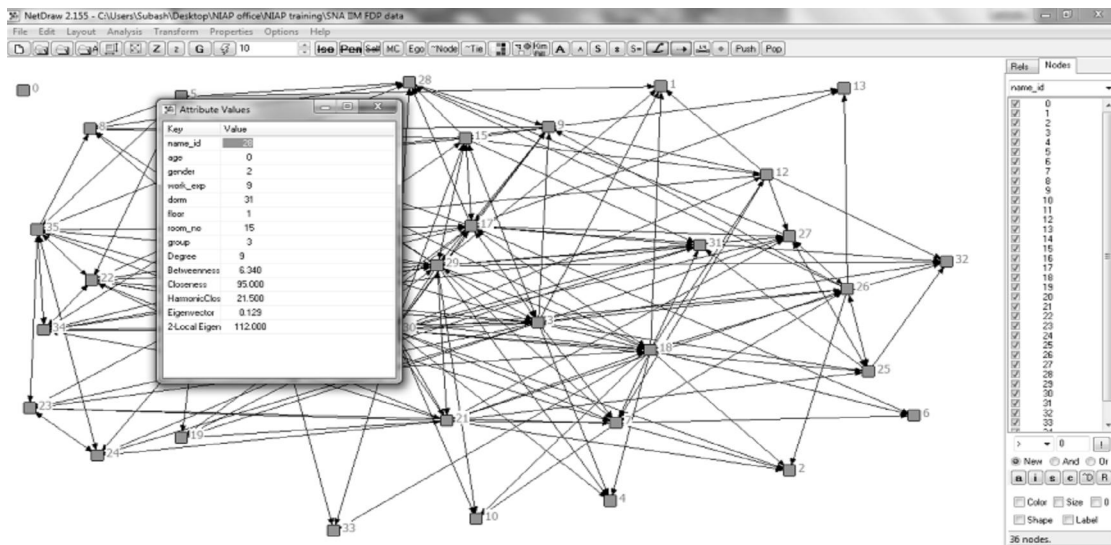


Fig 13

SNA using Gephi

Gephi is another open source package. It is difficult to learn but has a wider functionality specially with larger dataset.

Step1: Data structure in excel

Similar to step in netdraw the data collected from the survey need to be restructured. The data are divided into node data and edges data (Cf. Tie data in netdraw). The sheet in which node data exists have column names Nodes, ID and Label. The sheet in edges data is to be incorporated has column names Source, Targets and Type (undirected/directed) [Fig 14].

Step 2: Save excel file as CSV file

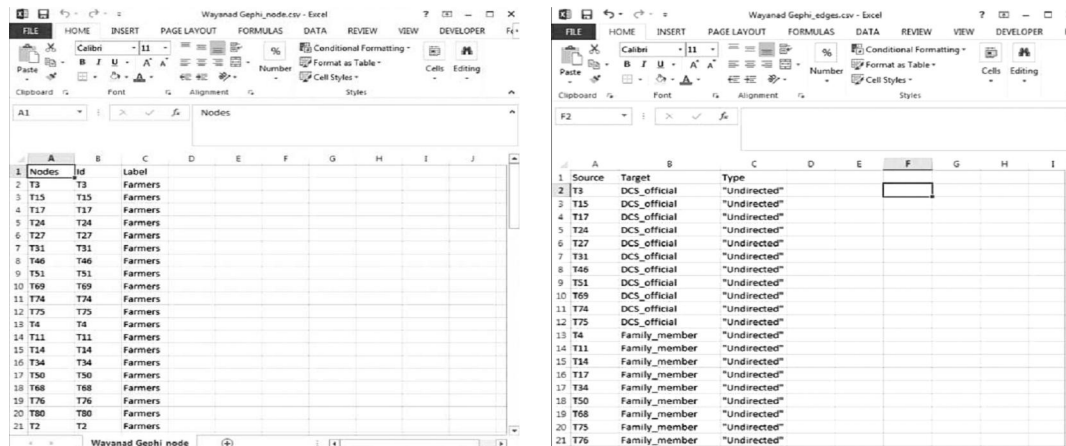


Fig 14

Step 3: Importing csv file to gephi

The csv data files could be imported into gephi by clicking data laboratory (Fig 15). The two tables (nodes and edges) are imported one after another.

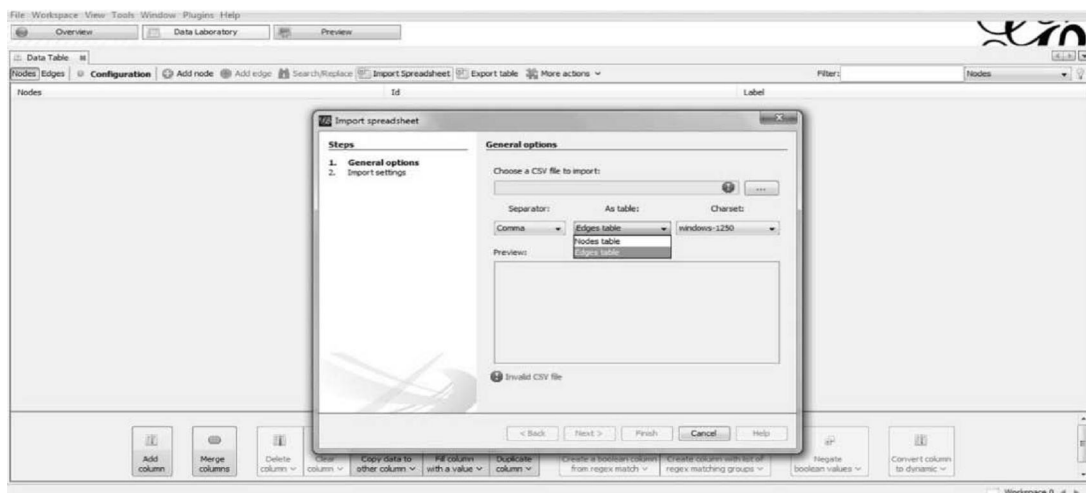


Fig 15

Step 4: Selecting layout

Once the data is imported, we could start analyzing by switching to overview window. In overview window, we could use different layouts (Force Atlas, Force Atlas 2, Random Layout etc.) for visualizing data (Fig 16).

Step 5: Graph adjustments

In the overview window, side and bottom task bar (Fig 16) in graph page, has several options to modify the graph.

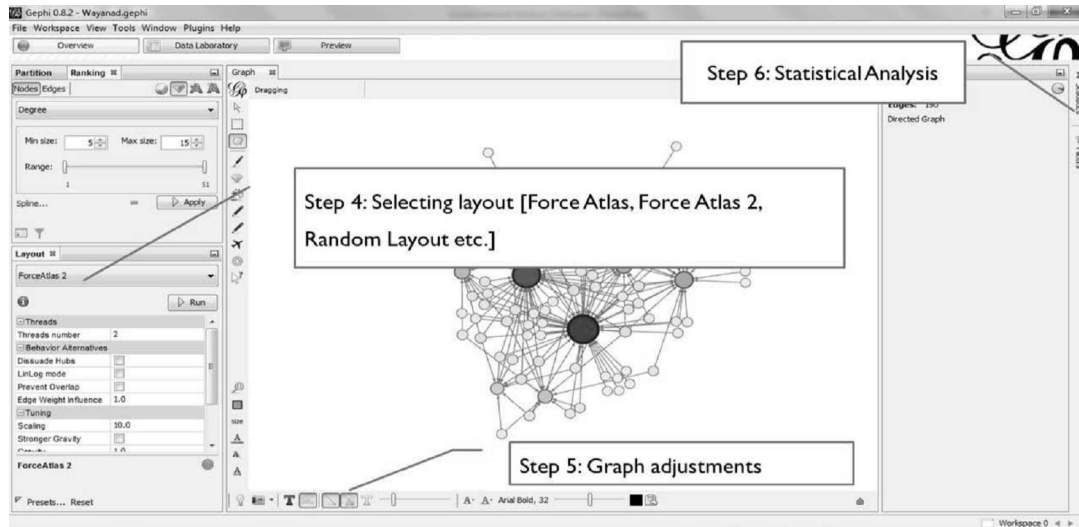


Fig 16

Step 6: Statistical analysis

Qualitative network measures as explained above could be carried out by clicking statistics option (Fig 17, 18). Click run and also selected either directed or un-directed option.

Step 7: Ranking notes/edges

The nodes and edges could be shaped based on the quantitative network measures. Use ranking option in the menu and select the measure based on which node/edges are to be modified (Fig 17).

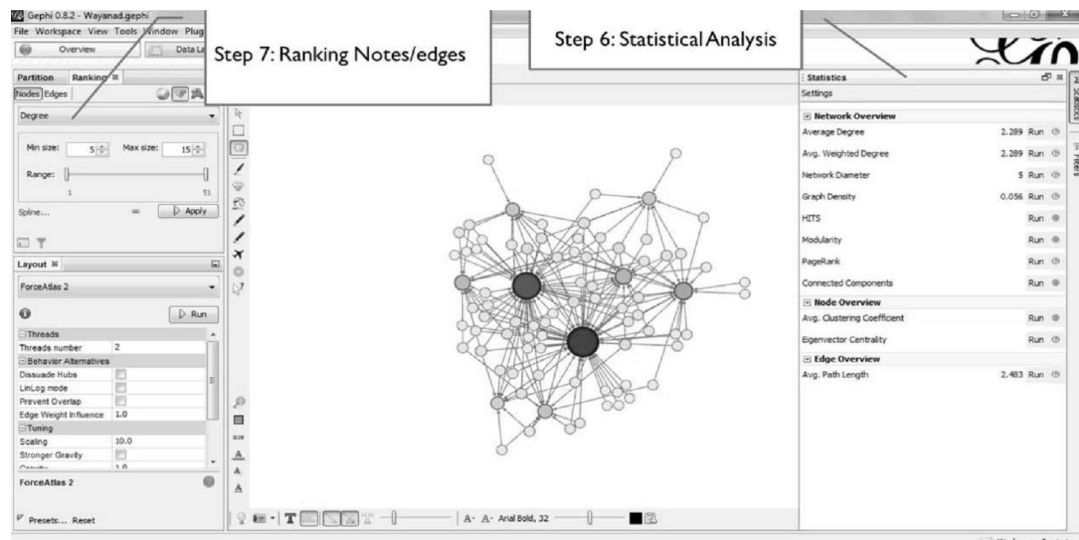


Fig 17

Step 8: Exporting analyzed data

We could also export the quantitative measures analyzed by going back to data laboratory page and clicking export table option (Fig 18).

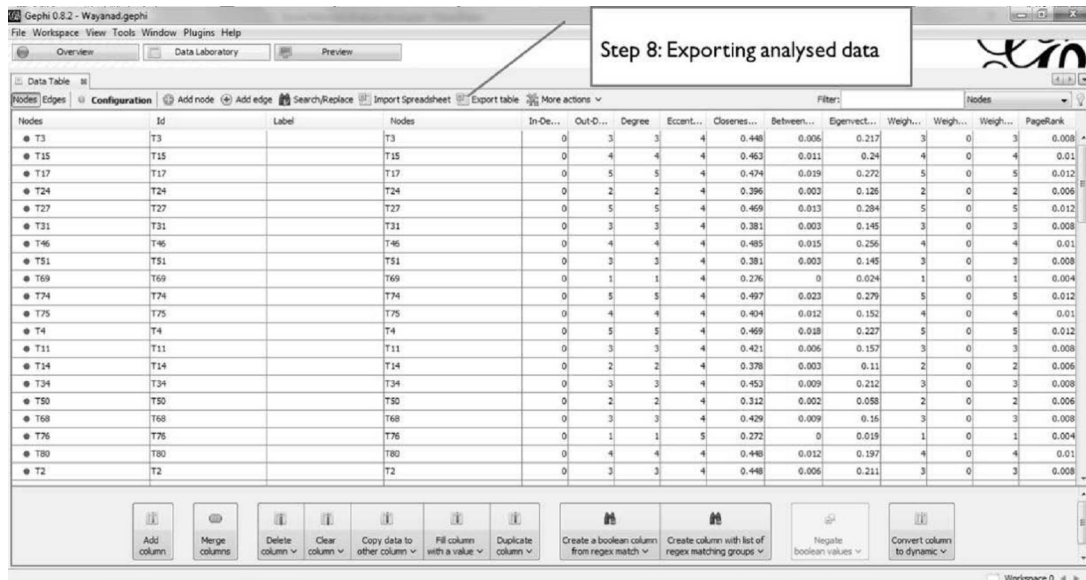


Fig 18

Step 9: Preview

We could see the final data by clicking preview option (Fig 19)

Step 10: Saving the visualization

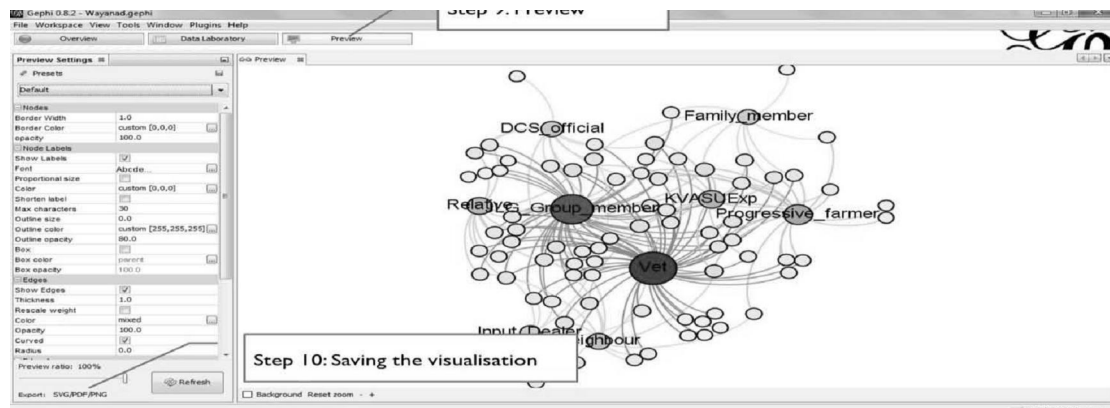


Fig 19

The final network map/figure could be saved either as pdf or as jpeg images (Fig 19).

REFERENCES

- Bartholomay, T. and S. Chazdon (2011), December 2011 Article Number 6FEA9 mapping extension's networks : using social network analysis to explore extension ' s outreach, *49*(6): 1–14.
- Bellamy, M. and R. Basole (2011), Network analysis of supply chain systems: a systematic review and future research. *systems engineering*, *14*(3): 305–326.
- Borgatti, S. P. (2005), Centrality and network flow. *Social Networks*, *27*(1):55–71
- Borgatti, S. P. (2006), Identifying sets of keyplayers in a social network. *Computational and Mathematical Organization Theory*, *12*(1):21–34
- Borgatti, S. P., A. Mehra, D. J. Brass and G. Labianca (2009), Network analysis in the social sciences. *Science*, *323*: 892-895.
- Borgatti, S. and X. Li (2009), On social network analysis in a supply chain context. *Journal of Supply Chain Management*, *45*(2): 1–17.
- Bonacich, P. and P. Lloyd (2001), Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*(3): 191–201. doi:10.1016/S0378-8733(01)00038-7
- Bonacich, P. (1972), Technique for analyzing overlapping memberships. *Sociological Methodology*, *4* :176-185.
- Clark, L. (2006a), Building farmers' capacities for networking (part II) : Strengthening agricultural supply chains in Bolivia using network analysis. *KM4D Journal: Knowledge Management for Development Journal*, *2*(2):19–32.
- Clark, L. (2006b), Network Mapping as a Diagnostic Tool. Centro Internacional de Agricultura Tropical – CIAT. ISBN: 958-694-086-1, La Paz, Bolivia.
- Costenbader, E. and T. W. Valente (2003), The stability of centrality measures when networks are sampled. *Social Networks*, *25*(4):283– 307.
- Douthwaite, B., A. Carvajal, S. Alvarez, E. Claros and L. A. Hernández (2006), Building farmers ' capacities for networking (Part I): Strengthening rural groups in Colombia through network analysis (Part I). *Group, KM4D Journal: Knowledge Management for Development Journal*, 4–18.
- Ekboir, J., G. B. Canto and C. Sette (2011), Monitoring the composition and evolution of the research networks of the CGIAR Research Program on Roots, Tubers and Bananas (RTB). Available online on http://www.cgiar-ilac.org/files/ilac_report_research_networks_rtb_0.pdf. Accessed on 11-07-2015.
- Freeman, L. C. (2000), Visualizing social networks. *Journal of social structure*, *1*(1): 4.
- Freeman, L. C. (1979), Centrality in Social Networks Conceptual Clarification, *Social Networks*. *1*: 215–239.
- Helms, R., R. Ignacio, S. Brinkkemper and A. Zonneveld (2010), Limitations of Network Analysis for Studying Efficiency and Effectiveness of Knowledge Sharing, *8*(1): 53–68.

- Krätke, S. (2010), Regional knowledge networks: A network analysis approach to the interlinking of knowledge resources. *European Urban and Regional Studies*, 17 (1): 83-97.
- Kiss, C. and M. Bichler (2004), Identification of Influencers - Measuring Influence in Customer Networks, (Bone 1995), 1–33.
- Landherr, A., B. Friedl and J. Heidemann (2010), A Critical Review of Centrality Measures in Social Networks. *Business & Information Systems Engineering*, 2(6): 371–385. doi:10.1007/s12599-010-0127-3
- Lazzarini, S., F. Chaddad and M. Cook (2001), Integrating supply chain and network analyses: The study of netchains. *Journal on Chain and Network Science*, 1(1): 7-22.
- Liebowitz, J. (2005). Linking social network analysis with the analytic hierarchy process for knowledge mapping in organizations. *Journal of Knowledge Management*, 9(1): 76–86.
- Magnan, N., D. J. Spielman and T. J. Lybbert (2015), Information Networks among Women and Men and the Demand for an Agricultural Technology in India. IFPRI Discussion paper 01411.
- Matuschke, I. (2008). Evaluating the impact of social networks in rural innovation systems: An overview. *IFPRI Discussion Paper*, (November), 26.
- Misra, S., R. Goswami, I. D. Basu and T. R. J. (2014), Application of Social Network Analysis in Livelihood System Study. *Space and Culture*, 24–46.
- Mittal, S., Subash, S. P., A. Ajay and A. Kumar (2015), Understanding the knowledge and social networks in India-Case study of Bihar. International Conference of Agricultural Economics. August 8-14. Milan, Italy.
- Mittal, S., S. P. Subash and A. Ajay (2018), Agricultural information and knowledge network in rural India: A case of Bihar. *Journal of Agricultural Education and Extension*. 24(5): 393-418.
- Müller-Prothmann, T. (2007), Social Network Analysis: A Practical Method to Improve Knowledge Sharing. Hands-On Knowledge Co-Creation And Sharing; Practical Methods And Techniques, A. S. Kazi, L. Wohlfahrt, P. Wolf, eds., pp. 219-233, Knowledge Board, Stuttgart, 2007. Available at SSRN: <http://ssrn.com/abstract=1467609> or <http://dx.doi.org/10.2139/ssrn.1467609>
- Ramirez, A. (2013), The Influence of Social Networks on Agricultural Technology Adoption. *Procedia - Social and Behavioral Sciences*, 79: 101–116. <http://doi.org/10.1016/j.sbspro.2013.05.059>
- Scott, J. P. (2000), Social Network Analysis. *Sage Publications India Pvt. Ltd*, New Delhi.

- Spielman, D. J., K. Davis, M. Negash and G. Ayele (2010), Rural innovation systems and networks: findings from a study of Ethiopian smallholders. *Agriculture and Human Values*, 28(2), 195–212. <http://doi.org/10.1007/s10460-010-9273-y>
- Subash, S.P. and S. Vishnu (2017), DO NETWORKS MATTER? A retrospective on the potential applications of social network analysis. AESA Blog 72. Available online <https://www.aesanetwork.org/wp-content/uploads/2018/10/AESA-BLOG-72.pdf>
- Tatlonghari, G., T. Paris, V. Pede, I. Siliphouthone and R. Suhaeti (2012), Seed and Information Exchange through Social Networks : The Case of Rice Farmers of Indonesia and Lao PDR, 2(2): 169–176.
- Trienekens, J. H. (2011), Agricultural value chains in developing countries a framework for analysis. *International Food and Agribusiness Management Review*, 14(2): 51–82.
- Valente, T. W., K. Coronges, C. Lakon and E. Costenbader (2008), How correlated are network centrality measures. *Connect (Tor)*. 28(1): 16-26.
- Vishnu, S., I. J. Gupta and S.P. Subash (2019), Social network structures among the livestock farmers vis a vis calcium supplement technology, *Information Processing in Agriculture*, 6(1): 170-182.
- Wood, B. A., H. T. Blair, D. I. Gray, P. D. Kemp, P. R. Kenyon, S. T. Morris and A. M. Sewell (2014), Agricultural science in the wild: A social network analysis of farmer knowledge exchange. *PLoS ONE*, 9(8). <http://doi.org/10.1371/journal.pone.0105203>

Chapter 33

CONSTRUCTION OF COMPOSITE INDEX

Prem Chand

INTRODUCTION

A composite index is a collection of large number of indicators or variables that are aggregated together to represent overall performance of a sector/market. It is a statistical tool, which can provide a useful measure for relative performance of the phenomenon over time or space. Composite Indices aim to combining all the identified individual variables/indicators to reflect overall status or progress or gaps from the desired levels. For example, the well-known Human Development Index (HDI) of United Nations Development Programme (UNDP) combines indicators of health, education and income give the performance of countries. The Composite Index is basically an attempt to find a function of from $R^n \rightarrow R$ corresponding to n-number of component indicators/variables where indicators are functions of variables to measure the extent to which a specified objective or outcome has been achieved. An indicator provides direct measure of a specified aspect of the objective. The composite indices can be used to summaries complex or multi-dimensional issues and facilitate ranking of regions/states, districts according to their performance.

While other techniques of performance evaluation/distance measures have their own utility, the indexing approach recognized as a simple and useful tool in policy analysis and public communication. Number of composite indices are available across the word. Bandura (2006) cites more than 160 composite indicators around the world. The composite index makes the interpretation easier than making common trends across many individual indicators.

Steps in Construction of Composite Index

Construction of Composite Index involves following major steps.

Step 1: Defining the conceptual framework

The first and foremost steps in construction of composite index is defining the conceptual framework. This provides the basis for choice and grouping of variables into a meaningful composite indicator under a fitness-for-purpose principle. The conceptual framework should be defined clearly stating what is being measured, who are the clients of index, whether is it static or dynamic etc. Therefore involvement of experts and stakeholders is very important at this stage. Any new index developed should add value to the composite indicator.

Step 2: Identifying dimensions and relevant Indicators

The second step in construction of composite index is identification relevant indicators and grouping the similar indicators under various dimensions. According to OECD, an indicator is quantitative/qualitative factor/variable that provides a simple and

reliable means to measure achievement/performance to reflect changes connected to an intervention/location, or to assess the performance of a development actor. Identify group of indicators having significant bearing on things to be measured, therefore indicators should be selected carefully.

The indicator selection should be based on the analytical soundness, requirement of data complexity, measurability, geographical coverage, cost effectiveness in collection of data, relevance of the indicators to theme of measurement and their relationship to each other. Wherever, data are not available directly for an indicator, proxy variables should be used. For example, data on number of farmers using tractor might not be available at micro or macro level and hence number of tractors can be used as proxy to that.

Step 3: Construction of Composite Index

The purpose of this step is to integrate the multiple indicators into single index for meaningful interpretation. This step is further divided into following three sub-steps.

A) Normalization of Indicators

It is obvious that the units/scale of measurement of different indicators selected are not same. Some of these might be expressed in monetary terms, some are in hectare, some are in percent, and so on. Comparing these different scaled indicators is like comparing oranges with apple. Therefore, before moving further treatment, they must be attuned, smoothed and to be put on a common basis. The process of normalisation is used to serves this purpose. The transformation of the indicators in order to bring a common scale among indicators is called normalisation. Various methods are available for normalisation serving different purposes and suited to different data properties. Therefore, selection of a suitable normalisation method is very much vital. Choosing normalisation method is steered by concerns such as availability of data in soft and hard form, importance of extreme values, variance in data, etc. For example, if data has extreme values, it is advised to use 'z' score/standardized score. The commonly used normalisation methods are described below.

(a) Min-max normalisation

This method of normalisation is applied when the purpose of the constructing sustainability index is to assess the relative measure. The most common example of this method of normalisation are UNDP's Human Development Index (UNDP, 1995). In India, number of researchers have use min-max method of normalization for assessing relative agricultural sustainability (Chand and Sirohi, 2012, Chand *et al.*, 2011; Singh and Hiremath, 2010; Sen and Hatai, 2007). Following equations are used to normalize the indicator. When indicator has direct relation with measured phenomenon (e.g. sustainability, vulnerability etc), we use equation 1 and in case of negatively related indicator equation 2 is applied.

$$I_{ijk} = \frac{X_{ijk} - \text{Min} X_{ijk}}{\text{Max} X_{ijk} - \text{Min} X_{ijk}} \dots \quad (1)$$

$$I_{ijk} = \frac{\text{Max} X_{ijk} - X_{ijk}}{\text{Max } X_{ijk} - \text{Min } X_{ijk}} \dots \quad (2)$$

Where, X_{ijk} = Value of i^{th} variable representing j^{th} component of sustainability of k^{th} region.

Min-max normalisation is also known as feature scaling where the values of a numeric range of a feature of data, i.e. a property, are reduced to a scale between 0 and 1. However, extreme values/or outliers could distort the transformed indicator. Therefore, indicators with extreme values are treated before normalisation using either truncation method or functional transformation. On the other hand, Min-Max normalisation could widen the range of indicators lying within a small interval, increasing the effect on the composite indicator more than the z-score transformation (OECD, 2008).

(b) Standardization (Z-score)

Standard score or z-score is the number of standard deviations from the mean a data point is. It is a measure of how many standard deviations below or above the population mean a raw score is. Z-score converts indicators to a common scale with a mean of zero and standard deviation of one. In order to use a z-score, one need to know the mean μ and the population standard deviation σ . The formula for z-score is give below.

$$z = (x - \mu) / \sigma$$

This is one of the most commonly used normalisation method especially when the exceptional behavior is to be rewarded or penalized as zero mean of normalized indicator takes care of distortions that may arise while aggregating the indicators with different means.

(c) Normalisation using Benchmarks

This method of normalisation can be used both for relative and absolute assessment of phenomenon. Under this method, the performance of an indicator value is decided by comparing it with specified reference points or benchmarks. These benchmark values may be decided based on scientific reasoning, stakeholders consultations, a target set by the district/region/state/country etc. considered for normalizing the indicator. The indicator value is normalized by dividing it by the benchmark value of the indicator. Using this denominator, the normalisation takes into account the evolution of indicators across time; alternatively, one can use a denominator that changes across time.

(d) Normalisation using thresholds

This method of normalisation is proposed by Chand *et al.* (2015). Under this method instead of using one reference value, two reference values are used. For each indicator an Upper Threshold (UT) and a Lower Threshold (LT) threshold values are identified. For indicators having direct link with measured variable, the upper threshold indicated

the ideal state while the lower threshold indicate the minimum desired level. While for indicator having inverse relation with measured variable the reverse is used. The indicator value is then scored as given below.

Example: If the value of selected indicator has positive effect on phenomenon measure then actual value (A) of indicator \geq upper threshold is scored full whereas actual value of indicator \leq lower threshold scores zero. Value in-between UT and LT was scored as follows:

$$(A - LT) / \left\{ \frac{UT - LT}{(\text{Maximum score})} \right\}$$

If the value of selected indicator has negative effect on phenomenon measure then, actual value of indicator \geq upper threshold gets zero score and actual value of indicator \leq lower threshold gets full score. Value in-between UT and LT is scored as follows:

$$\text{Maximum score} - \left[(A - LT) / \left\{ \frac{UT - LT}{(\text{Maximum score})} \right\} \right]$$

(e) *Normalisation using ranking*

It is simplest method and independent to outliers. In this method, the normalisation is done by simply ranking the observations (districts/regions/states/countries). Examples of this method are Information and Communications Technology Index (Fagerberg, 2001) and the medicare study on healthcare performance across the United States (Jencks *et al.*, 2003). The major drawback of this method is losses information on absolute levels and the impossibility to draw any conclusion about difference in performance.

B) *Assigning weights to different indicators*

The importance of all the indicators of measured variable (like sustainability, vulnerability etc.) may not always be equal. The overall score and ranking of the composite indicators rely on the weighting of the normalized values of the indicators. The weights are assigned to different indicators to reflect their economic significance, reliability, statistical adequacy, etc. The relative weights have a significant effect on the overall composite indicator and relative performance of rankings. Therefore, the assignment of weights should be explicit and transparent. A number of weighting techniques are available in the literature of these some are derived from statistical models while other are from participatory methods (OECD, 2008). The commonly used weighting techniques are listed below.

Weighting based on statistical tools

- i. Principal component analysis (PCA)
- ii. Data envelopment analysis
- iii. Unobserved component model
- iv. Regression analysis

Subjective Weight

- v. Expert/stakeholders opinion
- vi. Budget allocation process
- vii. Analytic Hierarchy Processes (AHP)
- viii. Conjoint analysis

The choice of weight depends mainly on the objective of constructing index, reliability and pattern of data and easiness of weighting methodology. For example, if the indicators are highly correlated, it is advised to use weight based on PCA. While some experts might choose weights based only on statistical methods, others might reward (or punish) components that are deemed more (or less) influential, depending on expert opinion, to better reflect policy priorities or theoretical factors. Most of the composite indicators rely on equal weights (OECD, 2008). The details of some of these are described in rest of the book. One may also follow the OECD's publication "Handbook on constructing composite indicators: Methodology and user guide" (OECD, 2008).

C) Construction of aggregate index

The aggregation methodologies are well documented in the Handbook on Constructing Composite Indices. The fundamental issue in aggregation is the compensability of indicators, which is defined as compensating for any indicator's dimension with a suitable surplus in another indicator's dimension (Talukdar *et al.*, 2017). Based on compensation, there are three broad types of aggregation methodologies namely additive aggregation, geometric aggregation and non-compensatory multi-criteria analysis.

Additive aggregation

The additive or linear aggregation range from summing up the rank based normalized scores of each indicator to aggregating weighted normalised indicators. This method is suitable when all the indicators have the same measurement unit. Information & Communication Technologies Development Index, Summary Innovation Index, Environmental Sustainability Index etc. are the few examples of additive aggregation. The additive methods rewards the indicators based on their relative weights. While this method is simple to calculate, its undesirable feature is full compensability. This feature implies that poor performing indicators can get compensation by their counterpart highly performing indicators.

Geometric aggregation

The geometric aggregation is the product of weighted indicators and therefore unlike the additive aggregation, it is a less compensatory approach. The geometric aggregations reward the high scorer indicators. This method of aggregation is appropriate when the analyst wants some degree of non-compensability between the indicators or dimensions.

Non-compensatory multi-criteria analysis

When the different goals are equally legitimate and vital, a compensatory logic might not be appropriate. In other words, both linear and multiplicative aggregations are unsuitable if the modeller decides that the loss of one dimension cannot be compensated by increasing the performance of other dimension/indicator. This is typically the case when extremely different dimensions are composited. A non-compensatory multi-criterion approach (MCA) could assure non-compensability by finding a compromise between two or more legitimate goals (Munda, 1995).

This method uses a mathematical formulation like Condorcet ranking to rank in a complete pre-order all the indicator values from the best to the worst after a pair-wise comparison of indicator values across all the indicators. A Condorcet method is an election method that elects the candidate that would win a majority of the vote in all of the head-to-head elections against each of the other candidates, whenever there is such a candidate. While this procedure tries to overcome some of the shortcoming of earlier methods, its major drawbacks includes dependence of irrelevant alternatives and its complexity.

ILLUSTRATION

Assessment of Sustainability of Smallholder Dairy Farming in Rajasthan

The sustainability of dairy farming was evaluated through construction of Sustainable Dairy Farming Index (SDFI) for each sample households. The steps for developing this index are as follows:

Step I: Selection of indicators

The indicators for each dimension and a method of assigning the scores to each component/indicators are as given below.

A) Economic Indicators

Dairy farms are business that produce just homogenous products with a limited range of cost factors making the definition of economic indicators for economic sustainability straightforward. At the same time, the use of 'classical' economic business is limited, as the accounting system of farms differs considerably from that of other enterprises. This study considers the economic dimension of sustainability to be based on two important pillars viz., production cost and input productivity viz. labour, capital and feed (Table 1).

B) Social Indicators

Basic social requirements of sustainable dairy farming and farms are that the farm family as a group earn sufficient income, that there should be no gender discrimination or gender bias and that workers work under the condition that do not put their health at risk. Therefore, this dimension of sustainability was reflected in three variables viz., family labour income, women empowerment and drudgery of work. The list of social indicators is give in table 1.

Table 1: List of indicators and their threshold values

Indicators	Lower Threshold	Upper Threshold
Production cost	Least cost of the milk production in the sample households	Average selling prices of milk
Productivity of inputs		
a) Labour productivity	Minimum wage rate of Rs. 73 per day	Average productivity of top ten sample households
b) Capital productivity	Rate of interest on fixed deposit (9%)	Ideal rate of return (20%) on dairy
c) Feed productivity	0.45 lit./kg of dry matter	1 lit./kg of dry matter
Family labour income as a percent of monthly per capita consumption expenditure	0 percent	100 percent
Women empowerment index	0	100
Drudgery of work		
a) Sharing of burden among number of person in each activity	0	100
b) Carrying of weight/load	5kg/km	25 kg/km
c) Days off work affordable by main worker	One day off per week (52days)	One day off per month (12 days)
Use of dung as fuel to manure (%)	0 percent	100 percent
Enteric methane emission	Minimum methane emission in study area = 26.26 gm/lit of milk produced	Average methane emission in the study area = 42 gm/lit of milk
Breeding practices (index)	0	100

C) Ecological Dimension

The interaction between dairy farming and the environment are manifold. The goal of a sustainability index is to include the most significant environmental impacts of dairy farming and the ones with the greatest variations between dairy farms. Therefore, environmental factors were reflected by three variables viz., use of dung, enteric methane emission and breeding practices.

Step-II: Assigning of scores

After quantifying the indicators in each of the three dimensions, viz., economic, social and ecological dimension, scores were assigned to the indicators using threshold method of normalization.

Step-III: Construction of index

Using the scores assigned to the quantitative values of each indicator, index was constructed for each dimension (economic, social and environmental) and an overall index incorporating these three dimensions. The indicator values were normalized using threshold method of normalization. Two types of weights were assigned to indicators namely equal weights and weights based on budget method.

Based on the indicators stated in the methodology section, the average score of economic sustainability computed for different herd size categories and the percent distribution of households based on the range of economic sustainability scores is given in Table 2. The table shows that the overall score of economic sustainability was on the lower side at 27.06. The inter-household variations were very wide (CV = 75.73%) as economic sustainability ranged from as low as 1.49 to as high as 79.45 on a 100-point scale.

Table 2: Distribution of households according to level of total economic sustainability score.

Level of Score	Weighting pattern							
	Equal weights (100-point scale)				Based on experts opinion (138-point scale)			
	Small (N=56)	Medium (N=39)	Large (N=25)	Overall	Small (N=56)	Medium (N=39)	Large (N=25)	Overall
Low	42 (75.00)	24 (61.54)	12 (48.00)	78 (65.00)	42 (75.00)	22 (56.41)	13 (52.00)	77 (64.13)
Moderate	11 (19.64)	15 (38.46)	9 (36.00)	35 (29.17)	11 (19.64)	17 (43.59)	9 (36.00)	37 (30.83)
High	3 (5.36)	-	4 (16.00)	7 (5.83)	3 (5.36)	-	3 (12.00)	6 (5.00)
Overall (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Average Score	21.69	28.61	36.67	27.06	30.12	39.69	50.62	37.50
Range	1.49 to 79.45	3.06 to 62.09	4.01 to 72.61	1.49 to 79.45	2.37 to 114.89	4.85 to 86.16	6.36 to 100.97	2.37 to 114.89
C.V.	93.61	65.74	55.24	75.73	94.87	66.31	55.14	76.32

Figures in parentheses are the percentages of total no of households in respective category

Low = Range of scores bottom one-third; Moderate = Range of scores middle one-third; High = Range of scores top one-third.

Similarly, social and ecological security index were developed and used for construction of composite index.

Overall Sustainable Dairy Farming Index (SDFI)

Finally, all the aspects of sustainability were weighted on overall Sustainable Dairy Farming Index (SDFI) using multiplicative aggregation methodology as mentioned below.

$$\text{SDFI} = \log (a*b*c)$$

Where,

- SDFI = Sustainable Dairy Farming Index
a = Economic dimension of sustainability,
b = Social dimension of sustainability and
c = Ecological dimension of sustainability.

In our example, the composite sustainability index was estimated by taking into account all the three dimensions discussed above. The full scores in each aspect of sustainability would results in a maximum of Sustainable Dairy Farming Index (SDFI) to 6, while minimum would be SDFI = 0. The average value of SDFI as well as distribution of households based on the value of SDFI across herd size category is presented in Table 3.

The overall average value of SDFI was 4.48, ranging from 2.86 to 5.46. The inter-household variations were low (CV = 12.35%) and lowest for large category of households (CV = 8.97%), followed by medium and small categories with coefficient of variations 8.97 and 10.89 percent, respectively.

Table 3: Distribution of households based on level of SDFI across herd size categories

Level of Score	Weighting pattern							
	Equal weights (100-point scale)				Based on experts opinion			
	Small (N=56)	Medium (N=39)	Large (N=25)	Overall	Small (N=56)	Medium (N=39)	Large (N=25)	Overall
Low	2 (3.57)	-	-	2 (1.67)	-	-	-	-
Moderate	49 (87.50)	29 (74.36)	15 (60.00)	93 (77.50)	50 (89.29)	28 (71.79)	15 (60.00)	93 (77.50)
High	5 (8.93)	10 (25.64)	10 (40.00)	25 (20.83)	6 (10.71)	11 (28.71)	10 (40.00)	27 (22.50)
Overall (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Average Score	4.26	4.59	4.81	4.48	4.37	4.67	4.89	4.58
Range	2.86 to 5.41	3.51 to 5.46	3.87 to 5.43	2.86 to 5.46	3.09 to 5.31	3.70 to 5.40	4.07 to 5.41	3.09 to 5.41
C.V.	14.22	10.89	8.97	12.93	12.35	9.38	7.64	11.27

Figures in parentheses are the percentages of total no of households in respective category
Low = (<3); Moderate (3-5); High (≥5)

REFERENCES

Bandura, R. (2006), A Survey of Composite Indices Measuring Country Performance: 2006 Update, United Nations Development Programme – Office of Development Studies, available at http://www.thenewpublicfinance.org/background/Measuring%20country%20performance_nov2006%20update.pdf

- Chand, P. and S. Sirohi (2012), District level sustainable livestock production index: Tool for livestock development planning in Rajasthan. *Indian Journal of Agricultural Economics*, 63 (2): 199-212.
- Chand, P., S. Sirohi and S. K. Sirohi (2011), Using Sustainable Livestock Production Index for Development of Livestock Sector: case of arid region in India. *Journal of Applied Animal Research*, 39(3):234-238.
- Chand, P., S. Sirohi and S.K. Sirohi (2015), Development and application of an integrated sustainability index for small-holder dairy farms in Rajasthan, India. *Ecological Indicators*, 56, 23-30.
- Fagerberg, J. (2001), Europe at the crossroads: The challenge from innovation-based growth. In: in Lundvall B. and Archibugi D. (eds.) *Globalising Learning Economy*, Oxford Press.
- Jencks, S. F., E. D. Huff and T. Cuerdon (2003), Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001, *Journal of the American Medical Association*, 289(3): 305-12.
- Munda, G. (1995), *Multicriteria Evaluation in a Fuzzy Environment*, Physica-Verlag, Contributions to Economics Series, Heidelberg.
- OECD (2008), *Handbook on constructing composite indicators: Methodology and user guide*. Available at <https://www.oecd.org/sdd/42495745.pdf>.
- Sen, C. and L. D. Hatai (2007), Agricultural Sustainability in Orissa: District-wise Analysis. Paper presented at the First Mediterranean Conference of Agro-Food Social Scientists. 103rd EAAE Seminar 'Adding Value to the Agro-Food Supply Chain in the Future Euromediterranean Space'. Barcelona, Spain, April 23-25.
- Singh, P. K. and B. N. Hiremath (2010), Sustainable Livelihood Security Index in a Developing Country: a Tool for Development Planning. *Ecological Indicators*, 10(2): 442-451.
- Talukder, B., I. D. Keith, W. Hipel and G. W. van Loon (2017), Developing Composite Indicators for Agricultural Sustainability Assessment: Effect of Normalization and Aggregation Techniques *Resources* 6 (4): 66.
- UNDP (1995), *Human Development Report 1995*, Oxford University Press, New York.

Chapter 34

BASIC SCALING TECHNIQUES IN SOCIAL SCIENCES

Sudipta Paul

INTRODUCTION

In social sciences much of the data consist of qualitative variables, like attitude, practice, status, esteem etc. and in order to develop a scientific understanding about these variables, they must be so arranged that they represent a quantitative series. Scaling techniques facilitate in ordering a series of such items along some sort of continuum. In other words, they act as methods of turning a series of qualitative facts or attributes into a quantitative series or variables. Therefore, scaling quintessentially hypothesizes the existence of a continuum of some kind, the nature of which is contingent upon the character of the items selected to construct the scale. Logically unrelated items, therefore, cannot be included in the same scale. Every scale consists of items which are only a sample of the universe of items. Therefore, the foremost important requirement in a scaling procedure is to develop a deep understanding and gather thorough knowledge in the subject under investigation. One must systematically exploit his own observations and those of others through a careful study of the literature. Achieving perfect accuracy in a scale is quite a formidable task. As the nature of the population which is to be scaled varies, a scale must be treated cautiously and must always be viewed as tentative when applied in a new and dissimilar population. The present chapter has been restricted with scaling of attitude.

Construction of an Attitude Scale

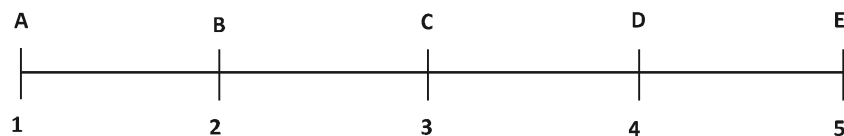
Attitude is a widely researched area of human behaviour and plenty of literature is available in the journals of social and behavioral sciences encompassing attitude and its measurement. Thurstone (1946) defined attitude as the degree of positive or negative affect associated with some psychological object. A psychological object according to him may be a symbol, phrase, slogan, person, institution, ideal and idea toward which people differ in terms of positive or negative affect. In the literature of psychology the terms affect and feeling have been used interchangeably, both indicating the same meaning. An individual who has negative affect or negative feeling toward some psychological object may be disliking of the object. In scaling procedures, such an attitude is termed as 'unfavourable attitude'. We may directly ask individuals how they feel about some psychological object and it seems quite logical to go for direct questioning to understand attitude of individuals toward a psychological object. We may be thereafter able to categorize individuals based upon their attitude. We may say individuals with favourable attitude, with unfavourable attitude and with neutral attitude.

Types of attitude scales

There are basically three major types of attitude scales- differential scales, summated scales and cumulative scales. Whereas, the first two are prominent scaling techniques, the third one is not as such a scaling procedure, rather can be described as a procedure to assess a set of statements of an already existing scale in order to confirm that the statements fulfill certain requirements of unidimensionality, as suggested by Guttman. A differential scale consists of items whose position on the scale is determined by some kind of ranking or rating process completed by judges. Thurstone's technique of equal appearing interval is a good example of differential scale. In case of summated rating, individuals indicate their agreement or disagreement with a set of items on a scale. In the process of summated rating, the total score obtained by an individual is determined by simple addition of scores obtained through item wise responses. Likert's method of summated rating falls under this category of scaling. As mentioned earlier also, the Cumulative or Guttman scale, consists of a relatively small set of homogeneous items that are or supposed to be unidimensional. A unidimensional scale measures only one variable at a time. Scalogram analysis is undertaken to determine whether or not the constituent items of a scale meet the requirements of unidimensionality as suggested by Guttman. The two most popularly followed scaling techniques, the method of equal appearing intervals and the method of summated rating have been discussed in details here.

THE METHOD OF EQUAL APPEARING INTERVALS

Interval or equal-appearing interval scales possess the characteristics of both nominal and ordinal scales, especially the rank-order characteristic. An interval scale can be viewed as follows:



The interval between A and C as depicted above is $3-1=2$ and that between C and D is $4-3=1$. If we now add these two intervals, we get $(3-1) + (4-3) = 2+1 = 3$. Note that the interval between A and D is $4-1=3$. Expressed in an equation: $(D-A) = (C-A) + (D-C)$. So, we can say that the intervals in an equal appearing interval scale can be added and subtracted. In addition, numerically equal distances on interval scales represent equal distances in the property being measured. If these intervals were some sort of scores of five students, measured through an interval scale, we could have safely concluded that the differences in scores between A and C and between B and D are equal, although we could not say that the score of D was twice as great as that of B. In order to do so, we require a measurement at a higher level called 'ratio level.' In an interval scale, it is not quantities or amounts that are added and subtracted, it is intervals or distances.

Steps in construction of an equal appearing interval scale

The following steps are involved in construction of an equal appearing interval scale.

Step 1: Collection of statements

The researcher needs to gather a large number of statements concerning the area under investigation. The method of statement collection involves thorough review of existing literature in the related areas of study, discussion with experts in the related fields, field observation and intuition.

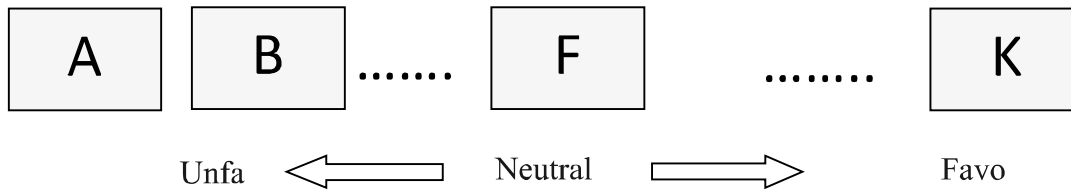
Step 2: Scrutiny and editing of statements

Wang (1932), Thurstone and Chave (1929) Likert (1932), Bird (1940), Edwards (1941) and Kilpatrick (1948) suggested various informal criteria for editing statements for attitude scale construction. Edwards (1969) summarized their suggestions as follows:

1. Avoid statements that refer to the past rather than to the present.
2. Avoid statements that are factual or capable of being interpreted as factual.
3. Avoid statements that may be interpreted in more than one way.
4. Avoid statements that are irrelevant to the psychological object under consideration.
5. Avoid statements that are likely to be endorsed by almost everyone or by almost no one.
6. Select statements that are believed to cover the entire range of the affective scale of interest.
7. Keep the language of the statements simple, clear, and direct.
8. Statements should be short, rarely exceeding 20 words.
9. Each statement should contain only one complete thought.
10. Statements containing universals such as all, always, none, and never often introduce ambiguity and should be avoided.
11. Words such as only, just, merely, and others of a similar nature should be used with care and moderation in writing statements.
12. Whenever possible, statements should be in the form of simple sentences rather than in the form of compound or complex sentences.
13. Avoid the use of words that may not be understood by those who are to be given the completed scale.
14. Avoid the use of double negatives.

Step 3: Sorting of statements

Thurstone and Chave (1929) suggested to print each attitude statement on a separate card for spreading them open to the judges who are to sort the cards into a predefined number of categories. Along with the cards containing attitude statements, therefore the judges are supplied with a set of 11 cards on which the letters A to K appear.



The A card which is placed at the extreme left is described as representing the card on which the statements that seem to express the most unfavourable feelings about the psychological object are to be placed. The K card is kept to the extreme right. Statements perceived to be expressing the most favourable feelings about the psychological object are asked to be placed on the K card. The middle or F card is described as the 'neutral' card. Statements expressing neither favourable nor unfavourable feelings regarding the psychological object are asked to be placed on the F card. As we move from cards, G to K, we should find statements with varying degrees of increasing favourableness. Similarly, as we move from cards, D to A, we should find statements with varying degrees of increasing unfavourableness. The F card or neutral interval is therefore, a zero point, an arbitrary zero point. So, basically in this particular step, each judge is asked to pile up cards containing attitude statements on the A-K continuum based upon the degree of favourableness or unfavourableness expressed by each statement as perceived by them. It is found that it takes around 45 minutes to judge 130 statements by a single judge.

Step 4: Computation of scale value and Q value

A table like Table 1 is drawn in order to compute scale value and Q value.

Table 1: Summary table for obtaining judgments

Statement	Measure	Sorting category										
		A	B	C	D	E	F	G	H	I	J	K
1	F											
	P											
	cp											
2	F											
	P											
	cp											
3	F											
	P											
	cp											
n	F											
	P											
	cp											

f – frequency, *p* – proportion, *cp* – cumulative proportion

As depicted in Table 1, three rows are used for each statement, the first row being representing frequency with which the statement was placed in each of the eleven categories. When we divide each frequency (f) by the total number of judges (n), what we get is proportions (f/n), shown in the second row. Proportion in a given category plus the sum of all of the proportions below the category, i.e., the cumulative proportions are given in the third row. The scale value is computed following the formula given below:

$$S = l + \left(\frac{0.50 - \sum pb}{pw} \right) i$$

where, S, median or scale value of the statement; l, lower limit of the interval in which the median falls; $\sum pb$, sum of all the proportions below the interval in which the median falls; pw, the proportion within the interval in which the median falls; I, width of the interval and is assumed to be equal to 1.0.

Thurstone and Chave (1929) used the interquartile range (Q) which is a measure of variation of the judgment distribution for a given statement. The interquartile range contains the middle 50 percent of the judgments. To determine the value of Q, two other point measures, the 75th centile and 25th centile are worked out. The 25th centile is obtained from the following formula:

$$C25 = l + \left(\frac{0.25 - \sum pb}{pw} \right) i$$

where, C25, 25th centile; l, lower limit of the interval in which the 25th centile falls; $\sum pb$, sum of all the proportions below the interval in which the 25th centile falls; pw, proportion within the interval in which the 25th centile falls; i is the width of the interval and is assumed to be equal to 1.0.

The same way, 75th centile can be obtained through the following formula:

$$C75 = l + \left(\frac{0.75 - \sum pb}{pw} \right) i$$

Where, C75, 75th centile; l, lower limit of the interval in which the 75th centile falls; $\sum pb$, sum of all the proportions below the interval in which the 75th centile falls; pw, proportion within the interval in which the 75th centile falls; I, width of the interval and is assumed to be equal to 1.0.

The interquartile range, Q is the difference between C75 and C25 and can be represented in the form of a equation: $Q = C75 - C25$. As, Q is a measure of spread of the middle 50 percent of judgments, the smaller the Q, higher the agreement between the judges regarding favourableness or unfavourableness of a statement. A larger Q value indicates that there lies some ambiguity with the statement which can be interpreted in more than one way.

Step 5: Retention of statements in scale

Approximately 20-22 statements are desirable in an equal appearing interval scale. The scale values of the statements should be such that they are relative equally spaced on the psychological continuum. The Q values of the statements should be relatively small. Thus, both S and Q are used as important criteria for the selection of attitude statements of an equal appearing interval scale. Sometimes, we need to make a choice among several statements having approximately the same scale values (S). Preference should obviously be attached to the statements with the lowest Q values. Any of the standard methods of reliability testing can be followed in obtaining the reliability coefficient. The reliability coefficients typically reported in literature for equal appearing interval scales are above 0.85.

Thurstone and Chave (1929) suggested another criterion, criterion of irrelevance in addition to Q, as a basis for rejecting statements in scales constructed by the method of equal-appearing intervals. The criterion of irrelevance is based upon the agreement or disagreement of respondents with statements having known scale values. The judgments offered by a group of judges regarding the degree of favorableness or unfavorableness of statements are not considered in finding out the criterion of irrelevance. Let us assume that a group of n number of respondents are equal in their attitude scores. All obtained 7.0. 'i' is any statement with scale value S_i , n_i is the number of respondents in the group of n who endorsed the statement i . As long as S_i takes a scale value closer to 7.0, we can expect that n_i will increase and approach maximum. Similarly, as we move farther from an S_i of 7.0, we should expect n_i to decrease. Here, n_i/n is the probability with which a given statement will be endorsed by the particular group comprising of n number of respondents. The probability value should be maximum for statements with the same scale values. This criterion however, has not been extensively used.

Step 6: Administration and scoring

The selected statements can be arranged in random order and presented to respondents with instructions to indicate those statements that they are willing to accept or agree with and those statements that they reject or disagree with. An attitude score can be obtained from the scale values of the statements agreed upon. An indication of the location of the respondent on the psychological continuum on which the statements have been scaled, can thus be obtained. So, basically the attitude score is based upon the arithmetic mean or median of the scale values of the statements agreed with.

Interpretation of attitude scores obtained through the method of equal appearing intervals

In equal appearing interval scales, the attitude score obtained by an individual has an absolute interpretation in terms of the psychological continuum of scale values of the statements, because the attitude score is taken as the median of the scale values of the statements with which the subject agrees. Each attitude score is thus itself a scale value on the psychological continuum on which the statements have been scaled. If we recall correctly, in scaling the statements, one end of the continuum was defined as unfavourable and the other end as favourable, with the middle category being defined as neutral. If an attitude score falls in the middle portion of the continuum, it can be described as neutral. Similarly, if it falls toward the favourable end of the continuum, it can be described as favourable, and if it falls toward the unfavourable end, the attitude can be described as unfavourable. This interpretation of an attitude score on an equal appearing interval scale can, thus be made independently of the distribution scores for a particular group of respondents.

THE METHOD OF SUMMATED RATING

A summated rating scale consists of a set of attitude items all of which are considered of approximately equal attitude value. The respondents need to respond to each of the statements with degrees of agreement or disagreement. The scores of the items of such a scale can be summed up and averaged to finally reach an individual's attitude score. The purpose of summated rating scale is to place an individual somewhere on an agreement continuum of the attitude toward a psychological object. Summated rating scales allow for the intensity of attitude expression. Respondents can agree or disagree, strongly agree or strongly disagree, even can remain neutral. Bird (1940) called this method 'the method of summated ratings'.

Steps in construction of a summated rating scale

The following steps are involved in construction of summated rating scale.

Step 1: Item collection

The researcher needs to gather a large number of statements concerning the area under investigation. The method of item collection involves thorough review of existing literature, discussion with experts in the related fields, field observation and intuition.

Step 2: Scrutiny and editing of items

Wang (1932), Thurstone and Chave (1929) Likert (1932), Bird (1940), Edwards (1941) and Kilpatrick (1948) suggested various informal criteria for editing statements to be used in the construction of attitude scales. Edwards (1969) summarized their suggestions which are dealt in step 2 of equal appearing interval methods.

Step 3: Selection of items

(a) Assignment of weightage and primary administration

A Likert type scale consists of two types of statements- favourable and unfavourable, with approximately the same number of statements of each type. The statements are given to a group of non-sample respondents who are asked to respond to each one of the statements in terms of their own agreement or disagreement with the statement. The agreement or disagreement is asked to be expressed in any of the five categories administered- strongly disagree, disagree, undecided, agree and strongly agree. For favourable statements, the strongly agree category is assigned with a weight of 4, agree with 3, undecided with 2, disagree with 1 and strongly disagree with 0. A reverse scoring pattern is adopted in unfavourable statements.

(b) Deciding upon the criterion groups

The total scores obtained by the non-sample respondents are calculated after summing up item wise scores. Twenty five percent of respondents with highest total score and twenty five percent respondents with lowest total score are then isolated. These two groups serve as the 'criterion groups' in evaluation of individual statements.

(c) Calculation of t values

The value of t is a measure of the extent to which a given statement is able to differentiate between the criterion groups, therefore, the t value of individual items is worked out using the following formula:

$$t = \frac{\bar{X}_H - \bar{X}_L}{\sqrt{\frac{S_H^2}{n_H} + \frac{S_L^2}{n_L}}}$$

where, \bar{X}_H , mean score on a given statement for the high group; \bar{X}_L , mean score on the same statement for the low group; S_H^2 , variance of the distribution of responses of the high group to the statement; S_L^2 , variance of the distribution of responses of the low group to the statement; n_H , number of respondents in the high group and n_L , number of respondents in the low group.

If, $n_H = n_L = n$, as will be the case if we select the same percentage of the total number of respondents in the high and low groups, the formula can be produced as:

$$t = \frac{\bar{X}_H - \bar{X}_L}{\sqrt{\frac{\sum(X_H - \bar{X}_H)^2 + \sum(X_L - \bar{X}_L)^2}{n(n-1)}}$$

where, $\sum(X_H - \bar{X}_H)^2 = \sum X_H^2 - \frac{(\sum X_H)^2}{n}$ and $\sum(X_L - \bar{X}_L)^2 = \sum X_L^2 - \frac{(\sum X_L)^2}{n}$

If we have 25 or more non-sample respondents both in the high as well as in the low group, we may regard as a crude and approximate rule of thumb any t value ≥ 1.75 as indicating that the average response between the high and low groups to a statement differs significantly.

Step 4: Reaching the final scale

After obtaining t value of all the statements, they are arranged in descending order. Top 20-25 statements, all with t value ≥ 1.75 are retained in the final scale. Approximately half of the selected statements should be favourable and the remaining half should consist of unfavourable statements. The reliability of the scale may be established after following any of the standard methods. The reliability coefficients typically reported for scales constructed by the method of summated ratings are above 0.85.

Interpretation of attitude scores obtained through the method of summated rating

A summated rating score corresponding to the zero or neutral point on a favourable-unfavourable continuum is not known as it is assumed to be known in case of the equal appearing interval scales. Therefore, interpretation of an attitude score on a summated rating scale cannot be made independently of the distribution of scores of some defined group. There is no evidence to indicate that the neutral point on a summated rating scale corresponds to the midpoint of the possible range of scores. So, we cannot say that a score of 50 obtained through the process of summated rating, comprising 25 statements each with a 0-4 continuum, is the 'zero point'.

Scaling techniques help in turning a series of qualitative facts or attributes into a quantitative series or variables. Scales are mostly used to measure attitude of subjects towards a psychological object. There are basically three major types of attitude scales - differential scales, summated scales and cumulative scales. Equal appearing interval scale of Thurstone is a good example of differential scales. Apart from scale values, q values (interquartile range) are important to consider for retention of statements in an equal appearing interval scale. The Likert scale which is an example of a summated scale, uses t values calculated using mean and variance of scores in criterion groups, for discarding and retaining statements in the final scale. Measurement and interpretation of attitude scores vary for the two types of scales. The scales are administered after checking for their internal consistency (reliability) and validity.

REFERENCES

- Bird, C. (1940), Social psychology. New York: Appleton-Centruy-Crofts.
Edwards, A. L. (1969), Techniques of attitude scale construction, Vakils, Feffer and Simons Pvt. Ltd.

- Kilpatrick, F. P. (1941), A Technique for the construction of attitude scales. *Journal of Applied Psychology*, 36: 34-50.
- Likert, R. A. (1932), Technique for the measurement of attitudes. *Archives of psychology*.
- Thurstone, L. L. (1946), Comment. *American Journal Social*, 52; 39-50
- Thurstone, L. L. and E. J. Chave (1929), The measurement of attitude. Chicago. Univ. Chicago Press.
- Wang, K. A. (1932), Suggested criteria for writing attitude statements. *Journal of Social Psychology*, 3: 367-373.

Chapter 35

ANALYTICAL HIERARCHY PROCESS: A MULTI-CRITERIA DECISION MAKING TECHNIQUE

Anirban Mukherjee, Mrinmoy Ray and Kumari Shubha

INTRODUCTION

Analytic hierarchy process (AHP) is one of multi criteria decision making method that was originally developed by Prof. Thomas L. Saaty (1980). In short, it is a method to derive ratio scales from paired comparisons. The AHP is a methodology for structuring, measurement and synthesis. It has been applied to a wide range of problem situations: selecting among competing alternatives in a multi-objective environment, the allocation of scarce resources, and forecasting. Although it has wide applicability, the axiomatic foundation of the AHP carefully delimits the scope of the problem environment (Saaty, 1986). It is based on the well-defined mathematical structure of consistent matrices and their associated right-eigenvector's ability to generate true or approximate weights, Mirkin (1979), Saaty (1980, 1994a). The input can be obtained from actual measurement such as price, weight etc., or from subjective opinion such as satisfaction feelings and preference. AHP allows some small inconsistency in judgment because human is not always consistent. The ratio scales are derived from the principal Eigen vectors and the consistency index is derived from the principal Eigen value.

The prime use of the AHP is the resolution of choice problems in a multi-criteria environment. In that mode, its methodology includes comparisons of objectives and alternatives in a natural, pairwise manner. The AHP converts individual preferences into ratio-scale weights that are combined into linear additive weights for the associated alternatives. These resultant weights are used to rank the alternatives and, thus, assist the decision maker (DM) in making a choice or forecasting an outcome.

The AHP employs three commonly agreed to decision making steps:

- (1) Given $i = 1, \dots, m$ objectives, determine their respective weights w_i ,
- (2) For each objective i , compare the $j = 1, \dots, n$ alternatives and determine their weights w_{ij} with respect to objective i , and
- (3) Determine the final (global) alternative weights (priorities) W_j with respect to all the objectives by $W_j = w_{1j}w_1 + w_{2j}w_2 + \dots + w_{mj}w_m$.

The alternatives are then ordered by the W_j , with the most preferred alternative having the largest W_j . The various decision methodologies are differentiated by the way they determine the objective and alternative weights, as prescribed by each one's axiomatic or rule-based structure. The general validity of the AHP, and the confidence placed

in its ability to resolve multi-objective decision situations, is based on the many thousands of diverse applications in which the AHP results were accepted and used by the cognizant decision makers, Saaty (1994b).

Development of the AHP

In the late 1960's, Thomas L. Saaty, an OR pioneer, was directing research projects for the Arms Control and Disarmament Agency at the U.S. Department of State. Saaty's research agenda, and very generous budget, enabled him to recruit some of the world's leading game and utility theorists and economists. In spite of the talents of the people recruited (three members of the team, Gerard Debreu, John Harsanyi, and Reinhard Selten, have since won the Nobel Prize), Saaty was disappointed in the results of the team's efforts. Years later, while teaching at the Wharton School, Saaty was still troubled by the apparent lack of a practical systematic approach for priority setting and decision making. He was thus motivated to develop a simple way to help decision makers to make complex decisions. The result was the AHP. There is ample evidence that the power and simplicity of the AHP has led to its widespread usage throughout the world. In addition to the popular Expert Choice software, there have been several other commercial implementations of the AHP. The AHP is now included in most OR/MS texts and is taught in numerous universities. It is used extensively in organizations that have carefully investigated the AHP's theoretical underpinnings, such as the Central Intelligence Agency.

Calculation of AHP

The AHP algorithm mainly consists of two parts: 1) construction of the pair-wise comparison and 2) prioritization of decision alternatives.

The steps of AHP methodology followed in this research are:

Step 1: Structuring of the decision problem into a hierarchical model

In this step decomposition of the decision problem into elements according to their common characteristics and the formation of hierarchical model are done.

Step 2: Constructing the pair-wise comparison matrix

Two types of pair wise comparisons are made in the AHP. The first one is between the factor pairs within the same hierarchical level which involves analyst inputs of relative importance ratings based on the pair wise comparative ratings in a scale of 1 to 9. The factors weights are computed and used in the final hierarchical merit aggregation process.

The matrices of pair-wise comparisons (Eq. 1) are obtained.

In this matrix, the element $a_{ij} = \frac{1}{a_{ji}}$ and thus, when $i=j$ $a_{ij} = 1$. Every element in an

upper level is used to compare with respect to the elements in the level below. This work was done by pair-wise comparison two by two and through dedicating numeral scores which shows priority and majority between two decision elements.

$$A = (a_{ij}) = \begin{pmatrix} 1 & \frac{w_1}{w_2} & \dots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & 1 & \dots & \frac{w_2}{w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \dots & 1 \end{pmatrix} \dots \quad (1)$$

Step 3: Calculating the consistency

The traditional Eigen vector method, is the weight vector that was our goal. It helps in measuring the consistency of the referees preference arranged in comparison matrix. The consistency index (CI) measures the degree of logical consistency among pair-wise comparisons. Saaty (1996) defined the consistency index (CI) as follows:

$$CI = \frac{\lambda_{\max} - 1}{n - 1} \dots \quad (2)$$

Where, n is the number of existing items in the judgment matrix problem.

Consistency ratio (CR) indicates that the amount of allowed inconsistency i.e. (0.1 or 10%). It is calculated using the following formula:

$$CR = \frac{CI}{RI} \dots \quad (3)$$

The value of the random index (RI) for matrices (Saaty, 2008) of order (n) was used in CR calculation.

Table 1: Pair-wise comparison scale used in the study

Definition	Intensity of importance	Explanation
Equal importance	1	Equal contribution of two parameters
Weak importance	3	Experience and judgement slightly favour one parameter over another
Strong importance	5	Experience and judgement strongly favour one parameters over another
Very strong or demonstrated importance	7	An parameter is favoured very strongly over another; its dominance demonstrated in practice
Extreme or absolute importance	9	The evidence favouring one parameter over another is of the highest possible order of affirmation

Definition	Intensity of importance	Explanation
Intermediate value	2,4,6,8	Interpolation of compromised judgement numerically because there is no good word to describe it

Source: Saaty (1996)

Table 2: Random index (RI)

N	1	2	3	4	5	6	7	8	9	10	11	12
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48

Source: Saaty (1996)

Fig 1 shows an example of two levels AHP. The structure of hierarchy in this example can be drawn as following:

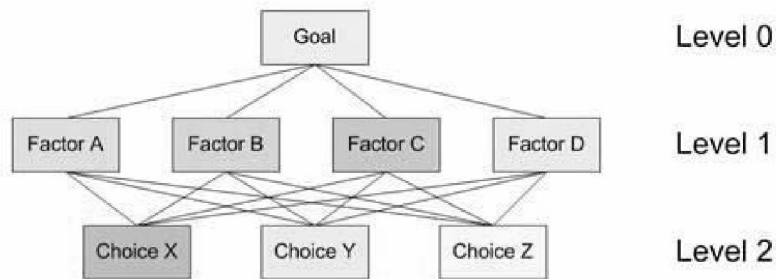


Fig 1: Two level structure of hierarchy of AHP

Level 0 is the goal of the analysis. Level 1 is multi criteria that consist of several factors. You can also add several other levels of sub criteria and sub-sub criteria. The last level (level 2 in figure above) is the alternative choices. You can see again Table 1 for several examples of goals, factors and alternative choices. The lines between levels indicate relationship between factors, choices and goal. In level 1 you will have one comparison matrix corresponding to pair-wise comparisons between 4 factors with respect to the goal. Thus, the comparison matrix of level 1 has size of 4 by 4. Because each choice is connected to each factor, and you have 3 choices and 4 factors, then in general you will have 4 comparison matrices at level 2. Each of these matrices has size 3 by 3. However, in this particular example, you will see that some weight of level 2 matrices are too small to contribute to overall decision, thus we can ignore them.

Based on questionnaire survey or your own paired comparison, you can make several comparison matrices. The diagonal is always 1 and the lower triangular matrix is filled using formula

$$a_{ji} = \frac{1}{a_{ij}}$$

Table 3: Paired comparison matrix level 1 with respect to the goal

Criteria	A	B	C	D	Priority Vector (%)
A	1.00	3.00	7.00	9.00	57.39
B	0.33	1.00	5.00	7.00	29.13
C	0.14	0.20	1.00	3.00	9.03
D	0.11	0.14	0.33	1.00	4.45
Sum	1.59	4.34	13.33	20.00	100.00

$$\lambda_{NA} = 4.2692, CI = 0.0897, CR = 9.97\% < 10\% \text{ (acceptable)}$$

Like wise it can be calculated for other levels. And finally combining all these you shall get a composite weight.

The screenshot shows an Excel spreadsheet titled 'AHP Fodder [Compatibility Mode] - Microsoft Excel'. The worksheet contains the following data and formulas:

	A	B	C	D	E	F	G	H
1								
2								
3								
4								
5		A	B	C	D		nth root of prod eigen vector	
6	A	1	=1/B7	=1/B8	=1/B9	=B6*C6*D6*E6	=F6^(1/4)	=G6/G10
7	B	2	1	0.5	1	=B7*C7*D7*E7	=F7^(1/4)	=G7/G10
8	C	3	2	1	0.5	=B8*C8*D8*E8	=F8^(1/4)	=G8/G10
9	D	2	1	2	1	=B9*C9*D9*E9	=F9^(1/4)	=G9/G10
10							=SUM(G6:G9)	
11								
12								
13								
14								
15								
16								
17								
18								
19								

Plate 1: The Formula of calculating AHP in Excel

The plate 1 indicates the formula to calculate the AHP. Here 4 variables are considered like A, B, C, D. To calculate the values follow the formula. To calculate the Eigen vector follow the column F, G carefully.

For more details you can refer some recent papers on AHP like Mukherjee *et al.*, 2018; Wasielewska and Ganzha, 2012 etc. or search in Google.

Why the AHP is so widely applicable

Any situation that requires structuring, measurement, and/or synthesis is a good candidate for application of the AHP. Broad areas in which the AHP has been successfully employed include: selection of one alternative from many; resource allocation; forecasting; total quality management; business process re-engineering; quality function deployment, and the balanced scorecard. The AHP, however, is rarely

used in isolation. Rather, it is used along with, or in support of, other methodologies. When deciding how many servers to employ in a queueing situation, the AHP is used in conjunction with queueing theory to measure and synthesize preference with respect to such objectives as waiting times, costs, and human frustration. When using a decision tree to analyze alternative choices -- chance situations -- the AHP is used to derive probabilities for the choice nodes of the decision tree, as well as to derive priorities for alternatives at the extremities of the decision tree.

Strengths

The advantages of AHP over other multi criteria methods are its flexibility, intuitive appeal to the decision makers and its ability to check inconsistencies. Generally, users find the pairwise comparison form of data input straightforward and convenient.

- Additionally, the AHP method has the distinct advantage that it decomposes a decision problem into its constituent parts and builds hierarchies of criteria. Here, the importance of each element (criterion) becomes clear.
- AHP helps to capture both subjective and objective evaluation measures. While providing a useful mechanism for checking the consistency of the evaluation measures and alternatives, AHP reduces bias in decision making.
- The AHP method supports group decision-making through consensus by calculating the geometric mean of the individual pairwise comparisons.
- AHP is uniquely positioned to help model situations of uncertainty and risk since it is capable of deriving scales where measures ordinarily do not exist.

Weaknesses

- Despite the popularity of the AHP, many authors have expressed concern over certain issues in the AHP methodology.
- The AHP-method can be considered as a complete aggregation method of the additive type. The problem with such aggregation is that compensation between good scores on some criteria and bad scores on other criteria can occur. Detailed, and often important, information can be lost by such aggregation.
- With AHP the decision problem is decomposed into a number of subsystems, within which and between which a substantial number of pairwise comparisons need to be completed. This approach has the disadvantage that if the number of pairwise comparisons to be made, may become very large ($n(n-1)/2$), and thus a lengthy task.
- Another critical disadvantage of the AHP method is the artificial limitation of the use of the 9-point scale. Sometimes, the decision-maker might find difficult to distinguish among them and tell for example whether one alternative is 6 or 7 times more important than the another. Also, the AHP method cannot cope with the fact that alternative A is 25 times more important than alternative C. So the

decision makers only indicated whether a criterion was more or less important or equally important to its partner.

REFERENCES

- Mirkin, B. G. (1979), Group Choice, John Wiley and Sons, New York.
- Mukherjee, A., T. Mondal, J. K. Bisht and A. Pattanayak (2018), Farmers' preference of fodder trees in mid hills of Uttarakhand: a comprehensive ranking using analytical hierarchy process. *Range Management and Agroforestry*, 39 (1) : 115-120.
- Saaty, T. L. (1980), The Analytic Hierarchy Process, McGraw-Hill Book Co., New York.
- Saaty, T. L. (2008), Decision making with the analytic hierarchy process. *International Journal Services sciences*. 1 (1): 83-98.
- Saaty, T. L. (1986), Axiomatic Foundation of the Analytic Hierarchy Process, *Management Science*, 32: 841-855.
- Saaty, T. L. (1994a), How to make a decision: the analytic hierarchy process. *Interfaces*, 24, 19-43.
- Saaty, T. L. (1994b), Fundamentals of Decision Making, RWS Publications, Pittsburgh, PA
- Wasielewska, K. and M. Ganzha (2012), Using Analytic Hierarchy Process Approach in Ontological Multicriterial Decision Making - Preliminary Considerations. https://www.researchgate.net/publication/258573199_Using_Analytic_Hierarchy_Process_Approach_in_Ontological_Multicriterial_Decision_Making_-_Preliminary_Considerations

Chapter 36

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND BIG DATA

Rajni Jain, Shabana Begam, Sapna Nigam and Vaijunath

INTRODUCTION

The term artificial intelligence (AI) was coined in 1956 by John McCarthy, an American computer and cognitive scientist, who organized the first international conference on AI at Dartmouth, New Hampshire. AI is a branch of science which deals with developing intelligent machines that think and act like humans. It is also defined as the capability of a machine to imitate the intelligent human behaviour. Today, with the advent of the computers and 50 years of research into Artificial Intelligence programming techniques, the dream of creating smart machines is becoming a reality. Intelligence is the ability to adapt one's behaviour to fit new circumstances. This consists of the ability to solve problems, think quickly, act with purpose, think rationally and associate effectively with the environment.

CLASSIFICATION OF AI

There are various ways of classifying AI. Three basis of classifying AI are:

- (1) On the basis of strength
- (2) On the basis of functionalities
- (3) On the basis of domain

On the basis of strength

Weak AI: It is focused on one narrow task, the phenomenon that machines which are not too intelligent to do their own work can be built in such a way that they seem smart. An example would be a poker game where a machine beats human where in which all rules and moves are fed into the machine. Here each and every possible scenario need to be entered beforehand manually. Each and every weak AI will contribute to the building of strong AI.

Strong AI: The machines that can actually think and perform tasks on its own just like a human being. There are no proper existing examples for this but some industry leaders are very keen on getting close to build a strong AI which has resulted in rapid progress.

On the basis of functionalities

1. **Reactive machines:** This is one of the basic forms of AI. It doesn't have past memory and cannot use past information to inform future actions. Example: IBM chess program that beat Garry Kasparov in the 1990s.
2. **Limited memory:** AI systems can use past experiences to inform future decisions. Some of the decision-making functions in *self-driving cars* have been designed this way. Observations used to inform actions happening in the not so distant future, such as a car that has changed lanes. These observations are not stored permanently and also Apple's Chatbot Siri.
3. **Theory of mind:** This type of AI should be able to understand people's emotion, belief, thoughts, expectations and be able to interact socially. Even though a lot of improvements are there in this field, this kind of AI is not complete yet.
4. **Self-awareness:** An AI that has its own conscious, super intelligent, self-awareness and sentient (In simple words a complete human being). Of course, this kind of bot also does not exist and if achieved it will be one of the milestones in the field of AI.

On the basis of domain

The domain of AI is classified into Formal tasks, Mundane tasks, and Expert tasks (Fig 1). Each of these tasks further leads to many categories of applications as shown in the Fig 1.

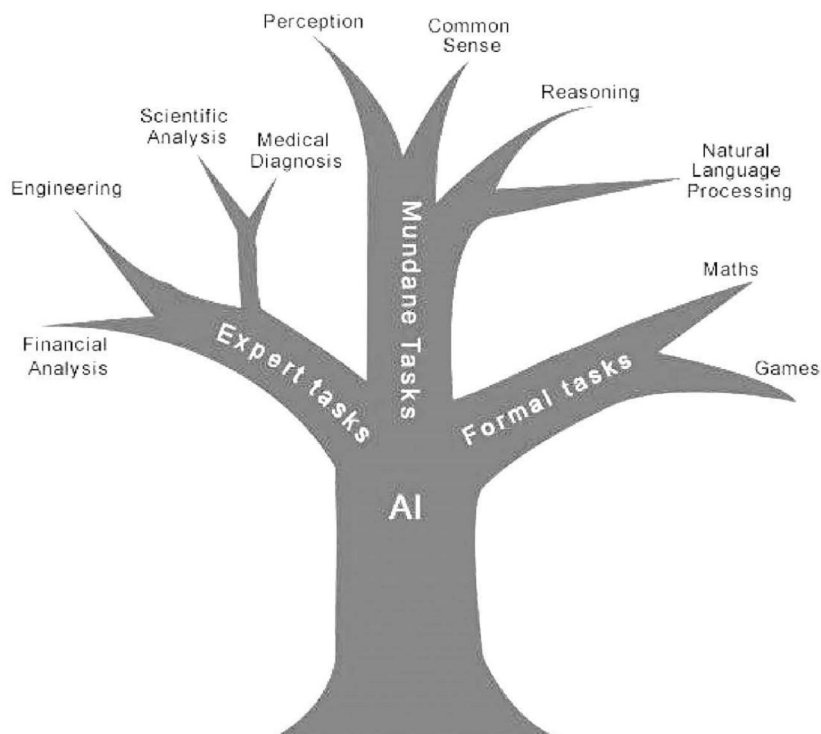


Fig 1: Tasks classification of AI

Applications of AI

Expert systems

A large area of application of artificial intelligence is in expert systems. AI programs that achieve expert-level competence in solving problems in specific task areas by bringing to bear a body of knowledge are called knowledge-based or expert systems. An expert system is software that uses a knowledge base of human expertise for problem solving, or to clarify uncertainties where normally one or more human experts would need to be consulted (Fig 2). The internal structure of an expert system can be considered as consisting of three parts: the knowledge base, inference engine, and user interface. The knowledge base captures the knowledge from the expert. Inference engine acquires and manipulates the knowledge from the knowledge base to arrive at a particular solution. User Interface provides interaction between the user and the Expert System itself. Most of these systems use IF-THEN rules to represent knowledge. Expert systems have been used to facilitate tasks in the fields of accounting, medicine, process control, financial services, production, and human resources among others.

Autonomous planning and scheduling

A hundred million miles from Earth, NASA's remote agent program became the first on-board autonomous planning program to control the scheduling of operations for a spacecraft. Remote agent generates plans from high-level goals specified from the ground. It monitors the operation of the spacecraft as the plans are executed, detecting, diagnosing, and recovering from problems as they occur.

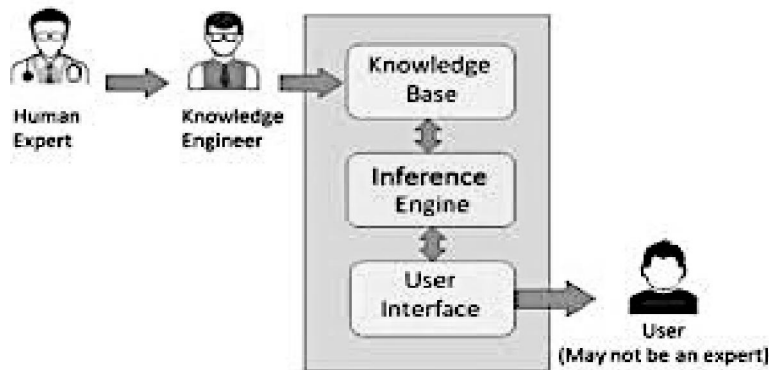


Fig 2: Expert System

Computer vision

Computer vision is a field of artificial intelligence that is used to obtain information from images or multi-dimensional data. Machine Learning algorithms such as K-means is used for Image Segmentation, Support Vector Machine is used for Image Classification and so on.

Therefore, Computer Vision makes use of AI technologies to solve complex problems such as Object Detection, Image Processing, etc.

Game playing

AI plays crucial role in strategic games such as chess, poker etc. Machines can think large numbers of possible positions/options based on heuristics knowledge. Deep Blue became the first computer program to defeat the world champion in a chess match when it bested Garry Kasparov by a score of 3.5 to 2.5 in an exhibition match.

Natural language processing

Natural language processing (NLP) refers to the Artificial Intelligence method that analyses natural human language to derive useful insights in order to solve problems.

Speech recognition

Systems capable of hearing and comprehend the language in terms of sentences and their meanings while human speaks. It can handle different accents, slang words, noise etc.

Fuzzy logic systems

Fuzzy logic is an approach to computing based on “degrees of truth” rather than the usual “true or false” (1 or 0) boolean logic on which the modern computer is based. Fuzzy logic Systems can take imprecise, distorted, noisy input information.

Robotics

It is the branch of AI, composed of mechanical, electrical engineering and computer science that deals with the design, construction, operation and use of robots as well as computer systems for their control, sensory feedback and information processing. Robots are the artificial agents acting in real world environments. The different areas of application of robotics includes industries, military, medicine, entertainment and others.

Artificial neural networks

ANN is the information processing paradigm based on the generalization of mathematical model of human brain. Artificial neural network will consist of three types of layers:

- **Input layer:** This layer receives all the inputs and forwards them to the hidden layer for analysis
- **Hidden layer:** In this layer, various computations are carried out and the result is transferred to the output layer. There can be n number of hidden layers, depending on the problem you’re trying to solve.

- Output layer: This layer is responsible for transferring information from the neural network to the outside world.

Artificial intelligence in economics

Artificial, or computational economics is a research discipline at the interface between computer science and economics. Within the area of computational economics, the field of Agent-based computational economics (ACE) belongs to the discipline of complex adaptive dynamic systems that studies economic processes, including whole economies, as dynamic systems of autonomous interacting agents. Large numbers of individual agents engage repeatedly in local interactions, giving rise to global regularities such as employment and growth rates, income distributions, market institutions, and social conventions. These global regularities in turn feed back into the determination of local interactions. The result is an intricate system of interdependent feedback loops connecting microeconomic behaviours, interaction patterns, and global regularities.

Some underlying structural parameters of an economy, such as demand, production, and cost functions, and therefore equilibrium prices and quantities, can be directly observed, rather than estimated, as would be required if real-world data were used. Point predictions of theoretical models can be computed. Moreover, in addition to observing the underlying structure of the economy, the researcher can specify and control it. There researcher can evaluate a change in one parameter while keeping all else constant and look at its effect in isolation.

The agents in an ACE model can be economic entities as well as social, biological, and physical entities. In ACE, the term agent refers broadly to an encapsulated piece of software that includes data together with behavioural methods that act on these data. Some of these data are publicly accessible to all other agents. Others are designated as private, and hence are not accessible by any other agents, or only accessible to a specified subset of other agents. Agents can communicate with each other through public and/or private channels, depending on the economy that is considered.

Examples of agents include individuals (e.g., consumers, workers), social groupings (e.g., families, firms, government agencies), institutions (e.g., markets, regulatory systems), biological entities (e.g., crops, livestock, forests), and physical entities (e.g., infrastructure, weather, and geographical regions). Thus, agents can range from active data gathering decision-makers with sophisticated learning capabilities, to passive structures with no cognitive functioning. Moreover, hierarchical constructions are permitted, so that, e.g, a firm might be composed of workers and managers.

MACHINE LEARNING

Machine learning is the field of computer science that gives the computers the ability to learn without being explicitly programmed. Machine learning algorithms use

computational methods to learn information directly from the data without relying on a predetermined equation as model (<https://www.expertsystem.com/machine-learning-definition/>). Three categories of machine learning methods are explained in Table 1.

Table 1: Categories of machine learning methods

Parameters	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	Labelled data	Unlabelled data	Agent based interactions and rewards
Type of problems	Regression and classification	Association and clustering	Reward based
Popular Algorithms	Linear regression, logistic regression, support vector machines, KNN, deep learning	K-means, C-means	Q-learning, SARSA
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understands patterns and discover output	Follow trial and error method

In machine learning, we do not have to define explicitly all the steps or conditions like any other programming application. On the contrary, the machine gets trained on a training dataset, large enough to create a model, which helps machine to take decisions based on its learning.

ILLUSTRATIONS

Example 1: Determination of species of a flower

To determine the species of a flower based on its petal and sepal length (leaves of a flower) using machine learning involves three steps (Fig 3).

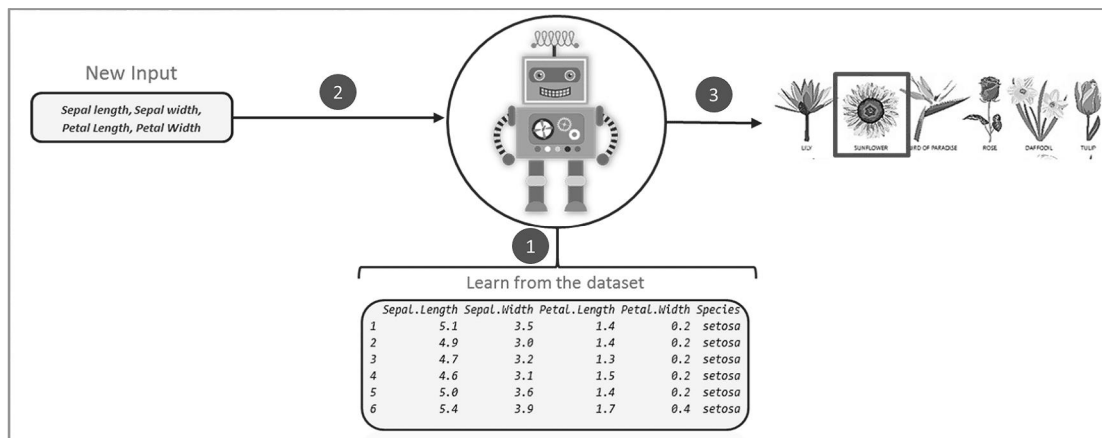


Fig 3: Determining species of flower

Three steps involved are

1. The flower data set is feed as an input which contains various characteristics of different flowers along with their respective species into machine (Figure 3). Using this input data set, the machine will create and train a model which can be used to classify flowers into different categories
2. Once the model has been trained, set of characteristics are passed as input to the model.
3. Finally, model will output the species of the flower present in the new input data set. This process of training a machine to create a model and use it for decision making is called machine learning.

Limitations of machine learning involves incapability of handling high dimensional data that is where input and output is quite large. Handling and processing such large type of data becomes very complex and resource exhaustive. This is termed as curse of dimensionality.

Deep learning: Deep learning is the most recent machine learning technique. It allows computational models that are composed of multiple processing layers to learn representations of data with multiple level of abstraction. Deep learning imitates the way human brain works. The cost of AI is getting cheaper in terms of computation power and in terms of tools. Each new tool/library is helping machine learning developers to spend less time on prediction problems. Three illustrative examples using machine learning are presented for understanding.

Example 2: Face recognition on facebook

Facebook uses Deep Face for face verification. It works on the face verification algorithm, structured by Artificial Intelligence (AI) techniques using neural network models (Fig 4).

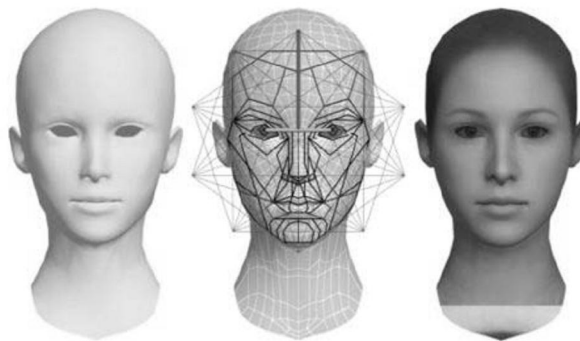


Fig 4: Face verification (Source: Edureka)

Input: Scans a wild form of photos with large complex data. This involves blurry images, images with high intensity and contrast.

Process: In modern face recognition, the process completes in 4 raw steps:

- Detect facial features
- Align and compare the features
- Represent the key patterns by using 3D graphs
- Classify the images based on similarity

Output: Final result is a face representation, which is derived from a 9-layer deep neural net

Training data: More than 4 million facial images of more than 4000 people

Result: Facebook can detect whether the two images represent the same person or not.

Example 3: Market basket analysis

Market basket analysis explains the combinations of products that frequently co-occur in transactions. For example, if a person buys bread, there is a 40% chance that he might also buy butter. By understanding such correlations between items, companies can grow their businesses by giving relevant offers and discount codes on such items. Market Basket Analysis is a well-known practice that is followed by almost every huge retailer in the market. The logic behind this is machine learning algorithms such as Association Rule Mining and Apriori algorithm:

- *Association rule mining is a technique that shows how items are associated with each other.*
- *Apriori algorithm uses frequent itemsets to generate association rules. It is based on the concept that a subset of a frequent itemset must also be a frequent itemset.*

For example, the above rule suggests that, if a person buys item A then he will also buy item B. In this manner the retailer can give a discount offer which states that on purchasing Item A and B, there will be a 30% off on item C. Such rules are generated using machine learning. These are then applied on items in order to increase sales and grow a business.

Example 4: Plant disease identification

AI can be used to implement image processing and classification techniques for extraction and classification of leaf diseases. Disease identification steps are explained below:

Image acquisition: The sample images are collected and stored as an input database.

Image pre-processing: Image pre-processing includes the following:

- Improve image data that suppresses unwanted distortion
- Enhance image features
- Image clipping, enhancement, colour space conversion
- Perform Histogram equalization to adjust the contrast of an image

Image segmentation: It is the process of partitioning a digital image into multiple segments so that image analysis becomes easier. Segmentation is based on image features such as colour, texture. A popular Machine Learning method used for segmentation is the K-means clustering algorithm.

Feature Extraction: This is done to extract information that can be used to find the significance of a given sample. The Haar Wavelet transform can be used for texture analysis and the computations can be done by using Gray-Level Co-Occurrence Matrix.

Classification: Finally, Linear Support Vector Machine is used for classification of leaf disease. SVM is a binary classifier which uses a hyperplane called the decision boundary between two classes. This results in the formation of two classes:

1. Diseased leaves
2. Healthy leaves

Therefore, AI can be used in Computer Vision to classify and detect disease by studying and processing images. This is one of the most profound applications of AI.

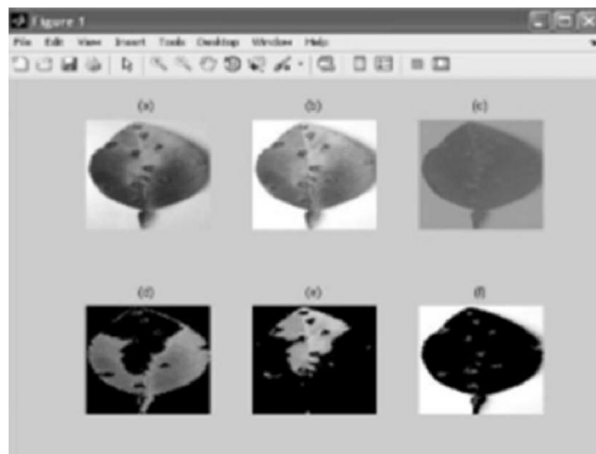


Fig 5: Image processing and disease identification

BIG DATA

Big data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently. These data come from many sources like - Social networking sites, E-commerce site, Weather Station, Telecom company, Share Market. The statistics shows that 500+terabytes of

new data get ingested into the databases of social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

Forms of Big Data: Big Data could be found in three forms: structured, unstructured and semi-structured.

Structured: Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data. Data stored in a relational database management system is one example of a ‘structured’ data (Table 2).

Table 2: Examples of structured data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured: Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. The output returned by ‘Google Search’ is an example of unstructured big data.

Semi-structured: Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file (Fig 6).

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Fig 6: XML file-Example of semi-structured data

Characteristics of big data

Big Data has certain characteristics and hence is defined using 4Vs namely volume, velocity, variety and veracity (Fig 7)

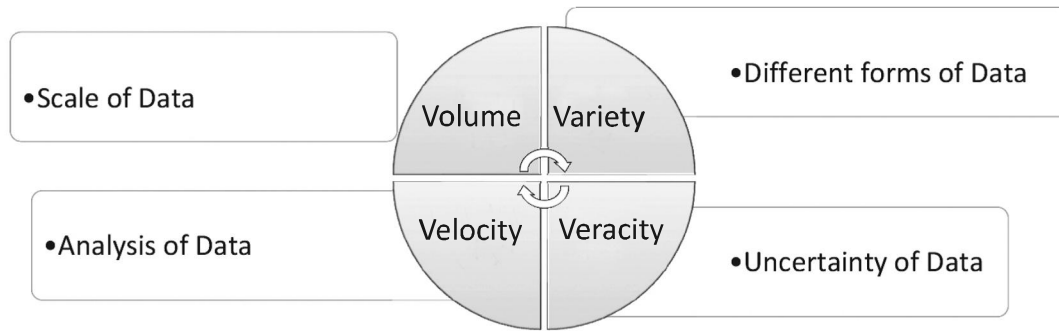


Fig 7: Four Vs of Big data

Volume: The amount of data that businesses can collect is really enormous and hence the volume of the data becomes a critical factor in Big Data analytics (Anuradha, 2015).

Velocity: The rate at which new data is being generated all thanks to our dependence on the internet, sensors, machine-to-machine data is also important to parse Big Data in a timely manner.

Variety: The data that is generated is completely heterogeneous in the sense that it could be in various formats like video, text, database, numeric, sensor data and so on and hence understanding the type of Big Data is a key factor to unlocking its value.

Veracity: Knowing whether the data that is available is coming from a credible source is of utmost importance before deciphering and implementing Big Data for business needs.

Clustered computing-Heart of Big data

Clustered computing is an important strategy employed by most big data solutions. Because of the qualities of big data, individual computers are often inadequate for handling the data at most stages. To better address the high storage and computational needs of big data, computer clusters are a better fit. Big data clustering software combines the resources of many smaller machines, seeking to provide a number of benefits like:

- (i) **Resource pooling:** Combining the available storage space to hold data is a clear benefit, but CPU and memory pooling is also extremely important. Processing large datasets requires large amounts of all three of these resources.
- (ii) **High availability:** Clusters can provide varying levels of fault tolerance and availability guarantees to prevent hardware or software failures from affecting access to data and processing. This becomes increasingly important as we continue to emphasize the importance of real-time analytics.
- (iii) **Easyscalability:** Clusters make it easy to scale horizontally by adding additional machines to the group. This means the system can react to changes in resource requirements without expanding the physical resources on a machine.

Using clusters requires a solution for managing cluster membership, coordinating resource sharing, and scheduling actual work on individual nodes. Cluster membership and resource allocation can be handled by software like **Hadoop's YARN** (which stands for Yet Another Resource Negotiator) or **Apache Mesos**.

Big data life cycle

The general categories of activities involved with big data processing are ingesting data into the system, persisting the data in storage, computing and analysing data, and visualizing the results (Kaisler, 2016) The steps and related technologies are briefly mentioned here.

Step 1: Ingesting data into the system

Data ingestion is the process of taking raw data and adding it to the system. The complexity of this operation depends heavily on the format and quality of the data sources and how far the data is from the desired state prior to processing. One way that data can be added to a big data system are dedicated ingestion tools. Technologies like **Apache Sqoop**, **Apache Flume** and **Apache Chukwa** can take existing data from relational databases and add it to a big data system. During the ingestion process, some level of analysis, sorting, and labelling usually takes place. This process is sometimes called ETL, which stands for extract, transform, and load.

Step 2: Persisting the data in storage

The ingestion processes typically hand the data off to the components that manage storage, so that it can be reliably persisted to disk. While this seems like it would be a simple operation, the volume of incoming data, the requirements for availability, and the distributed computing layer make more complex storage systems necessary. This usually means leveraging a distributed file system for raw data storage. Solutions like **Apache Hadoop's HDFS**, **Ceph** and **GlusterFS** filesystem allow large quantities of data to be written across multiple nodes in the cluster. This ensures that the data can be accessed by compute resources, can be loaded into the cluster's RAM for in-memory operations, and can gracefully handle component failures. Distributed databases, especially NoSQL databases, are well-suited for this role because they are often designed with the same fault tolerant considerations and can handle heterogeneous data.

Step 3: Computing and analysing data

Once the data is available, the system can begin processing the data to surface actual information. The computation layer is perhaps the most diverse part of the system as the requirements and best approach can vary significantly depending on what type of insights are desired. Data is often processed repeatedly, either iteratively by a single tool or by using a number of tools to surface different types of insights.

Batch processing is one method of computing over a large dataset. The process involves breaking work up into smaller pieces, scheduling each piece on an individual machine, reshuffling the data based on the intermediate results, and then calculating and assembling the final result. These steps are often referred to individually as splitting, mapping, shuffling, reducing, and assembling, or collectively as a distributed map reduce algorithm. This is the strategy used by **Apache Hadoop's MapReduce**. Batch processing is most useful when dealing with very large datasets that require quite a bit of computation. While batch processing is a good fit for certain types of data and computation, other workloads require more **real-time processing**. Real-time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available. One way of achieving this is **stream processing**, which operates on a continuous stream of data composed of individual items. Another common characteristic of real-time processors is in-memory computing, which works with representations of the data in the cluster's memory to avoid having to write back to disk.

Apache Storm, Apache Flink, and Apache Spark provide different ways of achieving real-time or near real-time processing. There are trade-offs with each of these technologies, which can affect which approach is best for any individual problem. In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly.

The above examples represent computational frameworks. However, there are many other ways of computing over or analyzing data within a big data system. These tools frequently plug into the above frameworks and provide additional interfaces for interacting with the underlying layers. For instance, **Apache Hive** provides a data warehouse interface for Hadoop and **Apache Pig** provides a high level querying interface. For straight analytics programming that has wide support in the big data ecosystem, both **R** and **Python** are popular choices.

Step 4: Visualizing the results

Visualizing data is one of the most useful ways to spot trends and make sense of a large number of data points. Real-time processing is frequently used to visualize application and server metrics. A visualization technology typically used for interactive data science work is a data “notebook”. It allows for interactive exploration and visualization of the data in a format conducive to sharing, presenting, or collaborating. Popular examples of this type of visualization interface are **Jupyter Notebook** and **Apache Zeppelin**. One of the best-known methods for turning raw data into useful information is what is known as MapReduce. MapReduce is a method for taking a large data set and performing computations on it across multiple computers, in parallel. It serves as a model for how to program and is often used to refer to the actual implementation of this model. In essence, MapReduce consists of two parts. The Map function does sorting and filtering, taking data and placing it inside of categories so that it can be analyzed. The Reduce function provides a summary of this data by combining it all together.

Hadoop: A tool for big data

Most influential and established tool for analyzing big data is known as Apache Hadoop (<https://www.javatpoint.com/what-is-hadoop>). Apache Hadoop is a framework for storing and processing data at a large scale, and it is completely open source (Figure 8). Hadoop can run on commodity hardware, making it easy to use with an existing data center, or even to conduct analysis in the cloud. Hadoop is broken into four main parts:

1. The Hadoop Distributed File System (HDFS), which is a distributed file system designed for very high aggregate bandwidth;
2. YARN (Yet Another Resource Negotiator), a platform for managing Hadoop's resources and scheduling programs that will run on the Hadoop infrastructure;
3. MapReduce, as described above, a model for doing big data processing;
4. A common set of libraries for other modules to use.

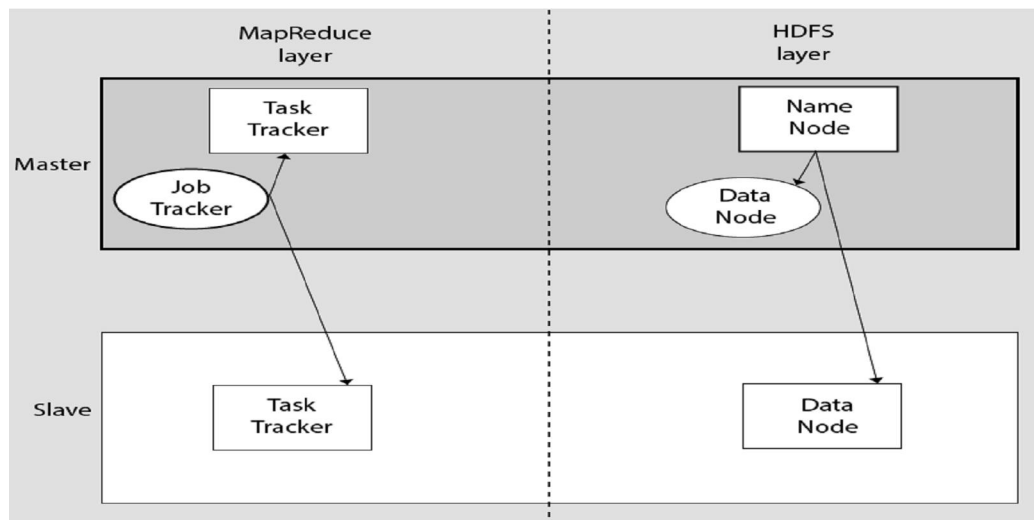


Fig 8: Hadoop architecture

Thus main issue in Big data huge unstructured data which needs to be stored, processed and analysed is solved using Hadoop. For huge data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle. Processing is done by Map Reduce paradigm by applying it to data distributed over network to find the required output. Pig, Hive can be used to analyse the data. Besides, Hadoop is open source so the cost is not an issue. A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.

The production environment of Hadoop is UNIX, but it can also be used in Windows using Cygwin. Java 1.6 or above is needed to run Map Reduce Programs. Further details on Hadoop can easily be browsed on internet.

Some other popular big data tools are mentioned below:

1. Apache Beam is “a unified model for defining both batch and streaming data-parallel processing pipelines”. It allows developers to write code that works across multiple processing engines.
2. Apache Hive is a data warehouse built on Hadoop. A top-level Apache project, it facilitates reading, writing, and managing large datasets ... using SQL.
3. Apache Impala is an SQL query engine that runs on Hadoop. It’s incubating within Apache and is touted for improving SQL query performance while offering a familiar interface.
4. Apache Kafka allows users to publish and subscribe to real-time data feeds. It aims to bring the reliability of other messaging systems to streaming data.
5. Apache Lucene is a full-text indexing and search software library that can be used for recommendation engines. It is also the basis for many other search projects, including Solr and Elasticsearch.
6. Apache Pig is a platform for analyzing large datasets that runs on Hadoop. Yahoo, which developed it to do MapReduce jobs on large datasets, contributed it to the ASF in 2007.
7. Apache Solr is an enterprise search platform built upon Lucene.
8. Apache Zeppelin is an incubating project that enables interactive data analytics with SQL and other programming languages.
9. Other open source big data tools you may want to investigate:
10. Elasticsearch is another enterprise search engine based on Lucene. It is part of the Elastic stack (formerly known as the ELK stack for its components: Elasticsearch, Kibana, and Logstash) that generates insights from structured and unstructured data.
11. Cruise Control was developed by LinkedIn to run Apache Kafka clusters at large scale.
12. TensorFlow is a software library for machine learning that has grown rapidly since Google open sourced it in late 2015. It has been praised for democratizing machine learning because of its ease-of-use.

As big data continues to grow in size and importance, the list of open source tools for working with it will certainly continue to grow as well.

Big Data Applications in Agriculture

Big data has significant potential to address the issues of modern societies, including the needs of consumers, financial analysts, marketing agents, producers, and decision makers. (Coble *et al.*, 2018). Big data has no shortage of uses within farming. Some

of the more prominent include yield prediction, risk management, food safety, and operations management (Bronson *et al.*, 2016; Wolfert *et al.*, 2017).

Yield prediction

Yield prediction sees the use of mathematical models to analyse data around yield, weather, chemicals, leaf and biomass index among others, with machine learning used to crunch the stats and power the making of decisions. Predicting yields in this way can allow a farmer to extract insight on what to plant as well as where and when to plant it. The use of sensors for collecting data means that only a small amount of manual work is required to hand each business an instruction manual on how to guarantee the best return from their crops.

Risk management

It is now possible for farmers to leverage a web of big data with a view to evaluating the chances of events like crop failure, and even improve feed efficiency within the production of livestock. The area of risk management created headlines in 2014 as advice from data scientists to Colombian rice farmers was said to have saved millions in damages caused by shifting weather patterns.

Food safety and spoilage prevention

A critical aspect of modern-day farming is allowing instant detection of microbes and incidents of contamination. The collection of data around things like humidity, temperature and chemicals will paint a picture of health around smart agricultural businesses. That level of insight should be of interest to organic farmers in the US, whose issues with GMO contamination between 2011-2014 was said to have caused damages of \$66,395 per affected business. Perhaps an earlier detection may have lowered the repair bill, or at least reduced some of the wastage.

Operation/equipment management

Finally, we cannot underestimate the role of big data in aiding various aspects of the everyday running of an agricultural business. Equipment manufacturers across the world have already made a good start with their fitting of sensors around vehicles to aid their providing of data. Farmers can then log into special portals to manage their fleet and maintenance of equipment in order to reduce downtime and keep everything productive.

REFERENCES

- Anuradha, J. (2015), A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science*, 48: 319-324.
- Bronson, K. and I. Knezevic (2016), Big Data in food and agriculture. *Big Data and Society*, 3(1): 2053951716648174.
- Coble, K. H., A. K. Mishra, S. Ferrell and T. Griffin (2018), Big data in agriculture: A challenge for the future. *Applied Economic Perspectives and Policy*, 40(1): 79-96.
- Kaisler, S. H., F. J. Armour and J. A. Espinosa (2016), Introduction to the Big Data and Analytics: Concepts, Techniques, Methods, and Applications Minitrack. In *HICSS*(pp. 1059-1060).
- Wolfert, S., L. Ge, C. Verdouw and M. J. Bogaardt (2017), Big data in smart farming—a review. *Agricultural Systems*, 153, 69-80.
- <https://www.javatpoint.com/what-is-hadoop>
- <https://www.expertsystem.com/machine-learning-definition/>

LIST OF CONTRIBUTORS

Abdulla

Research Associate, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: abdullahalig007@gmail.com

Abimanyu Jhahria

Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: abhimanyujhahria@gmail.com

Achal Lama

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: achal.lama@icar.gov.in

Ankur Bisvas

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: ankur.bckv@gmail.com

Anirban Mukherjee

Scientist, ICAR Research Complex for Eastern Region, Patna
e-mail id: anirban.extn@gmail.com

Anuja A. R.

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: Anuja.AR@icar.gov.in

Arathy Ashok

Scientist (Sr. Scale), ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: arathyashok@gmail.com

Arpan Bhowmik

Scientist (Sr. Scale), ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: arpan.bhowmik@icar.gov.in

Balaji S. J.

Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: balajiniap@gmail.com

Bishal Gurung

Scientist (Sr. Scale), ICAR-Indian Agricultural Statistics Research Institute, New Delhi

e-mail id: Bishal.Gurung@icar.gov.in

Chandra Sen

Professor (Retired), Banaras Hindu University, Varanasi

e-mail id: chandra.sen@bhu.ac.in

Deepak Singh

Scientist (Sr. Scale), ICAR-Indian Agricultural Statistics Research Institute, New Delhi

e-mail id: deepaksingh2112@gmail.com

Dharam Raj Singh

Principal Scientist, ICAR-Indian Agricultural Research Institute, New Delhi

e-mail id: drsingh@iari.res.in

Gaurav Kumar Vani

Assistant Professor, JNKVV, Jabalpur, Madhya Pradesh

e-mail id: kumaragri.vani1@gmail.com

Girish Kumar Jha

Principal Scientist, ICAR-Indian Agricultural Research Institute, New Delhi

e-mail id: gkjha@iari.res.in

Harish Kumar H. V.

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi

e-mail id: Harishkumar.HV@icar.gov.in

Jaiprakash Bishen

Scientist, ICAR-National Rice Research Institute, Cuttack, Odisha.

e-mail id: jpbisen.iari@gmail.com

Jaspal Singh

Consultant, NITI Aayog, New Delhi

e-mail id: Punjabimatti82@gmail.com

K. N. Singh

Head of Department, Forecasting & Agricultural Systems Modelling, ICAR-Indian Agricultural Statistics Research Institute, New Delhi

e-mail id: Kn.Singh@icar.gov.in

K. S. Aditya

Scientist, ICAR-Indian Agricultural Research Institute, New Delhi

e-mail id: adityaag68@gmail.com

Kanchan Sinha

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: kanchan.sinha@icar.gov.in

Kingsly I. T.

Scientist (Sr. Scale), ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: k.immanuelraj@icar.gov.in

Krishan Lal

Principal Scientist (Retired), ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Kumari Shubha

Scientist, Division of Crop Research, ICAR - Research Complex for Eastern Region, Patna

L. M. Bhar

Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: lm.bhar@icar.gov.in

M. Balasubramanian

Scientist, ICAR-Indian Agricultural Research Institute, New Delhi
e-mail id: bala.sbrmn@gmail.com

M. S. Raman

Research Associate, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: ramnomics@hotmail.com

Mrinmoy Ray

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: mrinmoy4848@gmail.com

N. Sivaramane

Principal Scientist, National Academy of Agricultural Research Management, NAARM, Hyderabad
e-mail id: sivaramane@gmail.com

P. Adhiguru

Principal Scientist, ICAR – Agricultural Extension Division, KAB-I, New Delhi
e-mail id: padhiguru@gmail.com

P. Sethuraman Sivakumar

Principal Scientist, ICAR - Central Tuber Crop Research Institute, Thiruvananthapuram
e-mail id: sethu_73@yahoo.com

Philip Kuriachen

PhD Scholar, ICAR-Indian Agricultural Research Institute, New Delhi
e-mail id: philipkuriachen@gmail.com

Prabhat kishor

Scientist, ICAR-National Institute of Agricultural Economics and Policy Research,
New Delhi
e-mail id: prabhat.kishore@icar.gov.in

Prem Chand

Scientist (Sr. Scale), ICAR-National Institute of Agricultural Economics and Policy
Research, New Delhi
e-mail id: prem.chand@icar.gov.in

Rajender Parsad

ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: rajender.parsad@icar.gov.in

Rajesh T.

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: Rajesh.T@icar.gov.in

Rajni Jain

Principal Scientist, ICAR - National Institute of Agricultural Economics and Policy
Research, New Delhi
e-mail id: rajnijain67@gmail.com

Raju Kumar

Principal Scientist, ICAR - Indian Agricultural Statistics Research Institute,
New Delhi
e-mail id: raju.kumar@icar.gov.in

Raka Saxena

Principal Scientist, ICAR - National Institute of Agricultural Economics and Policy
Research, New Delhi
e-mail id: raka.saxena@icar.gov.in

Ramasubramanian V.

Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute,
New Delhi
e-mail id: R.Subramanian@icar.gov.in

Ranjit Kumar Paul

Scientist (Sr. Scale), ICAR-Indian Agricultural Statistics Research Institute,
New Delhi
e-mail id: ranjitstat@gmail.com

Ravindra Singh Shekhawat

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: ravindra.shekhawat@icar.gov.in

Rohit Kumar

Research Associate, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: rohitjnvproduct@gmail.com

Sapna Nigam

PhD Scholar, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: sapna.nigam1010@gmail.com

Seema Jaggi

Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: seema.jaggi@icar.gov.in

Shabana Begam

PhD Scholar, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: shaba.shb@gmail.com

Shinoj Parappurathu

Senior Scientist, ICAR-Central Marine Fishery Research Institute, CMFRI, Kochi
e-mail id: pshinoj@gmail.com

Shiv Kumar

Principal Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: shivkumardull@gmail.com

Shivaswamy G. P.

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: Shivaswamy.GP@icar.gov.in

Shivendra Kumar Srivastava

Scientist (Sr. Scale), ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: sk.srivastava@icar.gov.in

Subash S. P.

Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi
e-mail id: subashspar@gmail.com

Sudipta Paul

Scientist (Sr. Scale), ICAR-Indian Agricultural Research Institute, New Delhi
e-mail id: sudiptaiari@gmail.com

Sujit Sarkar

Scientist (Sr. Scale), ICAR-Indian Agricultural Research Institute, Regional Station,
Kalimpong, West Bengal
e-mail id: Sujit.Sarkar@icar.gov.in

Sukanta Dash

Scientist (Sr. Scale), ICAR-Indian Agricultural Statistics Research Institute,
New Delhi
e-mail id: sukanta.dash@icar.gov.in

Suresh Kumar

Scientist (Sr. Scale), ICAR-Indian Institute of Soil and Water Conservation,
Dehradun
e-mail id: skdcswrcrti@gmail.com

Suresh Pal

Director, ICAR-National Institute of Agricultural Economics and Policy Research,
New Delhi
e-mail id: suresh.pal20@gov.in

Vaijunath kumasagi

PhD Scholar, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
e-mail id: vaijunathsk21@gmail.com

Venkatesh P.

Senior Scientist, ICAR-Indian Agricultural Research Institute, New Delhi
e-mail id: venkatesh1998@gmail.com

Vinayak Nikam

Scientist (Sr. Scale), ICAR-National Institute of Agricultural Economics and Policy
Research, New Delhi
e-mail id: vinayakrnkam@gmail.com

Vinita Kanwal

Research Associate, ICAR-National Institute of Agricultural Economics and Policy
Research, New Delhi
e-mail id: vinitakanwal888@gmail.com

