TB-ICN: 299/2023

NAHEP

World Bank – ICAR funded

National Agricultural Higher Education Project

Centre for Advanced Agricultural Science and Technology (CAAST)

On

Genomics Assisted Crop Improvement and Management

Training Manual

सामाजिक विज्ञान हेतु उन्नत अनुसंधान विधियाँ एवं आवश्यक कौशल

Advanced Research Methods and Essential Skills for Social Sciences

December 12 – 22, 2022

कृषि अर्थशास्त्र संभाग
भा.कृ.अ.प.–भारतीय कृषि अनुसंधान संस्थान
नई दिल्ली–110012

Division of Agricultural Economics
ICAR – Indian Agricultural Research Institute
New Delhi – 110012
www.nahep-caast.iari.res.in

NAHEP sponsored training programme

on

Advanced Research Methods and Essential Skills for Social Sciences

Course Director

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

Coordinators

R R Burman
Principal Scientist
Division of Agricultural Extension
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail: burman_extn@hotmail.com

Praveen K V
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail: veenkv@gmail.com

Asha Devi S S
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail:ash.nibha@gmail.com

TB-ICN: 299/2023

© 2023- ICAR- Indian Agricultural Research Institute, New Delhi- 110012

Division of Agriculture Economics
ICAR-Indian Agricultural Research Institute
New Delhi-110012

About NAHEP-CAAST at IARI, New Delhi

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

Objectives:

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

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

Core-Team Members:

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

Associate Team

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

Preface

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

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

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

Alka Singh
R.R. Burman
Praveen K.V.
Asha Devi S.S.

Date: 10.12.2022

Acknowledgments

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

Contents

An Introduction to STATA Software and Extraction and Handling of Unit-level Data Sets of NSSO and ASI

Nithyashree M.L. and Asha Devi S.S.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: nithya.econ@gmail.com; ash.nibha@gmail.com

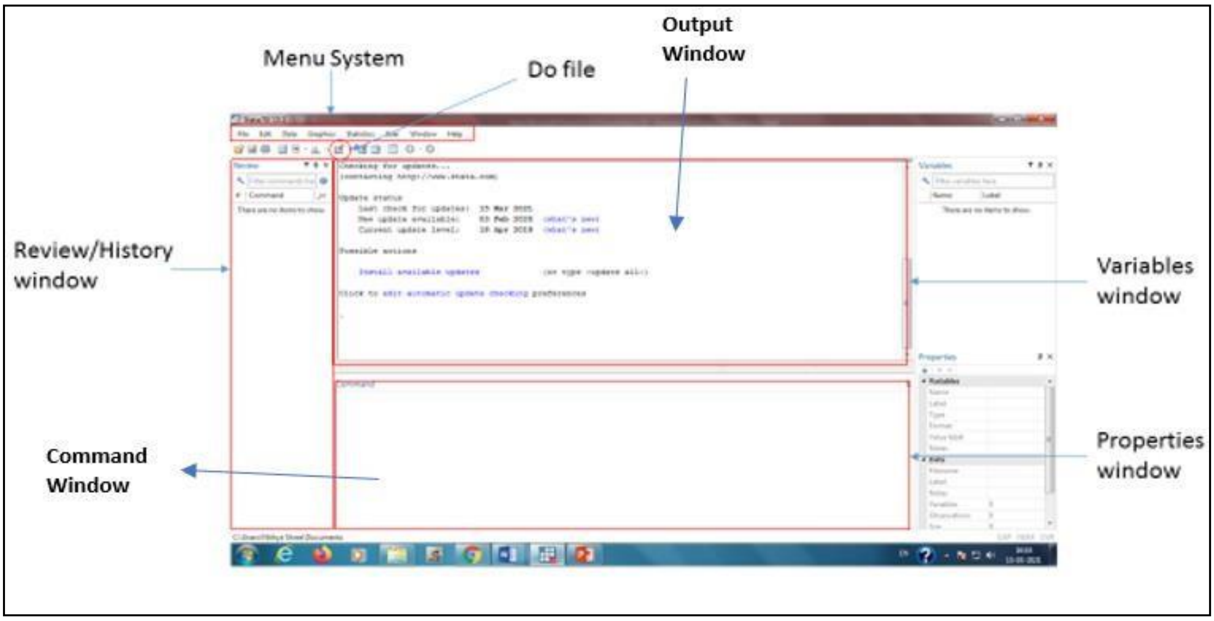
An Introduction to STATA

This chapter describes an overview and basic commands used in the STATA software, which will help beginners to get familiar with and hold a grip on using the software for further statistical analysis. Also, this chapter aims to introduce the different unit-level data sets available and how they are handled. Those who don’t have access to STATA software can avail of the short-term student license service by using the link Student short-term license request | Stata.

Stata is a multi-purpose statistical package that is widely used in social science research to explore, summarize and analyze datasets. As shown below, the STATA's main interphase (Figure 1) has five panes/windows. The result pane displays the commands and resulting outputs, while the command window is meant to type different commands to carry out the analysis. The review pane keeps a record of commands entered in the current session of stata which can be clicked back to the command pane if needed. The variables pane showcases the list of variables in the current dataset, and the variables can be brought into the command window by double-clicking on the particular variable. The properties pane displays the dataset properties.

Operations in Stata can be carried out either through a menu system or a command system. A menu system enables the users to perform the task by interactive UI; drop-down menu. Alternatively, there is also a command window with which users can write commands directly for statistical analysis. Stata commands are case-sensitive and are in lowercase. For example, to summarize data, by typing summarize variable/variables name, we can get the result, and one can also do it by using the drop-down menu by going to current statistics (Figure 2). As a beginner, one can start by exploring the possible option available by using the drop-down menu; after getting comfortable with the software, one can directly type the command to perform any required analysis, which is more efficient and professional, apart from this the executed commands can be copied from the history/review window and saved in do file, which can be easily shared with other researchers and this enables single-click execution.

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



```
Checking for updates...
(contacting http://www.stata.com)

Update status
    Last check for updates:  [illegible]
    New update available:    [illegible]  (what's new)
    Current update level:    [illegible]  (what's new)

Possible actions

    Install available updates          (or type -update all-)

Click to edit automatic update checking preferences
```

Figure 1. Main interphase of STATA

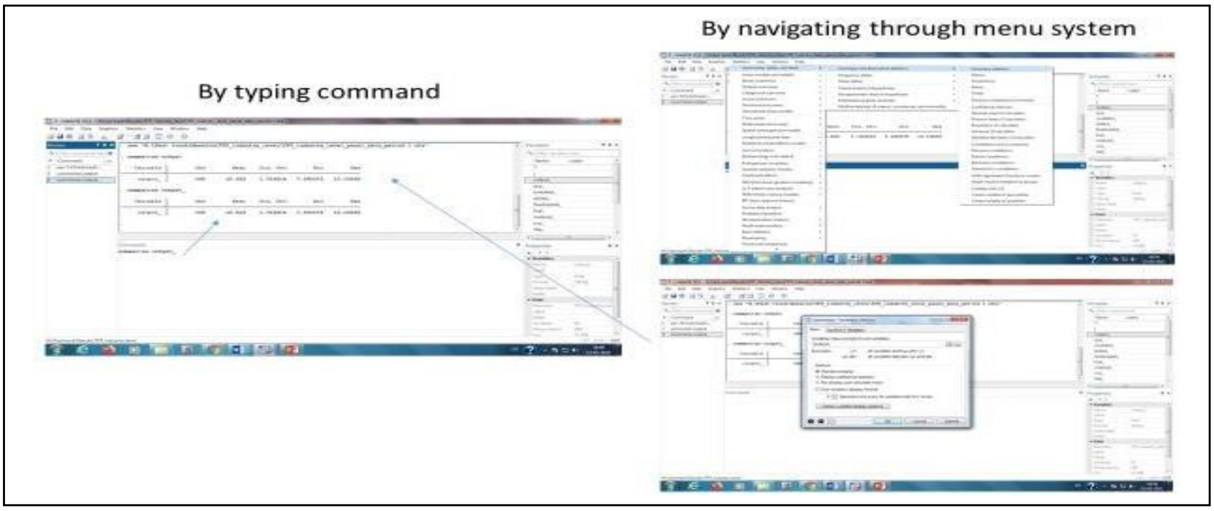
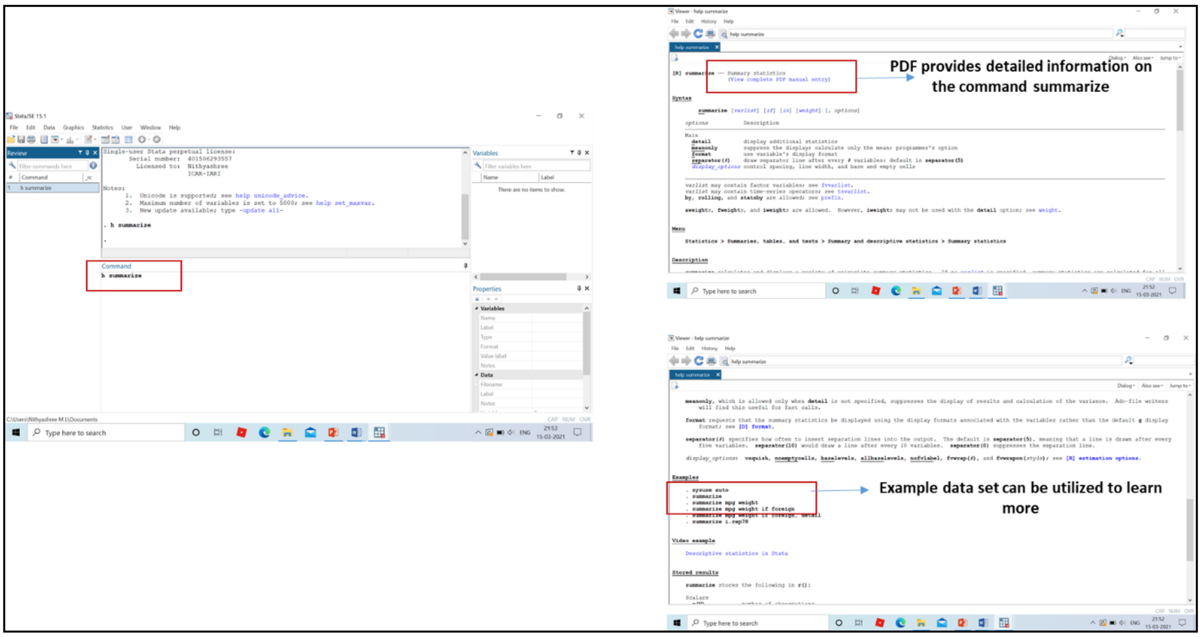


Figure 2. Two ways of performing task in STATA

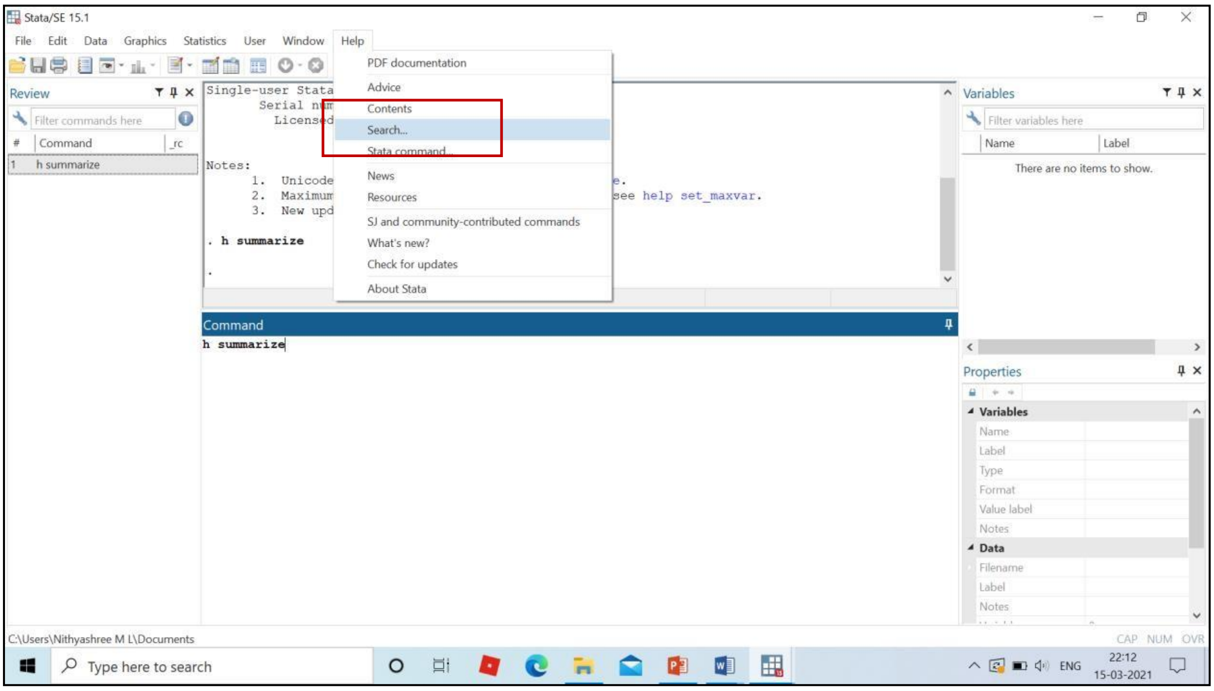


```
. h summarize
```

PDF provides detailed information on the command summarize

Example data set can be utilized to learn more

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



```
Single-user Stata
    Serial nu
      Licensed

Notes:
      1.  Unicode
      2.  Maximum                          see help set_maxvar.
      3.  New upd

. h summarize

.
```

Command: h summarize

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

Mathematical and logical operators in STATA is similar to those used in MS excel

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

Handling the Data set

We can import data of different formats to stata, such as excel, CSV, text etc., or one can use built-in datasets available with stata for practice using *sysuse* command. Besides, there is a large collection of data available in stata website, which can be accessed using *webuse* command and are accessible through this link https://www.stata-press.com/data/r17/. To import data, go to the File menu, click on import, and select your data as per data format.

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

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

2. **For labelling the variable:** to create a label to the variable, write the command label variable and type the variable name need to be labelled then write the label with in the invited comma
   Example: label variable ln_wage "Log of hourly wage"

3. **The replace command:** Replace command generally helps in editing value of already existing or generated variable, for this command ***replace*** can be used.
   Example: replace gender=0 if missing(gender)
   replace gender=0 if gender==.

4. To sort the data in ascending order: use command ***sort variable name***
5. Command that can convert string to numeric variables: ***Tostring and destring***

Few Commands for Basic Statistical analysis

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

Few useful user written commands

h. *tatable2*: Calculate group wise mean value and test the significance
i. *orth_out*: Perform t-test for any number of variables at once
j. *dea*: Data envelopment analysis
k. *acfest*: Production function est. using Ackerberg-Caves-Frazer method
l. *levpet*: Production function est. using Levinsohn and Petrin approach
m. doubleb: Perform Double Bound Contingent Valuation.
n. clustersampsi: perform power calculations for RCT

Exercise-1

1. Import the dataset “stata_intro”, which has been shared with you already.
2. Summarize the data so that you see the means, standard deviation, min, and max of each variable
3. generate the logarithmic transformation of the variable wage as ln_wage
4. label variable ln_wage, wage, collgrad and union as Log of hourly wage, Hourly wage, College graduate and Union member respectively.
5. define label values for the variable collgrad
6. encode racecat as race
7. obtain mean wage sd wage for the variable union
9. draw histogram for the variable ln_wage and superimpose normal curve
10. generate scatter plot for the variables wage tenure
11. generate scatter plot for the variables wage tenure by union
12. obtain a liner regression equation of ln_wage and tenure
13. obtain a liner regression equation ln_wage and tenure along with by considering any one of the categorical variable and also an interaction component
14. save the commands in do file and results in log file.

Extraction and Handling of Unit-level Data

As a beginner before getting the hands-on unit-level data, it is important to go through the key indicators or summary results. This helps to understand the purpose of the survey and the data coverage in the particular dataset. Besides, this will also help the user to comprehend the sample size, sampling design and estimation procedure which is necessary for further analysis and interpretation of the results. Since the key indicators/summary results, by and large, provides the aggregate estimate, the use of unit-level data enables the researchers to work with the basic unit of survey i.e. household in case of agriculture and firm with regard to the industry sector. The unit-level data sets are generally available in text format, to convert these files into a usable format, some of the steps need to be understood and they are discussed by using two unit-level datasets *viz.*, Key Indicators of Situation of Agricultural Households in India of National Sample Survey Organization (NSSO) and Annual Survey of Industries (ASI). For handling any unit-level data it is essential to use suitable statistical software, here we use STATA (version 15) software for the illustration and the summary of the steps to be followed is shown in Figure 1.

The first and most important step in using unit-level data is to get a though understanding of the supporting documents they mainly comprise the layout which instructs, how the data arranges in the text file and other details such as tabulation programme, code list, concept and definition and schedule, etc. For extracting data from the text file one has to write the commands in *STATA* by using the information given in the layout.

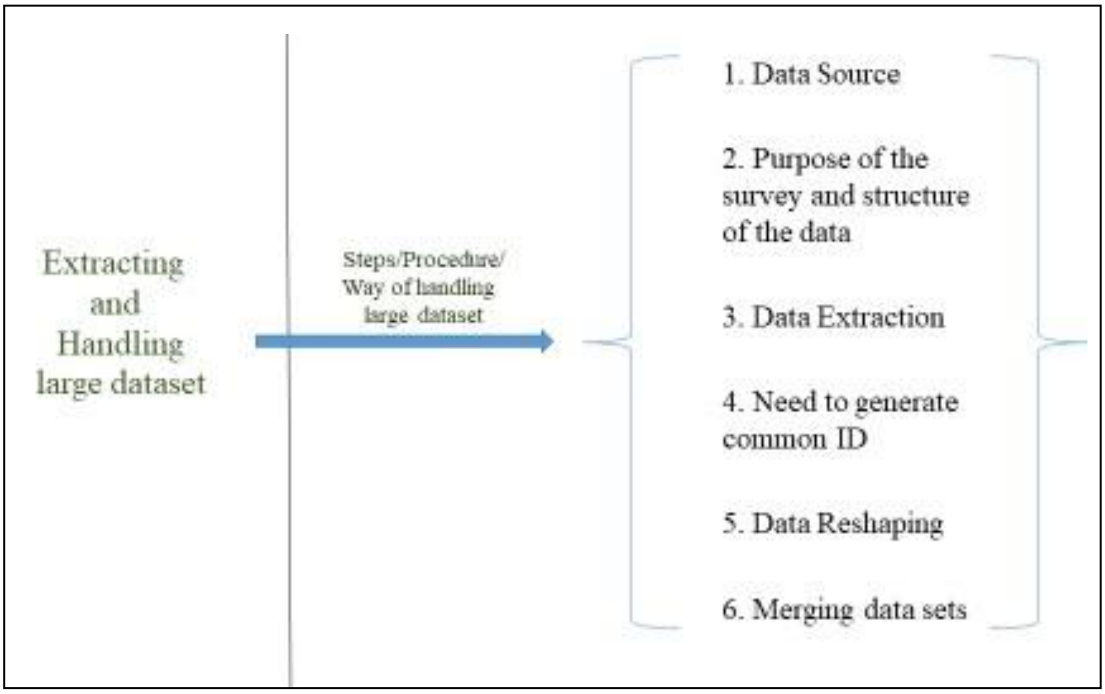

Extracting and Handling large dataset
Steps/Procedure/ Way of handling large dataset
1. Data Source
2. Purpose of the survey and structure of the data
3. Data Extraction
4. Need to generate common ID
5. Data Reshaping
6. Merging data sets

Figure 1. Steps to extract and handle the unit-level data

For example, for the NSSO data set it can be written as:

```
infix Centercode 1-3 FSU 4-8 Round 9-10 Schedule 11-13 Sample 14 Sector 15 NSS_Region 16-18 District 19-20 Stratum
21-22 Sub_Stratum 23-24 Sub_Round 25 Sub_sample 26 FOD_subRegion 27-30 Hamlet 31 SecondStageStratum 32 ///
SampleHHno 33-34 VisitNo 35 Level 36-37 Filler 38-40 HouseholdSize 43-44 Religion 45 Social_Group 46 Dwelling 47
StructureType 48 WaterSource 49 HH_OwnLand_YN 50 LandType 51 PossesLandOutsideVillage 52 LandOperated 53
Land_Own_ha 54-62 Land_Leasedin_ha 63-71 Land_neitherLeased 72-80 Land_LeasedOut_ha 81-89
Land_total_possessed 90-98 Cultivation_whetherPerformed 99 Cultivation_incomesource 100
Livestock_whetherPerformed 101 Livestock_incomesource 102 OtherAgr_whetherPerformed 103
OtherAgr_incomesource 104 NonAgr_whetherPerformed 105 NonAgr_incomesource 106 Wage_whetherPerformed 107
Wage_incomesource 108 Pension_whetherPerformed 109 Pension_incomesource 110 Remittance_whetherPerformed 111
Remittance_incomesource 112 Others_whetherPerformed 113 Others_incomesource 114 MGNREGACard_YN 115
RationCard_YN 116 RationCard_Type 117 NSS 127-129 NSC 130-132 MLT 133-142 using "C:\Users\Nithyashree M
L\Desktop\Stata_Workshop\Unit level data\AH0233V1.TXT"
```

where the command infix is written to extract the text file into *STATA* file and at the end data path has been specified. Similarly, for the ASI data it can be written as:

```
infix year 1-4 Factory 5-9 State 10-11 str C 12 str Block 13 Scheme_code 14 NIC4digit 15-18 NIC5digit 19-23 R_U 24
RO_SRO 25-29 noofunits 30-32 Statusofunit 33-34 Manufacturingdays 35-37 Non_Manufacturingdays 38-40 Total 41-43
Costofproduction 44-57 directlyexported 58-60 Multiplierfactor 61-73 using "C:\Users\Nithyashree M
L\Desktop\Stata_Workshop\Unit level data\ASI\OASBLA16.TXT"
```

After extracting the data files, it is important to create the common ID or unique ID in each data file. This will help to identify the household across the different data sets and also enables to combine the information available in different data files. For creating a common id, the following command in *STATA* is written as:

```
NSSO
egen id=concat (FSU Hamlet SecondStageStratum SampleHHno)
ASI
egen id=concat(Factory State C)
```

To combine the information of different data files, the researcher has to ensure that the observations are uniquely identified across the data sets, if not the data needs to be rearranged. By observing the data structure, it can be grouped as wide-format or long format. For example, in Figure 2, students' ids and results are written in the wide-format in Table 1 and the same students' subject-wise grades are arranged in the long format in Table 2. To combine the information of grade and results, Table 2 has to be reshaped/rearranged into a wide format, which can be done by using the command reshape in STATA i.e. reshape wide Grade, i (ID) j(J).

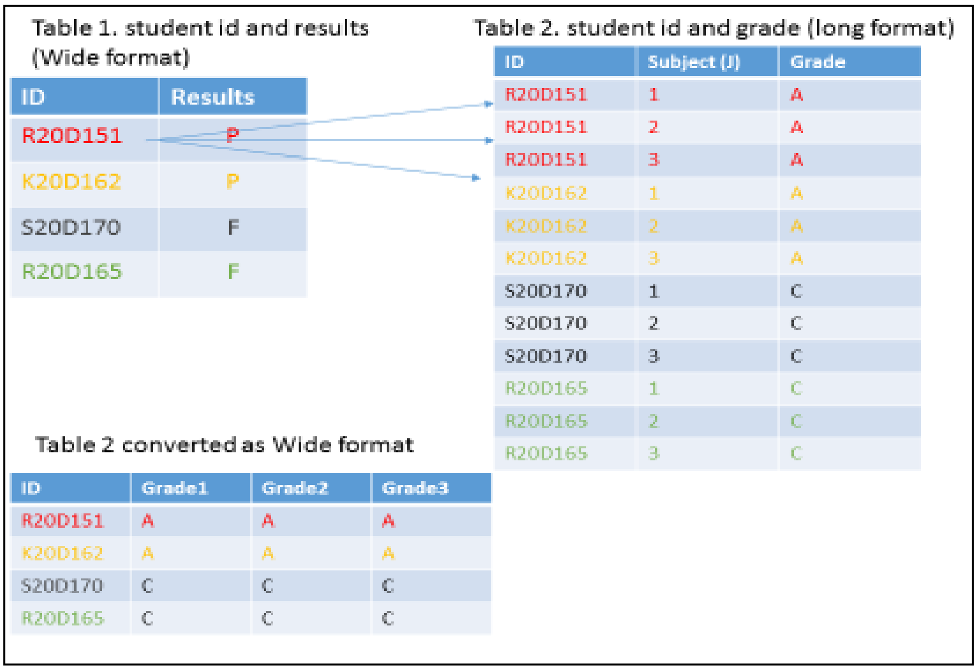


Table 1. student id and results (Wide format)

ID	Results
R20D151	P
K20D162	P
S20D170	F
R20D165	F

Table 2. student id and grade (long format)

ID	Subject (J)	Grade
R20D151	1	A
R20D151	2	A
R20D151	3	A
K20D162	1	A
K20D162	2	A
K20D162	3	A
S20D170	1	C
S20D170	2	C
S20D170	3	C
R20D165	1	C
R20D165	2	C
R20D165	3	C

Table 2 converted as Wide format

ID	Grade1	Grade2	Grade3
R20D151	A	A	A
K20D162	A	A	A
S20D170	C	C	C
R20D165	C	C	C

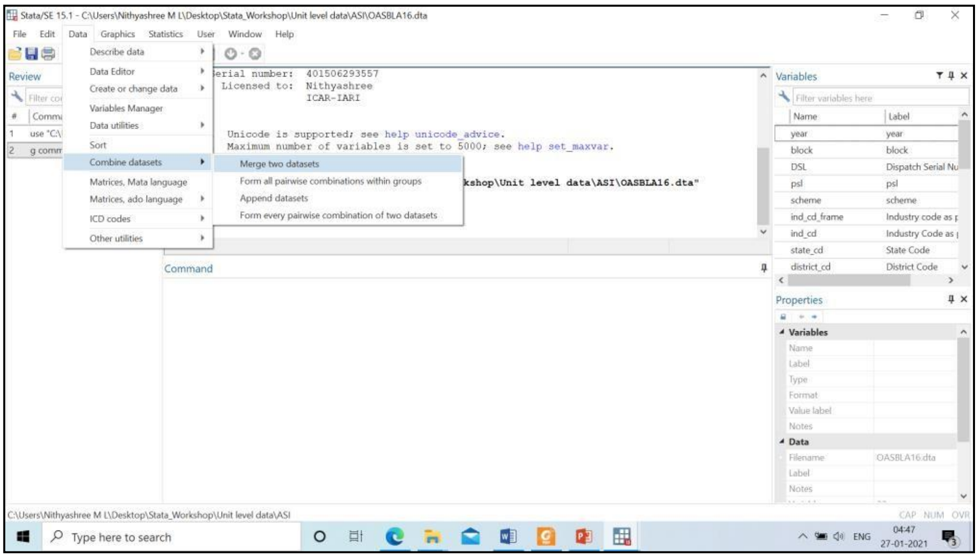
Figure 2. Visualization of wide and long format data

Similarly, to the unit-level data set command can be written as:

NSSO
reshape wide CropCode- PreHarvestSaleValue, i(id) j(SerialNo)

ASI
reshape wide GrossValueOpening - NVC , i(DSL) j(S_no)

In the next step, data files can be merged by using the common ID with the merge option one-to-one on key variable, as shown below



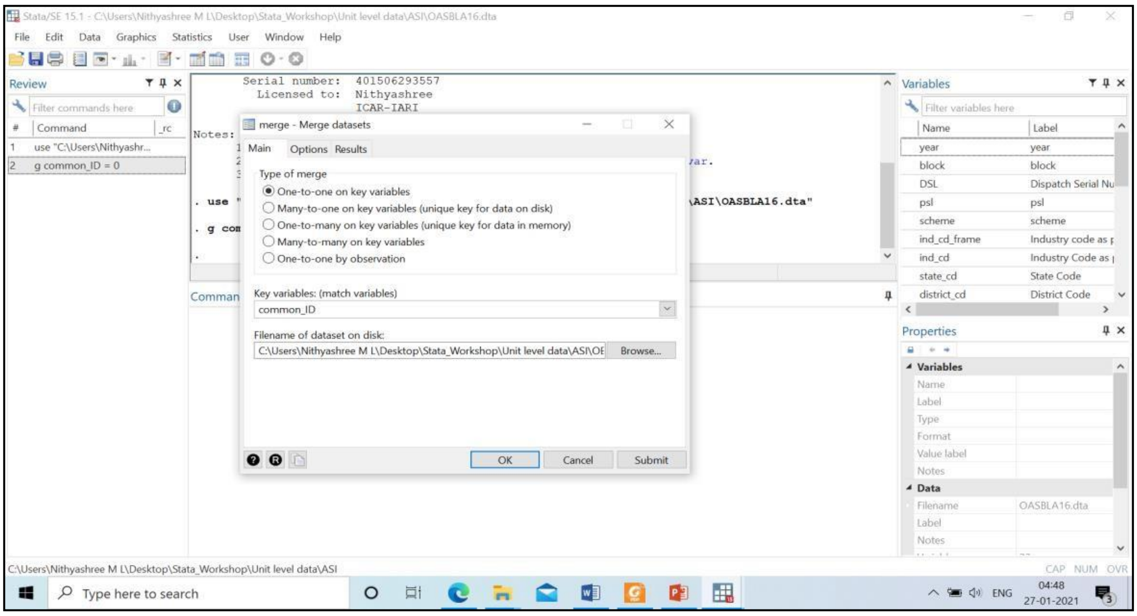
Stata/SE 15.1 - C:\Users\Nithyashree M L\Desktop\Stata_Workshop\Unit level data\ASI\OASBLA16.dta

```
Serial number:  401506293557
 Licensed to:  Nithyashree
               ICAR-IARI

 Unicode is supported; see help unicode_advice.
 Maximum number of variables is set to 5000; see help set_maxvar.

kshop\Unit level data\ASI\OASBLA16.dta"
```

Name	Label
year	year
block	block
DSL	Dispatch Serial Nu
psl	psl
scheme	scheme
ind_cd_frame	Industry code as p
ind_cd	Industry Code as p
state_cd	State Code
district_cd	District Code

Filename OASBLA16.dta

C:\Users\Nithyashree M L\Desktop\Stata_Workshop\Unit level data\ASI



Stata/SE 15.1 - C:\Users\Nithyashree M L\Desktop\Stata_Workshop\Unit level data\ASI\OASBLA16.dta

```
      Serial number:  401506293557
        Licensed to:  Nithyashree
                      ICAR-IARI

Notes:

. use "...\ASI\OASBLA16.dta"

. g com...

.
```

#	Command
1	use "C:\Users\Nithyashr...
2	g common_ID = 0

merge - Merge datasets

Main | Options | Results

Type of merge
- ● One-to-one on key variables
- ○ Many-to-one on key variables (unique key for data on disk)
- ○ One-to-many on key variables (unique key for data in memory)
- ○ Many-to-many on key variables
- ○ One-to-one by observation

Key variables: (match variables)
common_ID

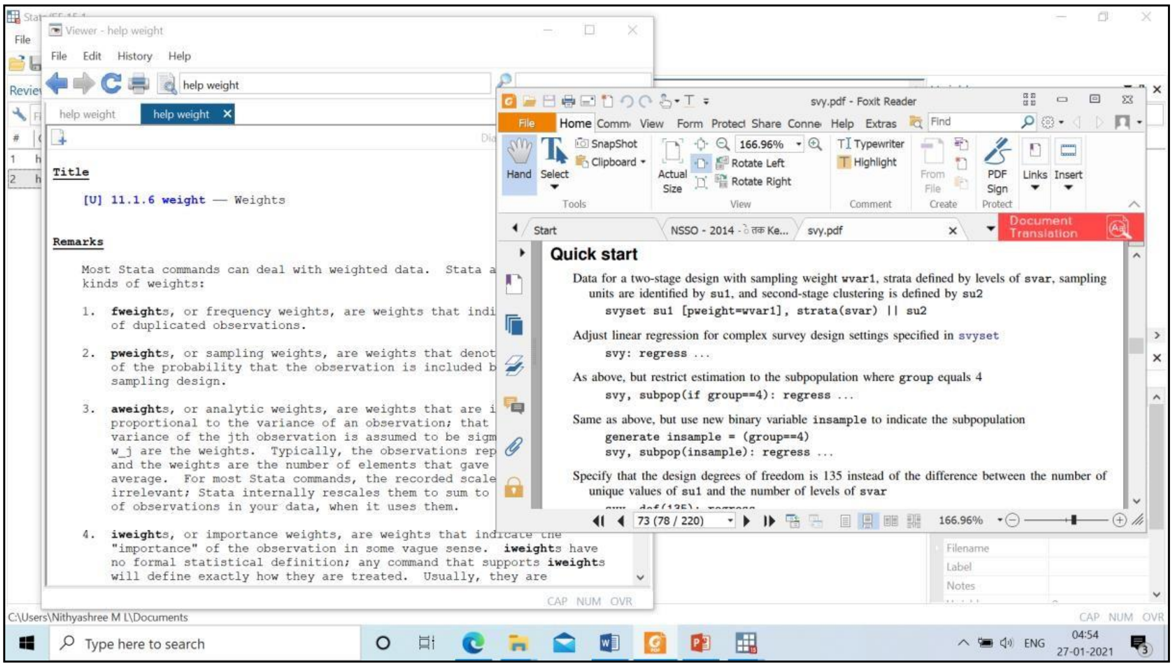
Filename of dataset on disk:
C:\Users\Nithyashree M L\Desktop\Stata_Workshop\Unit level data\ASI\OE

Name	Label
year	year
block	block
DSL	Dispatch Serial Nu
psl	psl
scheme	scheme
ind_cd_frame	Industry code as
ind_cd	Industry Code as
state_cd	State Code
district_cd	District Code

Filename: OASBLA16.dta

C:\Users\Nithyashree M L\Desktop\Stata_Workshop\Unit level data\ASI

Based on the need for interpretation, multiplier options can be further explored by using the help command as shown below.



Viewer - help weight

help weight

```
Title

    [U] 11.1.6 weight — Weights


Remarks

    Most Stata commands can deal with weighted data.  Stata a
    kinds of weights:

    1.  fweights, or frequency weights, are weights that indi
        of duplicated observations.

    2.  pweights, or sampling weights, are weights that denot
        of the probability that the observation is included b
        sampling design.

    3.  aweights, or analytic weights, are weights that are i
        proportional to the variance of an observation; that
        variance of the jth observation is assumed to be sigm
        w_j are the weights.  Typically, the observations rep
        and the weights are the number of elements that gave
        average.  For most Stata commands, the recorded scale
        irrelevant; Stata internally rescales them to sum to
        of observations in your data, when it uses them.

    4.  iweights, or importance weights, are weights that indicate the
        "importance" of the observation in some vague sense.  iweights have
        no formal statistical definition; any command that supports iweights
        will define exactly how they are treated.  Usually, they are
```

svy.pdf - Foxit Reader

Quick start

Data for a two-stage design with sampling weight wvar1, strata defined by levels of svar, sampling units are identified by su1, and second-stage clustering is defined by su2

```
svyset su1 [pweight=wvar1], strata(svar) || su2
```

Adjust linear regression for complex survey design settings specified in svyset

```
svy: regress ...
```

As above, but restrict estimation to the subpopulation where group equals 4

```
svy, subpop(if group==4): regress ...
```

Same as above, but use new binary variable insample to indicate the subpopulation

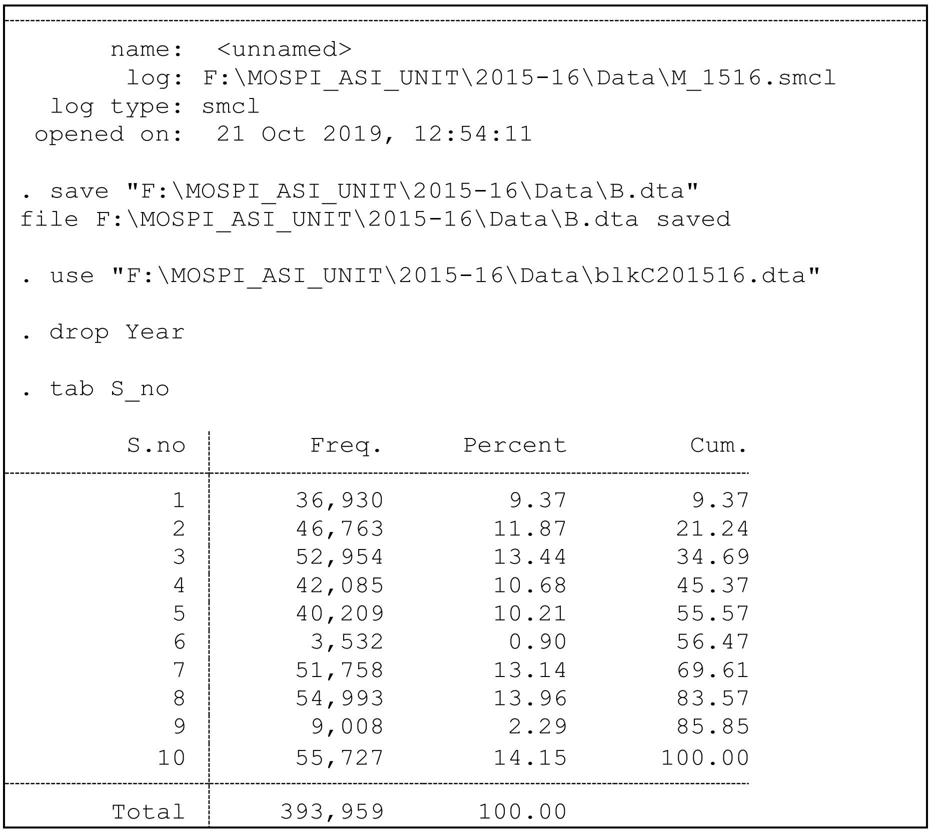
```
generate insample = (group==4)
svy, subpop(insample): regress ...
```

Specify that the design degrees of freedom is 135 instead of the difference between the number of unique values of su1 and the number of levels of svar

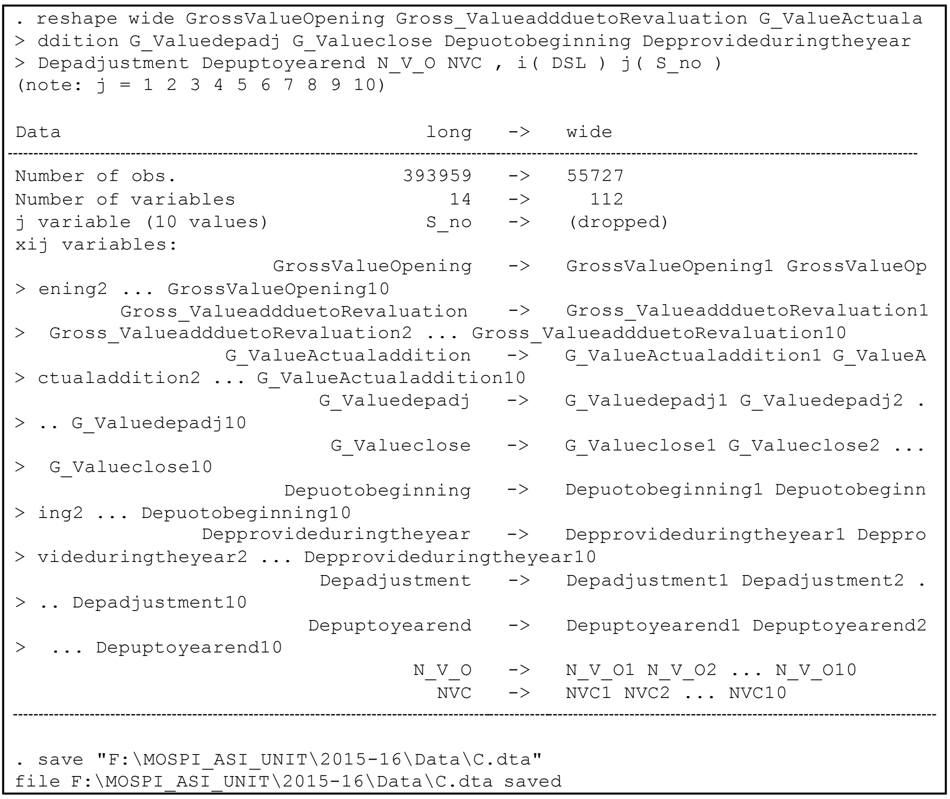
73 (78 / 220)   166.96%

C:\Users\Nithyashree M L\Documents

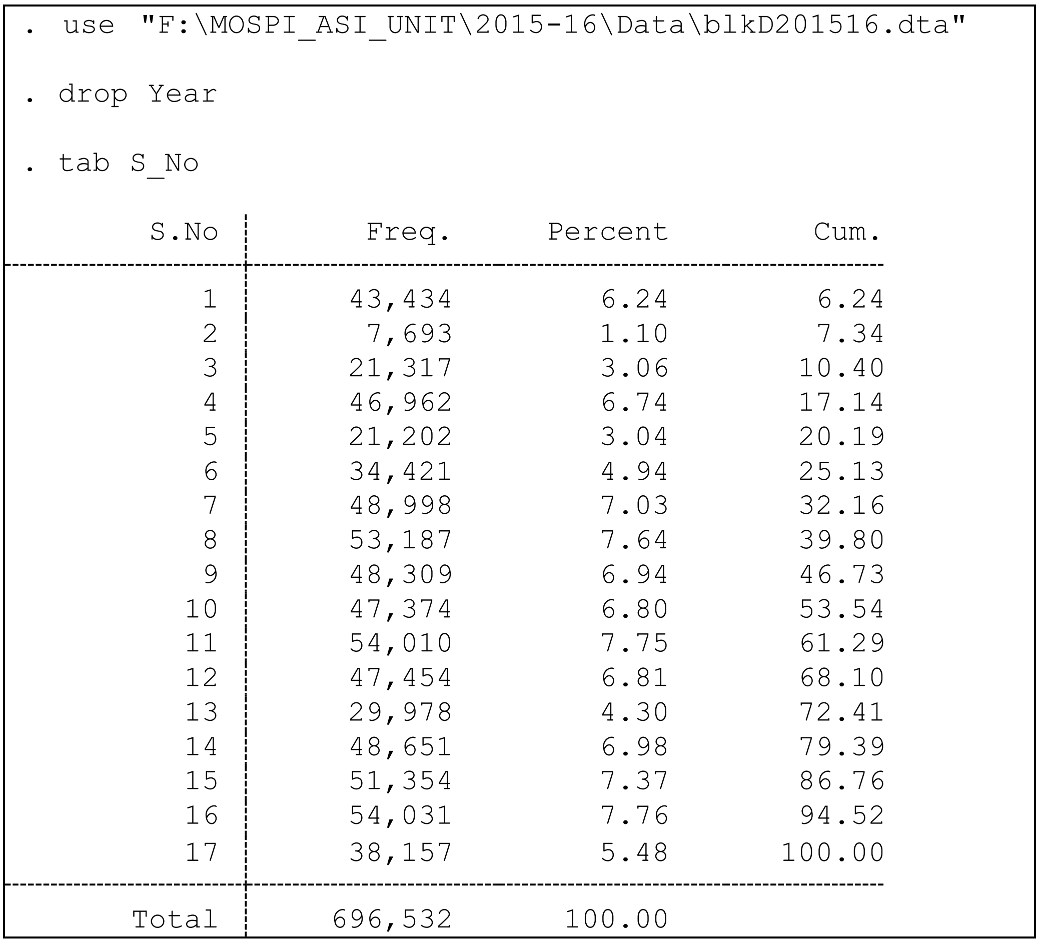
The detailed syntax for the ASI survey is given below keeping students in mind, which will be to practice reshaping and merging.



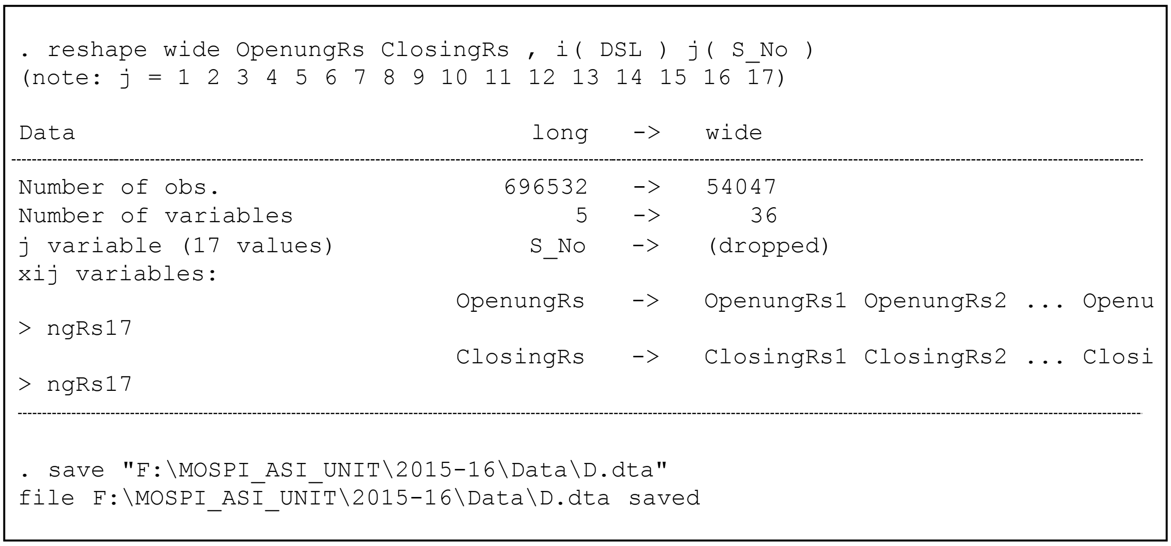
```
      name:  <unnamed>
       log:  F:\MOSPI_ASI_UNIT\2015-16\Data\M_1516.smcl
  log type:  smcl
 opened on:  21 Oct 2019, 12:54:11

. save "F:\MOSPI_ASI_UNIT\2015-16\Data\B.dta"
file F:\MOSPI_ASI_UNIT\2015-16\Data\B.dta saved

. use "F:\MOSPI_ASI_UNIT\2015-16\Data\blkC201516.dta"

. drop Year

. tab S_no

       S.no |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     36,930        9.37        9.37
          2 |     46,763       11.87       21.24
          3 |     52,954       13.44       34.69
          4 |     42,085       10.68       45.37
          5 |     40,209       10.21       55.57
          6 |      3,532        0.90       56.47
          7 |     51,758       13.14       69.61
          8 |     54,993       13.96       83.57
          9 |      9,008        2.29       85.85
         10 |     55,727       14.15      100.00
------------+-----------------------------------
      Total |    393,959      100.00
```



```
. reshape wide GrossValueOpening Gross_ValueaddduetoRevaluation G_ValueActuala
> ddition G_Valuedepadj G_Valueclose Depuotobeginning Depprovideduringtheyear
> Depadjustment Depuptoyearend N_V_O NVC , i( DSL ) j( S_no )
(note: j = 1 2 3 4 5 6 7 8 9 10)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                   393959   ->   55727
Number of variables                  14   ->     112
j variable (10 values)             S_no   ->   (dropped)
xij variables:
                      GrossValueOpening   ->   GrossValueOpening1 GrossValueOp
> ening2 ... GrossValueOpening10
         Gross_ValueaddduetoRevaluation   ->   Gross_ValueaddduetoRevaluation1
>  Gross_ValueaddduetoRevaluation2 ... Gross_ValueaddduetoRevaluation10
                  G_ValueActualaddition   ->   G_ValueActualaddition1 G_ValueA
> ctualaddition2 ... G_ValueActualaddition10
                          G_Valuedepadj   ->   G_Valuedepadj1 G_Valuedepadj2 .
> .. G_Valuedepadj10
                           G_Valueclose   ->   G_Valueclose1 G_Valueclose2 ...
>  G_Valueclose10
                       Depuotobeginning   ->   Depuotobeginning1 Depuotobeginn
> ing2 ... Depuotobeginning10
                Depprovideduringtheyear   ->   Depprovideduringtheyear1 Deppro
> videduringtheyear2 ... Depprovideduringtheyear10
                          Depadjustment   ->   Depadjustment1 Depadjustment2 .
> .. Depadjustment10
                         Depuptoyearend   ->   Depuptoyearend1 Depuptoyearend2
>  ... Depuptoyearend10
                                  N_V_O   ->   N_V_O1 N_V_O2 ... N_V_O10
                                    NVC   ->   NVC1 NVC2 ... NVC10
-----------------------------------------------------------------------------

. save "F:\MOSPI_ASI_UNIT\2015-16\Data\C.dta"
file F:\MOSPI_ASI_UNIT\2015-16\Data\C.dta saved
```



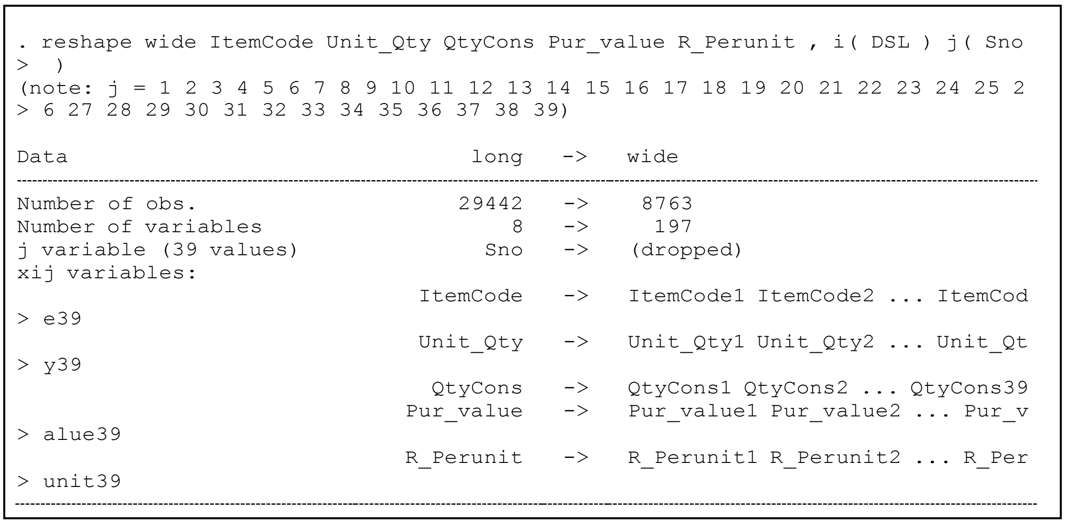
```
. use "F:\MOSPI_ASI_UNIT\2015-16\Data\blkD201516.dta"

. drop Year

. tab S_No

       S.No |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     43,434        6.24        6.24
          2 |      7,693        1.10        7.34
          3 |     21,317        3.06       10.40
          4 |     46,962        6.74       17.14
          5 |     21,202        3.04       20.19
          6 |     34,421        4.94       25.13
          7 |     48,998        7.03       32.16
          8 |     53,187        7.64       39.80
          9 |     48,309        6.94       46.73
         10 |     47,374        6.80       53.54
         11 |     54,010        7.75       61.29
         12 |     47,454        6.81       68.10
         13 |     29,978        4.30       72.41
         14 |     48,651        6.98       79.39
         15 |     51,354        7.37       86.76
         16 |     54,031        7.76       94.52
         17 |     38,157        5.48      100.00
------------+-----------------------------------
      Total |    696,532      100.00
```



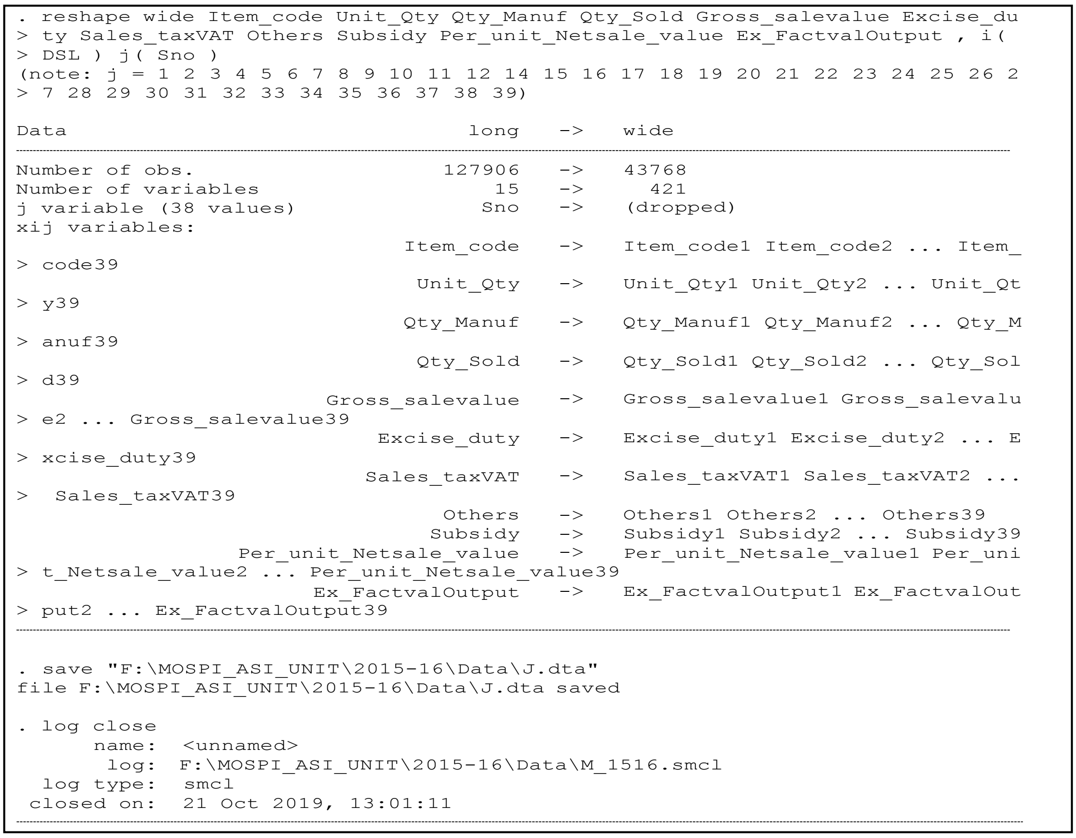
```
. reshape wide OpenungRs ClosingRs , i( DSL ) j( S_No )
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                   696532   ->   54047
Number of variables                   5   ->      36
j variable (17 values)             S_No   ->   (dropped)
xij variables:
                              OpenungRs   ->   OpenungRs1 OpenungRs2 ... Openu
> ngRs17
                              ClosingRs   ->   ClosingRs1 ClosingRs2 ... Closi
> ngRs17
-----------------------------------------------------------------------------

. save "F:\MOSPI_ASI_UNIT\2015-16\Data\D.dta"
file F:\MOSPI_ASI_UNIT\2015-16\Data\D.dta saved
```

```
. use "F:\MOSPI_ASI_UNIT\2015-16\Data\blkE201516.dta"

. br

. drop Year

. reshape wide MandaysWorkedManuf MandaysWorkedNonManuf MandaysWorkedTotal Ave
> NumberPersonwork NoofMandayspaid WagessalariesRs , i( DSL ) j( S_No )
(note: j = 1 2 3 4 5 6 7 8 9)

Data                                long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                    338475   ->   54346
Number of variables                    9   ->      56
j variable (9 values)               S_No   ->   (dropped)
xij variables:
                      MandaysWorkedManuf   ->   MandaysWorkedManuf1 MandaysWork
> edManuf2 ... MandaysWorkedManuf9
                   MandaysWorkedNonManuf   ->   MandaysWorkedNonManuf1 MandaysW
> orkedNonManuf2 ... MandaysWorkedNonManuf9
                      MandaysWorkedTotal   ->   MandaysWorkedTotal1 MandaysWork
> edTotal2 ... MandaysWorkedTotal9
                     AveNumberPersonwork   ->   AveNumberPersonwork1 AveNumberP
> ersonwork2 ... AveNumberPersonwork9
                         NoofMandayspaid   ->   NoofMandayspaid1 NoofMandayspai
> d2 ... NoofMandayspaid9
                         WagessalariesRs   ->   WagessalariesRs1 WagessalariesR
> s2 ... WagessalariesRs9
-----------------------------------------------------------------------------
```

```
. reshape wide ItemCode Unit_Quantity_code QtyCons Purchase_Value Rate_PerUnit
>  , i( DSL ) j( Sno )
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 2
> 6 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
>  52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76
> 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101
>  102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
>  121 122 123 124 125 126 127 128 129 130 131 132 133)

Data                                long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                    541009   ->   53406
Number of variables                    9   ->     668
j variable (133 values)              Sno   ->   (dropped)
xij variables:
                                ItemCode   ->   ItemCode1 ItemCode2 ... ItemCod
> e133
                      Unit_Quantity_code   ->   Unit_Quantity_code1 Unit_Quanti
> ty_code2 ... Unit_Quantity_code133
                                 QtyCons   ->   QtyCons1 QtyCons2 ... QtyCons13
> 3
                          Purchase_Value   ->   Purchase_Value1 Purchase_Value2
>  ... Purchase_Value133
                            Rate_PerUnit   ->Rate_PerUnit1 Rate_PerUnit2 ...
>  Rate_PerUnit133
-----------------------------------------------------------------------------
```



```
. reshape wide ItemCode Unit_Qty QtyCons Pur_value R_Perunit , i( DSL ) j( Sno
>  )
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 2
> 6 27 28 29 30 31 32 33 34 35 36 37 38 39)

Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                        29442   ->    8763
Number of variables                       8   ->     197
j variable (39 values)                  Sno   ->   (dropped)
xij variables:
                                   ItemCode   ->   ItemCode1 ItemCode2 ... ItemCod
> e39
                                   Unit_Qty   ->   Unit_Qty1 Unit_Qty2 ... Unit_Qt
> y39
                                    QtyCons   ->   QtyCons1 QtyCons2 ... QtyCons39
                                  Pur_value   ->   Pur_value1 Pur_value2 ... Pur_v
> alue39
                                  R_Perunit   ->   R_Perunit1 R_Perunit2 ... R_Per
> unit39
-----------------------------------------------------------------------------
```



```
. reshape wide Item_code Unit_Qty Qty_Manuf Qty_Sold Gross_salevalue Excise_du
> ty Sales_taxVAT Others Subsidy Per_unit_Netsale_value Ex_FactvalOutput , i(
> DSL ) j( Sno )
(note: j = 1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 18 19 20 21 22 23 24 25 26 2
> 7 28 29 30 31 32 33 34 35 36 37 38 39)

Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       127906   ->   43768
Number of variables                      15   ->     421
j variable (38 values)                  Sno   ->   (dropped)
xij variables:
                                  Item_code   ->   Item_code1 Item_code2 ... Item_
> code39
                                   Unit_Qty   ->   Unit_Qty1 Unit_Qty2 ... Unit_Qt
> y39
                                  Qty_Manuf   ->   Qty_Manuf1 Qty_Manuf2 ... Qty_M
> anuf39
                                   Qty_Sold   ->   Qty_Sold1 Qty_Sold2 ... Qty_Sol
> d39
                            Gross_salevalue   ->   Gross_salevalue1 Gross_salevalu
> e2 ... Gross_salevalue39
                                 Excise_duty  ->   Excise_duty1 Excise_duty2 ... E
> xcise_duty39
                                Sales_taxVAT  ->   Sales_taxVAT1 Sales_taxVAT2 ...
>  Sales_taxVAT39
                                     Others   ->   Others1 Others2 ... Others39
                                    Subsidy   ->   Subsidy1 Subsidy2 ... Subsidy39
                     Per_unit_Netsale_value   ->   Per_unit_Netsale_value1 Per_uni
> t_Netsale_value2 ... Per_unit_Netsale_value39
                           Ex_FactvalOutput   ->   Ex_FactvalOutput1 Ex_FactvalOut
> put2 ... Ex_FactvalOutput39
-----------------------------------------------------------------------------

. save "F:\MOSPI_ASI_UNIT\2015-16\Data\J.dta"
file F:\MOSPI_ASI_UNIT\2015-16\Data\J.dta saved

. log close
      name:  <unnamed>
       log:  F:\MOSPI_ASI_UNIT\2015-16\Data\M_1516.smcl
  log type:  smcl
 closed on:  21 Oct 2019, 13:01:11
-------------------------------------------------------------------------------
```

References

Annual Survey of Industries. (2015-16). *Summary results*, Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

National Sample Survey Organization. (2014). *Key Indicators of Situation of Agricultural Households in India*, January – December 2013. NSS 70th Round. Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

Sampling Techniques for Program Evaluation

Girish K. Jha
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: girish.stat@gmail.com

Introduction

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

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

Concepts and Definitions

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

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

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

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

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

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

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

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

Various Sampling Methods

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

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

Simple Random Sampling

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

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

(i) Lottery Method,
(ii) Use of Random Number Tables.

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

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

(a) Tippett's random number Tables
(b) Fisher and Yates Tables
(c) A million random digits Table.

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

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

Stratified Random Sampling

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

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

Cluster Sampling

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

Multi-Stage Sampling

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

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

Systematic Sampling

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

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

For example, suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts.

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate then simple random sampling. Another drawback of systematic sampling is that it is not possible to get an unbiased estimation of the variance of the estimator.

Varying Probability Sampling

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

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

Planning and Execution of Sample Surveys

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

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

- Field work

The different aspects listed above are inter-dependent.

(i) Formulation of data requirements:

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

(ii) Survey: Ad-hoc or repetitive:

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

(iii) Method of collecting primary data:

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

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

(iv) Questionnaire vs. schedule:

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

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

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

(v) Survey, reference and reporting periods:

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

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

(vi) Choice of sampling design:

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

(vii) Pilot surveys:

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

- Testing out provisional schedules and related instructions
- Evolving suitable procedure for field and tabulation work, and
- Training field and tabulation staff
- Potential sources of measurement error
- Likely non-response rate

- Sensitive issues or sources of ambiguity
- Difficulties of access to chosen sample members

(viii) Field work:

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

Determination of sample size

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

References

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

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

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

An Introduction to Regression Analysis

Asha Devi S.S. and Praveen K.V.

Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012

Email: ash.nibha@gmail.com

In social science, it is interesting to study how two variables are related so that we can predict the change in one variable from another variable (s). In many cases, the relationship between two variables is inexact. For example, the yield of crops and temperature. There are many factors that influence the yield, and it may be due to the variety, location, soil characteristics, inputs, irrigation and other such factors or can be due to measurement errors. So, there is no unique relationship between crop yield and temperature. But we can find an average yield reduction for a temperature rise beyond the optimal level. This average observed yield for a given level of observed temperature is called a regression curve of yield on temperature. The curve is straight; it is called linear regression; otherwise, it is non-linear.

Therefore, regression is the linear or non-linear relationship between two or more random variables. "*Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter*" (Gujarati,2004). It means simply that regression is a statistical measure to determine the average relationship between two or more variables.

The functional relationship between two variables can be expressed in the form of an equation as *Y=f(X).* Here Y denotes the dependent variable (or can be called as response/Outcome/Predictand/Regressand/Explained Variable), and X indicates the independent variable (also known as Predictor/Regressor/Covariates/control variables). Thus, the above equation indicates the value of Y for a particular value of X. But we cannot observe exact mathematical relationships most of the time in social science research due to sampling. In those situations, the functional relationship between two variables can be represented in the form of a statistical model as follows;

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

where ε indicates the error term, which is a proxy for all omitted variables which are not included in the model.

Let us consider a basic linear regression model with one explanatory variable.

$$Yi = \beta 0 + \beta 1Xi + \varepsilon i$$

Where Yi is the value of the response variable for the i th individual and Xi is the observation of independent variable for the ith individual. β0 and β1 are parameters to be estimated called as regression co-efficients. β0 is the intercept of the model, which means the mean probability distribution of Y when X=0. β1 is the slope of the regression line which indicates the mean probability distribution of y per unit increase in x. εi is a random error term with mean zero and variance σ^{2}.

The aim of any researcher would be to estimate β0 and β1 co-efficients accurately so that the estimated and observed values of Yi would be as close as possible. For this, we use the 'Principle of Least Squares'. We find β0 and β1 that minimises the sum of squared residuals which is the difference between observed and estimated Y.

$$S = \sum_{i=1}^{n}(Y_i - Yi^\wedge)\,2$$

$$S = \sum_{i=1}^{n}(Y_i - \beta 0^\wedge - \beta 1^\wedge Xi)\,2$$

To minimize S, we differentiate S with respect to each parameter and equate to zero. We get as many equations as the number of parameters. By solving these equations simultaneously, we get the estimates of parameters.

Regression Diagnosis

Once we estimate the co-efficients we have to verify if we have not violated the assumptions or if there are any departures.

1. Linearity

Linear regression assumes that the model is linear in both parameters and variables. If the actual model is non-linear, and if we use a linear model for prediction, the prediction may go wrong, and it is inappropriate to use such a model. The linearity assumption can be tested by plotting the graph between the residuals against the fitted values.

2. Normality of errors

This assumption means that the residuals should be normally distributed. Normality is necessary only for hypothesis tests to be valid, estimation of the coefficients only requires that the errors be identically and independently distributed. Normality is a problem when we analyse data from a small sample. This can be checked by plotting the histogram of residuals. Also, there are different tests to test the distribution of residuals.

3. Homoscedasticity

The variables around the regression line is the same for all values of the predictor variable. Visualization by plotting residuals against the predictor variable or against the fitted values is helpful to study if the variance of the error terms is constant. However, many statistical tests are available in the literature for testing constancy of error variance, such as White's test, Breusch-Pagan test, etc.

4. Independence of error terms

The errors associated with one observation are not correlated with the errors of any other observation. This correlation between the errors are observed in the case of time series data and it can be tested using Durbin Watson test.

5. Model specification

The model should be properly specified including all relevant variables, and excluding irrelevant variables to avoid bias.

6. Presence of outlier observations

Although it is not an assumption of regression, it is a good practice to check for outliers or extreme values as it can affect the standard error and hence lead to biased confidence intervals. This can be tested by using cook's distance measure.

References

Gujrati, D.N. (2003). Basic Econometrics, Mc Graw Hill. New York.

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

Achal Lama[1], Girish K. Jha[2] and Asha Devi[2]
[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012*
[2]*ICAR-Indian Agricultural Research Institute, New Delhi-110012*
Email: girish.stat@gmail.com

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

Time Series Methods

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

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

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

1. Linear time series models

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

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

1.1 Autoregressive (AR) model

A stochastic model that can be extremely useful in the representation of certain practically occurring series is the autoregressive model. In this model, current value of the process is expressed as a finite, linear aggregate of previous values of the process and a shock ε_t. Let us denote the values of a process at equally spaced time epochs $t, t-1, t-2, \ldots$ by $y_t, y_{t-1}, y_{t-2}, \ldots$ then y_t can be described as

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

If we define an autoregressive operator of order *p* by

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

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

1.2 Moving Average (MA) model

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

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

If we define a moving average operator of order *q* by

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

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

1.3 Autoregressive Moving Average (ARMA) model

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

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

or

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

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

1.4 Autoregressive Integrated Moving Average (ARIMA) model

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

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

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

2. Non-linear Models: ARCH and GARCH Models

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

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

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

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

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

$$\varepsilon_t = \xi_t h_t^{1/2} \qquad (1)$$

where ξ_t ~ N (0,1). A sufficient condition for the conditional variance to be positive is

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

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

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

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

Model Building

Step 1: Determine whether the time series is stationary

The series being analysed must be stationary. A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time. Thus, if we consider different subsets of a realization (TS 'sample') the different subsets will typically have means, variances and autocorrelation functions that do not differ significantly which means that stationary time series has the property that its statistical properties such as the mean and variance are constant over time. The presence of stationarity in the data can be obtained by simply plotting the raw data or by plotting the autocorrelation and partial autocorrelation function. Statistical tests like Dickey- Fuller test, augmented Dickey-Fuller test, KPSS (Kwiatkowski, Phillips, Schmidt, and Shin) test, Philips-Perron test are also available to test the stationarity.

Step 2: Identify the model

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

Step 3: Estimate the model parameters and diagnostic checking

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

The Ljung-Box statistic is given by:

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

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

Step 4: Select the most suitable ARIMA model

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

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

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

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

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

Step 5: Determination of residuals and heteroscedasticity test

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

Step 6: Residuals and diagnostic checking

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

Step 7: Estimation of parameters

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

5. Illustration

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

5.1 Fitting of the cotlook A index

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

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

$$y_t = 141.9264 \; –1.3905 \; y_{t-1} + 0.4538\, y_{t-2} + \varepsilon_t$$
$$(3.94) \quad (0.05) \qquad\quad (0.05)$$

where

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

and h_t satisfies the variance equation

$$h_t = 8.470 + 0.208\ \varepsilon_{t-1}^{2} + 0.215\ h_{t-1}$$
$$(1.97) \qquad (0.09) \qquad (0.079)$$

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

Table 1. Forecast of the cotlook A index series

Month	**Actual Value**	**Forecast Arima (1,1,0)**	**Forecast AR(2)-Garch (1,1)**
Feb-11	469.98	408.34(8.30)	389.59(26.46)
Mar-11	506.34	416.47(15.56)	371.55(25.74)
Apr-11	477.56	421.40(22.35)	348.54(25.05)
May-11	364.91	424.53(28.55)	324.69(24.39)
Jun-11	317.75	426.66(34.17)	301.98(23.75)
Jul-11	268.96	428.23(39.29)	281.25(23.13)
Aug-11	251.55	429.49(43.97)	262.76(22.54)
Sep-11	257.63	430.57(48.29)	246.50(21.97)
Oct-11	243.85	431.55(52.30)	232.32(21.42)
Nov-11	230.78	432.48(56.05)	220.01(20.90)

Dec-11	210.43	433.37(59.58)	209.35(20.39)
Jan-12	222.91	434.25(54.45)	200.15(19.91)
Feb-12	222.12	435.12(57.13)	192.21(19.44)
Mar-12	219.36	435.99(59.68)	185.37(19.01)

Table 2. Forecast evaluation of the cotlook A index series

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

6. R Code for Analysing a Time Series Data

```
library(“tseries”)

library(“forecast”)

library(“fgarch”)

setwd("C:/Users/ACHAL/Desktop") # Setting of the work directory

data<-read.table("data.txt") # Importing data

datats<-ts(data,frequency=12,start=c(1982,4)) # Converting data set into time series

plot.ts(datats) # Plot of the data set

adf.test(datats) # Test for stationarity

diffdatats<-diff(datats,differences=1) # Differencing the series

datatsacf<-acf(datats,lag.max=12) # Obtaining the ACF plot

datapacf<-pacf(datats,lag.max=12) # Obtaining the PACF plot

auto.arima(diffdatats) # Finding the order of ARIMA model

datatsarima<-arima(diffdatats,order=c(1,0,1),include.mean=TRUE) # Fitting of ARIMA model

forearimadatats<-forecast.Arima(datatsarima,h=12) # Forecasting using ARIMA model

plot.forecast(forearimadatats) # Plot of the forecast
```

```
residualarima<-resid(datatsarima) # Obtaining residuals

archTest(residualarima,lag=12) # Test for heteroscedascity

# Fitting of AR-GARCH model

garchdatats<-garchFit(formula = ~ arma(2)+garch(1, 1), data = datats, cond.dist = c("norm"),
include.mean = TRUE, include.delta = NULL, include.skew = NULL, include.shape = NULL,
leverage = NULL, trace = TRUE,algorithm = c("nlminb"))

# Forecasting using AR-GARCH model

forecastgarch<-predict(garchdatats, n.ahead = 12, trace = FALSE, mse = c("uncond"),
plot=FALSE, nx=NULL, crit_val=NULL, conf=NULL)

plot.ts(forecastgarch) # Plot of the forecast
```

References

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

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

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

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

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

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

Long Memory Process

Chiranjit Majumder
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: majumder.chira@gmail.com

Introduction

Time-series analysis is a crucial statistical method that serves as the foundation for human and automatic planning across a wide range of fields of study (Gooijer and Hyndman, 2006). Prediction or forecasting is a difficult field of study. Time-series methods are commonly used for price forecasting. Multiple time-series forecasting models have been developed and refined during the past few decades. Most studies in time-series analysis operate on the assumption that observations with large temporal gaps between them are completely unrelated. However, it is evident that distant observations are dependent in many agricultural data, especially daily commodity price data, meaning that the data set has the defining trait of long memory or long-range dependence. In statistics, long range dependency, sometimes known as "long memory of a time series," is a phenomena in which statistical reliance holds between observations that are widely spaced apart in time. In contrast to the exponential decay seen in the standard Autoregressive integrated moving average (ARIMA) model, a time series process is said to have a long memory if its autocorrelation function decays extremely slowly towards zero. The autocorrelation function of a process with a lengthy memory has a persistency structure that is not typical of either an I(1) or an I(0) process. By permitting a non-integer or fractional differencing value, the Autoregressive fractionally integrated moving average (ARFIMA) model may be used to represent time-series processes with extended memory in the mean equation (Granger and Joyeux, 1980). The ARFIMA model has been used to predict the cost of agricultural commodities by Paul (2014) and Paul et al. (2014a, 2015). The model has proven its worth in terms of both explaining and predicting variability.

Long Memory

Studies in time-series analysis often operate on the assumption that observations spaced far apart in time are unrelated to one another. Nonetheless, numerous real-world scenarios reveal that many empirical economic series demonstrate that the distant observations are dependent, although in modest but not insignificant ways. The statistical dependence of any time-series data is generally measured by plotting the ACF of the dataset. Let $X_t;\ (t = 1,2, \ldots)$ be a stationary time-series process and the autocorrelation function of the time-series with a time lag of k is given as

$$\rho_k = \text{cov}\ (x_t, x_{t-1})/\text{var}\ (x_t)$$

The series X; $(t = 1,2, \dots)$ is said to have short memory if the autocorrelation coefficient at lag k approaches to zero as k tends to infinity, i.e., $\lim_{k\to\infty} \rho_k = 0$. The autocorrelation functions of most of stationary and invertible (ARMA) time-series process decay very rapidly at an exponential rate, so that $\rho_k \approx |m|k$, where $|m| < 1$.

For long memory processes, decaying of autocorrelations functions occur at much slower rate (hyperbolic rate) which is consistent with $\rho_k \approx Ck^{2d-1}$, as k increases indefinitely, where C is a constant and d is the long memory parameter. Long memory processes have autocorrelation functions that show persistency structure inconsistent with either an I(1) or an I(0) process.

In other words, a short memory process is defined as

$$\sum_{k=0}^{\infty} |\rho_k| < \infty$$

And a long memory process is defined as

$$\sum_{k=0}^{\infty} |\rho_k| = \infty$$

where ρ_k is the coefficient of autocorrelation with lag of k.

In frequency domain a long memory is defined in terms of rates of explosion of low frequency spectra as

$$f_x(\omega) = \mathrm{g}\omega^{-2\,\mathrm{d}} \text{ as } \omega \to 0^+$$

In general, the low-frequency spectral definition of long memory is simply as

$$f_x(\omega) = \infty \text{ as } \omega \to 0^+$$

Berran (1995a) have discussed some properties of a stationary long memory process.

Arfima Model

For modelling time series with a long memory, the Autoregressive Fractionally Integrated Moving-Average (ARFIMA) model (Granger and Joyeux, 1980) is utilized. Fractional integration is a generalization of integer integration. Here, time-series is usually presumed to be integrated of order zero or one. For example, an autoregressive moving-average process integrated of order d [denoted ARFIMA (p, d, q)] can be represented as

$$\phi(B)y_t == (1 - B)^{-d}\theta(B)u_t$$

where, μ_t is an independently and identically distributed (i.i.d.) random variable having zero mean and constant variance, B denotes the lag operator; $\phi(B)$ and $\theta(B)$ denote finite AR and MA polynomials in the lag operator having roots outside the unit circle. For $d = 0$, the process is stationary, for $-0.5 < d < 0.5$ the process y_t is stationary and invertible, for $d \in (0, \frac{1}{2})$ the process is said to have long memory. For any value of d we have

$$(1 - B)^d = 1 - dB + \frac{B^2 d(d-1)}{2} + \cdots + \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j B^j$$

with binomial coefficients

$$\binom{d}{j} = \frac{d!}{j!\,(d-j)!} = \frac{r(d+1)}{r(j+1)\Gamma(d-j+1)}$$

where Γ (.) represents the gamma function.

Testing of Long Memory

Following approaches are generally used for estimating long memory parameter

(i) Parametric method - Maximum Likelihood method of estimation (MLE)(Berran, 1995b).

(ii) Semi-parametric method - Whittle (Robinson, 1994), GPH (Geweke and Porter-Hudak, 1983) etc.

(iii) Heuristic method - R/S statistic (Hurst, 1951), ACF plot, variance plot, log-log correlogram and least square regression in spectral domain.

In dealing with the long memory processes, the first and foremost step is to have an idea for the possible presence of long memory behavior in time series. Long memory can be measured by calculating the so-called Hurst exponent. A stationary stochastic process $\{y_t\}$ is called a long memory process if there exists a real number H and a finite constant C such that the autocorrelation function York $\rho(r)$ has the following rate of decay (Alptekin, N.2007).

$\rho(k) = C_c^{2H-2}$ as $r \to \infty$

The parameter H, Hurst Exponent, display the long memory property of the time series. The Hurst exponent takes values from 0 to $1(0 \leq H \leq 1)$. If $0.5 < H < 1$, the series indicates persistent behavior or long memory.

Rescaled range statistic (R/S)

The first test for long memory was used by the hydrologist Hurst (1951) for the design of an optimal reservoir for the Nile River, of where flow regimes were persistent. Hurst gave the following formula:

$$(R/S)_n = cn^H$$

R/S is the rescaled range statistic measured over a time index n, c is a constant and H the Hurst exponent. The aim of the R/S statistic is to estimate the Hurst exponent which can characterize a series. Estimation of Hurst exponent can be done by transforming to:

$$\log (R/S)_n = \log (c) + H\log (n)$$

Rescaled range statistic (R/S) is defined as the range of partial sums of deviation of a time series from its mean, rescaled by its standard deviation. Consider the sample $\{x_1, x_2, \ldots \ldots \ldots, x_n\}$ from a stationary long memory process $\{ x_n;\ t = 1,2, \ldots . N\}$, and let the partial sums of x_k is $\sum_{j=1}^{k} x_j$; k = 1,2, … , n

Let $\bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$, and $S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$ be the sample mean and sample variance respectively. The Rescaled range Statistic(R/S) is defined as:(Magsood,A. and Aqil ,S.M.2014) :

$$\text{R/S} = \frac{1}{S_n}[\text{Max}_{1\le k\le n}\ (\textstyle\sum_{j=1}^{k}(x_j - \bar{x}_n)) - \text{Min}_{1\le k\le n}\ (\textstyle\sum_{j=1}^{k}(x_j - \bar{x}_n))]$$

Where,

n: number of observations

S_n : the standard derivation

In addition, the Hurst coefficient H can be used to estimate the fractional differencing parameter d by the equation: $d = H - 0.5$

This method is considered as semi parametric of estimating ARFIMA models.

Aggregated variance method

This method divides the original time series $X = \{Xi, i \geq 1\}$ into blocks of size m and average within each block, that is, consider the aggregated series:

$$\text{X}^{(\text{m})}(\text{k}) = \frac{1}{\text{m}}\sum_{\text{i}=(\text{k}-1)\text{m}+1}^{\text{km}} \text{X}(\text{i})\ \ \text{k} = 1,2, \ldots \text{q}$$

Then, divide the data, $X_1, \ldots, X_N$, into N/m blocks of size m, and calculate its sample variance(Sun ,R., Chen, Y. and Li, Q.2007).

$$\text{Var}X^{(m)} = \frac{1}{N/m}\sum_{k=1}^{N/m}(X^{(m)}(k))^2 - (\frac{1}{N/m}\sum_{k=1}^{N/m}X^{(m)}(k))^2$$

$\mathrm{VarX}^{(m)} \sim \sigma^2 m^{\beta}$, as $m \to \infty$

Where σ is the scale parameter and $\beta = 2H - 2 < 0$, the sample variance $\hat{\mathrm{V}}\mathrm{arX}^{(\mathrm{m})}$ should be asymptotically proportional to m^{2H-2} for large N/m and m, the resulting points should form a straight line with slope $\beta = 2H - 2, -1 \leq \beta < 0$

Semi-parametric estimation methods

Geweke and PorterHudak (1983) introduced the most common and commonly used semi-parametric estimating approach, which is based on an estimated regression equation generated from the logarithm of the spectral density function (Joe and Sisir, 2014). The GPH estimation procedure is a two-step procedure which begins with the estimation of d (Paul, 2014) and is based on the following log-periodogram regression (Bhardwaj and Swanson, 2006).

$$\ln\left[I(\omega_j)\right] = \beta_0 + \beta_1 \ln\left[4\sin^2\left(\omega_j^2\right)\right] + v_j$$

where

$$\omega_j = \frac{2\pi j}{T}, J = 1,2,\dots,m$$

The estimate of d, say $\hat{d}_{\mathrm{GPH}}$; is $\hat{\beta}_1$, ω_j represents the $m = \sqrt{T}$ Fourier frequencies, and $I(\omega_j)$ denotes the sample periodogram defined as:

$$I(\omega_j) = \frac{1}{2\pi T} \left|\sum_{t=1}^{T} y_t e^{-\omega j t}\right|^2$$

The second step of the GPH estimation procedure involves fitting an ARMA model to data according to box and jenkins method , given the estimate of d. the GPH estimate is denoted as $\hat{d}_{GPH}$. Reisen and Lopes (1999) modified the GPH procedure by replacing the periodogram by a "smoothed" estimate of the spectral density.

$f_m(x) = \frac{1}{2\pi} \sum_{s=-m}^{m} k\left(\frac{s}{m}\right) \hat{p}(s) \cos(sx)$ Henceforth, the smoothed periodogram estimate of d is denote as $\hat{d}_{\mathrm{Sperio}}$. Both $\hat{d}_{\mathrm{GPH}}$ and $\hat{d}_{\mathrm{Sperio}}$ although simpler to implement are inefficient in the non-stationary region i.e., $|d| > 1 = 2$.

Another semi-parametric estimator, the local Whittle estimator, is also often used to estimate d. This estimator was proposed by Kuensch (1987) and modified by Robinson (1995). The local Whittle estimator of d; say $\hat{d}_w$ is obtained by maximizing the local

Whittle log likelihood at Fourier frequencies close to zero, given by (Bhardwaj and Swanson, 2006):

$$\Gamma(d) = -\frac{1}{2\pi m}\sum_{j=1}^{m}\frac{I(\omega_j)}{f(\omega_j; d)} - \frac{1}{2\pi m}\sum_{j=1}^{m} f(\omega_j; d)$$

Parametric Estimation Methods

The Parametric methods estimate all parameters of the ARFIMA process in one step. Exact Maximum Likelihood (EML), proposed by Sowell (1992).

Let **y** be the sample time series. The log-likelihood of the estimation is simply, based on the normality assumption and with a procedure to compute the autocovariances in the T × T covariance matrix $\sum\sigma_s^2 R$ of a T × 1 vector of observations y, the log-likelihood for the ARFIMA (p, d, q) model (4) with k regressors is (Lildholdt, P. 2000; Doornik, J.A. and Ooms, M. 2004) :

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

$$\sigma_s^2 - \frac{1}{2}\ln|\mathrm{R}| - \frac{1}{2\sigma_s^2}\mathrm{Z}^{-1}\mathrm{Z}$$

where $Z = y - X\beta$

When σ_s^2 and β are concentrated out, the resulting normal profile likelihood function becomes:

$$\log L(d, \phi, \theta) = c - \frac{1}{2}\ln|\mathrm{R}| - \frac{T}{2}\ln\,[\hat{\hat{Z}}^{2\,-1\,\hat{}}\,\mathrm{Z}]$$

Where $z\hat{} = y - X\beta\hat{}$

and

$$\beta = (\mathrm{X}^{-1}\mathrm{X}^{-1})'\mathrm{X}^{-1}\mathrm{Y}$$

Evaluation of the Forecasting Performance

This subsection presents the tools for comparing the forecasting performances of different models. In this paper, the forecasts from the estimated models are compared using the root mean squared error (RMSE), the mean absolute error (MAE) and the mean absolute percentage error (MAPE) are given below,

Where $\hat{y}_t \& y_t$ are the estimated and actual values respectively n is the number of data.

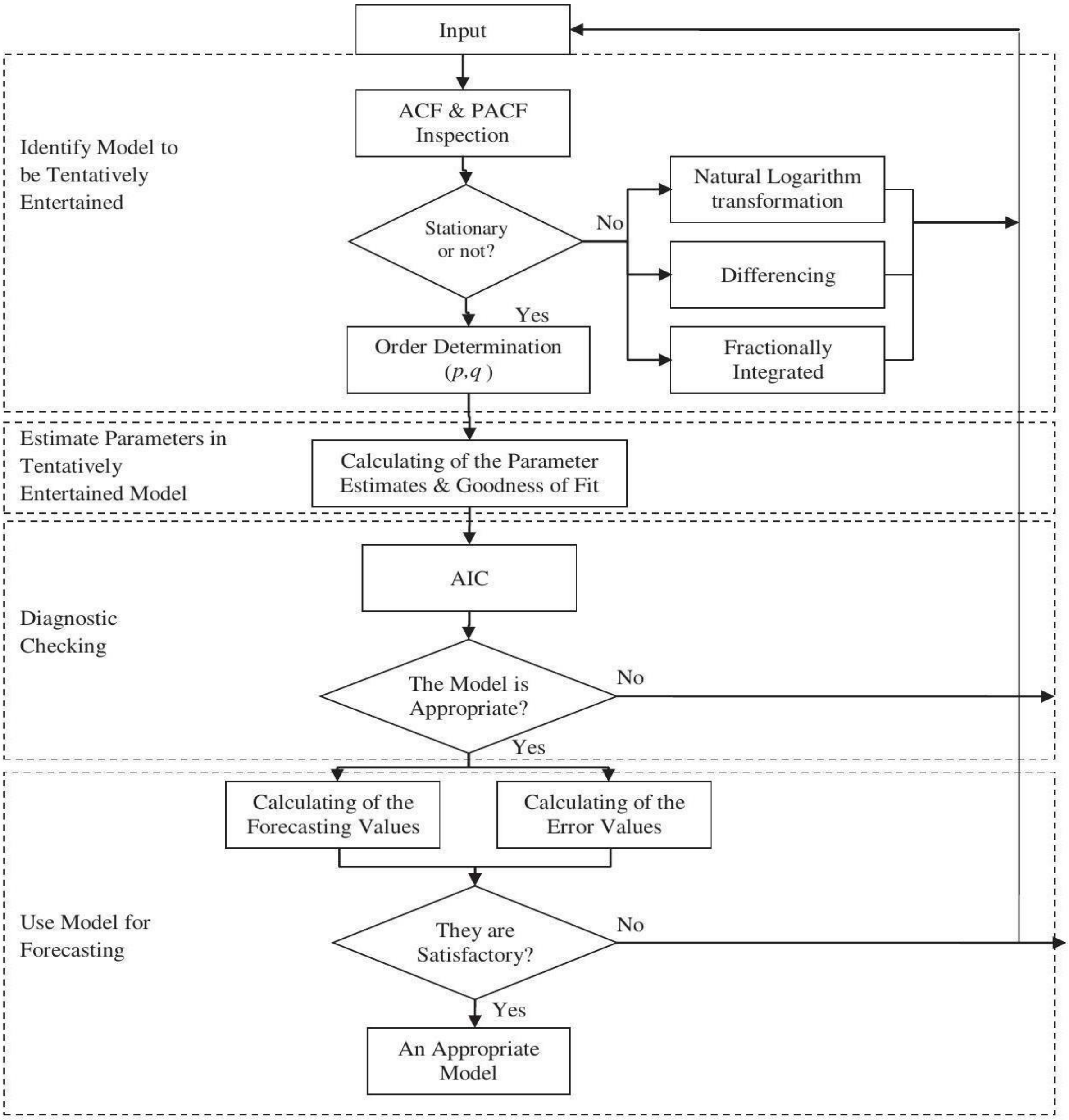

Input
Identify Model to be Tentatively Entertained
ACF & PACF Inspection
Stationary or not?
No
Natural Logarithm transformation
Differencing
Fractionally Integrated
Yes
Order Determination (p,q)
Estimate Parameters in Tentatively Entertained Model
Calculating of the Parameter Estimates & Goodness of Fit
Diagnostic Checking
AIC
The Model is Appropriate?
No
Yes
Use Model for Forecasting
Calculating of the Forecasting Values
Calculating of the Error Values
They are Satisfactory?
No
Yes
An Appropriate Model

Figure 1. Flo chart of the Box and Jenkins model-building strategy.

Application of Discriminant Analysis in Social Science Research

P. Venkatesh
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: venkatesh1998@gmail.com

Introduction

Discriminant Analysis (DA) is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature. DA undertakes the same task as multiple linear regressions by predicting an outcome. However, multiple linear regressions is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values. But many interesting variables are categorical, such as political party voting intention, migrant/non-migrant status, making a profit or not, holding a particular credit card, owning, renting or paying a mortgage for a house, employed/unemployed, satisfied versus dissatisfied employees, which customers are likely to buy a product or not buy, what distinguishes Stellar Bean clients from Gloria Beans clients, whether a person is a credit risk or not, etc.

Objectives

- Development of discriminant functions
- Examination of whether significant differences exist among the groups, in terms of the predictor variables.
- Determination of which predictor variables contribute to most of the intergroup differences
- Evaluation of the accuracy of classification

Discriminant Analysis Linear Equation

DA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D= v_1 X_1+ v_2 X_2+ v_3X_3+ \quad ... \quad + v_iX_i + a$$

Where D = discriminate function
v = the discriminant coefficient or weight for that variable
X = respondent’s score for that variable
a = a constant
i = the number of predictor variables

This function is similar to a regression equation or function. The v's are unstandardized discriminant coefficients analogous to the b's in the regression equation. These v's maximize the distance between the means of the criterion (dependent) variable. Standardized discriminant coefficients can also be used like beta weight in regression. Good predictors tend to have large weights. What you want this function to do is maximize the distance between the categories, i.e. come up with an equation that has strong discriminatory power between groups. After using an existing set of data to calculate the discriminant function and classify cases, any new cases can then be classified. The number of discriminant functions is one less the number of groups. There is only one function for the basic two group discriminant analysis.

Assumptions of Discriminant Analysis

- the observations are a random sample;
- each predictor variable is normally distributed;
- each of the allocations for the dependent categories in the initial classification are correctly classified;
- there must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group);
- each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division;
- for instance, three groups taking three available levels of amounts of housing loan;
- the groups or categories should be defined before collecting the data;
- the attribute(s) used to separate the groups should discriminate quite clearly between
- the groups so that group or category overlap is clearly non-existent or minimal;
- group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

Applications

- Agriculture extension research: Adoption behavior
- Market research: Market segmentation
- Financial research: Default behavior
- Human resources: High performers

Steps of Discriminant Analysis in SPSS

1. ***Analyse*** >> ***Classify*** >> ***Discriminant***

2. Select '*dependent variable*' as your grouping variable and enter it into the ***Grouping Variable Box***
3. Click ***Define Range*** button and enter the lowest and highest code for your groups (
4. Click ***Continue***.
5. Select your predictors (IV's) and enter into ***Independents*** box (Fig. 25.6) and select ***Enter Independents Together***. If you planned a stepwise analysis you would at this point select ***Use Stepwise Method*** and not the previous instruction.
6. Click on ***Statistics*** button and select ***Means, Univariate Anovas, Box's M, Unstandardized*** and ***Within-Groups Correlation***
7. ***Continue*** >> ***Classify***. Select ***Compute From Group Sizes, Summary Table, Leave One Out Classification, Within Groups***, and all ***Plots***
8. ***Continue*** >> ***Save*** and select ***Predicted Group Membership*** and ***Discriminant Scores***
9. ***OK***.

Interpretation

Group statistics tables

In discriminant analysis we are trying to predict a group membership, so firstly we examine whether there are any significant differences between groups on each of the independent variables using group means and ANOVA results data. The Group Statistics and Tests of Equality of Group Means tables provide this information. If there are no significant group differences it is not worthwhile proceeding any further with the analysis. A rough idea of variables that may be important can be obtained by inspecting the group means and standard deviations

Log determinants and Box's M tables

In ANOVA, an assumption is that the variances were equivalent for each group but in DA the basic assumption is that the variance-co-variance matrices are equivalent. Box's M tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent. The researcher wants this test not to be significant so that the null hypothesis that the groups do not differ can be retained. For this assumption to hold, the log determinants should be equal. When tested by Box's M, we are looking for a non-significant M to show similarity and lack of significant differences.

Table of eigenvalues

This provides information on each of the discriminate functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus 1. We are only using two groups here, namely 'smoke' and 'no smoke', so only one function is displayed. The canonical correlation is the multiple correlation between the predictors and the discriminant function. With only one function it provides an index of overall model fit which is interpreted as being the proportion of variance explained (R2).

Wilks' lambda

Wilks' lambda indicates the significance of the discriminant function. This table indicates a proportion of total variability not explained, i.e. it is the converse of the squared canonical correlation.

The standardized canonical discriminant function coefficients table

The interpretation of the discriminant coefficients (or weights) is like that in multiple regressions. This table provides an index of the importance of each predictor like the standardized regression coefficients (beta's) did in multiple regression. The sign indicates the direction of the relationship.

The structure matrix table

It provides another way of indicating the relative importance of the predictors and it can be seen below that the same pattern holds. Many researchers use the structure matrix correlations because they are considered more accurate than the Standardized Canonical Discriminant Function Coefficients. The structure matrix table shows the correlations of each variable with each discriminate function. These Pearson coefficients are structure coefficients or discriminant loadings. They serve like factor loadings in factor analysis. By identifying the largest loadings for each discriminate function the researcher gains insight into how to name each function. Generally, just like factor loadings, 0.30 is seen as the cut-off between important and less important variables.

The canonical discriminant function coefficient table

These unstandardized coefficients *(b)* are used to create the discriminant function (equation). It operates just like a regression equation. The discriminant function coefficients *b* or standardized form *beta* both indicate the partial contribution of each variable to the discriminate function controlling for all other variables in the equation. They can be used to assess each independent variables unique contribution to the discriminate function and therefore provide information on the relative importance of each variable. If there are any dummy variables, as in regression, individual beta weights cannot be used and dummy variables must be assessed as a group through hierarchical DA running the analysis, first without the dummy variables then with them. The difference in squared canonical correlation indicates the explanatory effect of the set of dummy variables.

Group centroids table

A further way of interpreting discriminant analysis results is to describe each group in terms of its profile, using the group means of the predictor variables. These group means are called centroids. Cases with scores near to a centroid are predicted as belonging to that group.

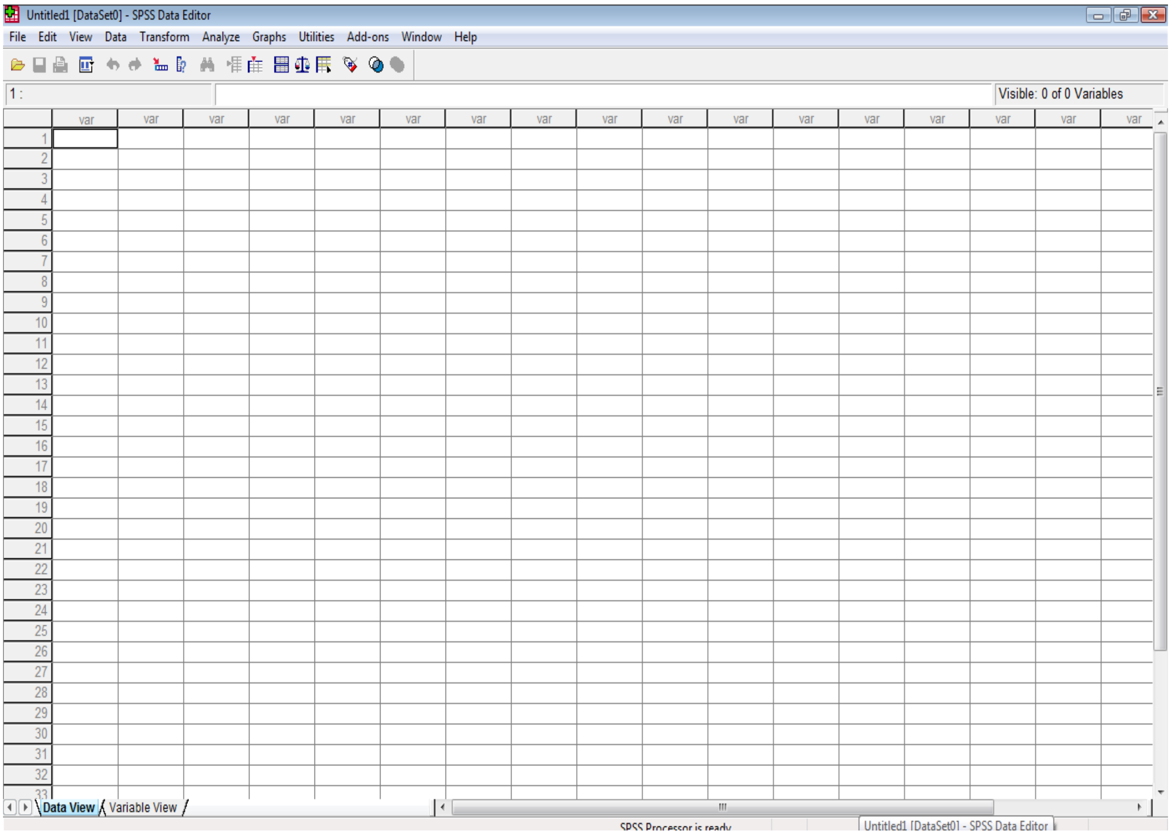
Classification table

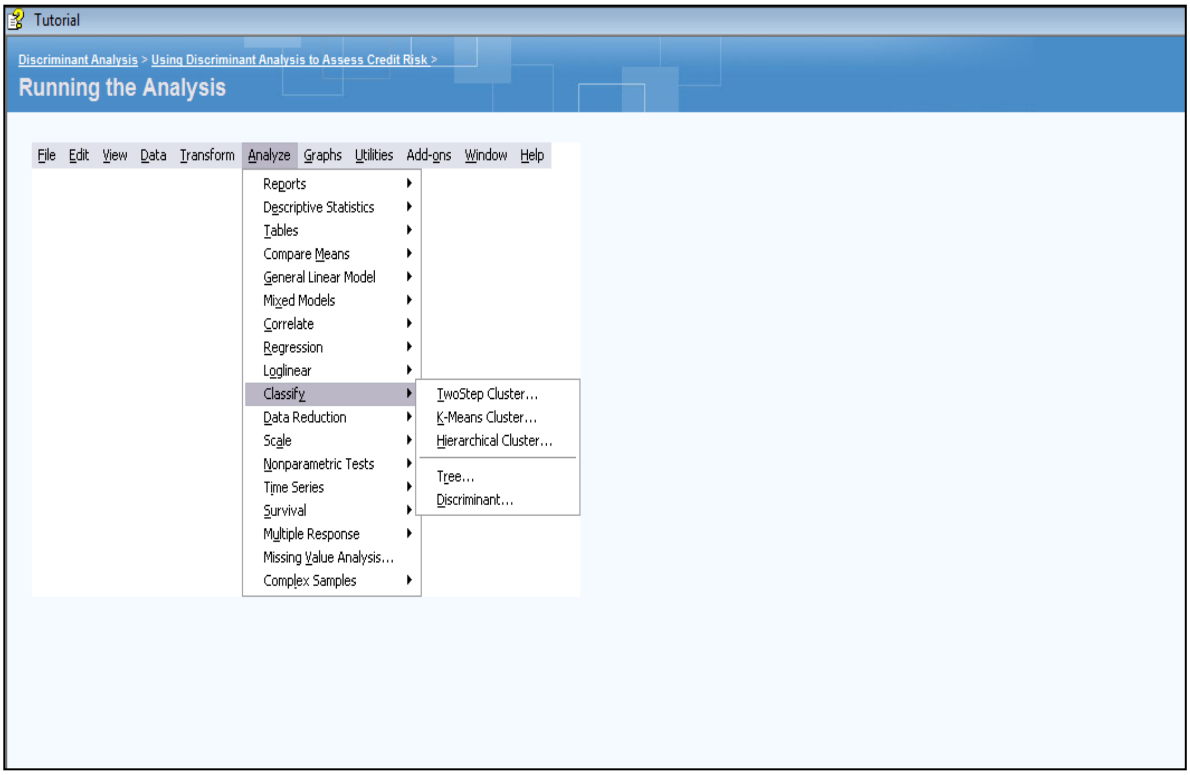
Finally, there is the classification phase. The classification table, also called a confusion table, is simply a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications.

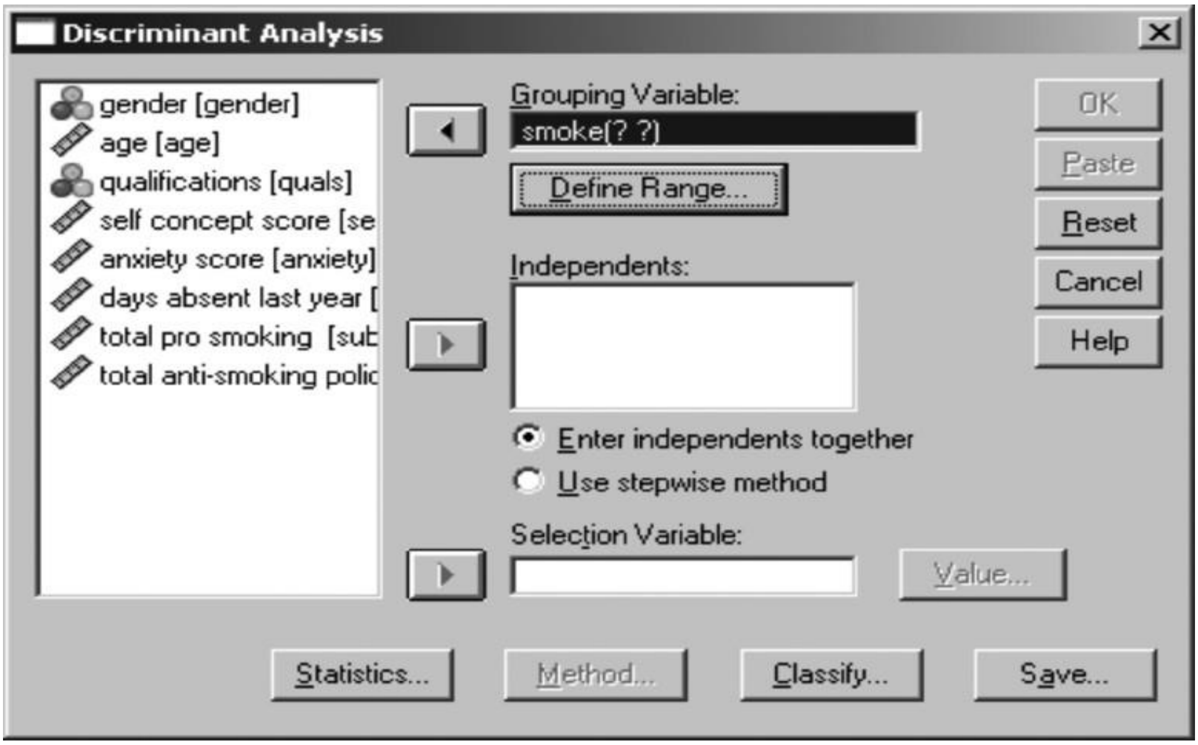
Acknowledgement

This lecture notes is largely drawn from SAGE‘s open access publication on discriminant analysis (www.uk.sagepub.com).

<u>Steps in SSP</u>

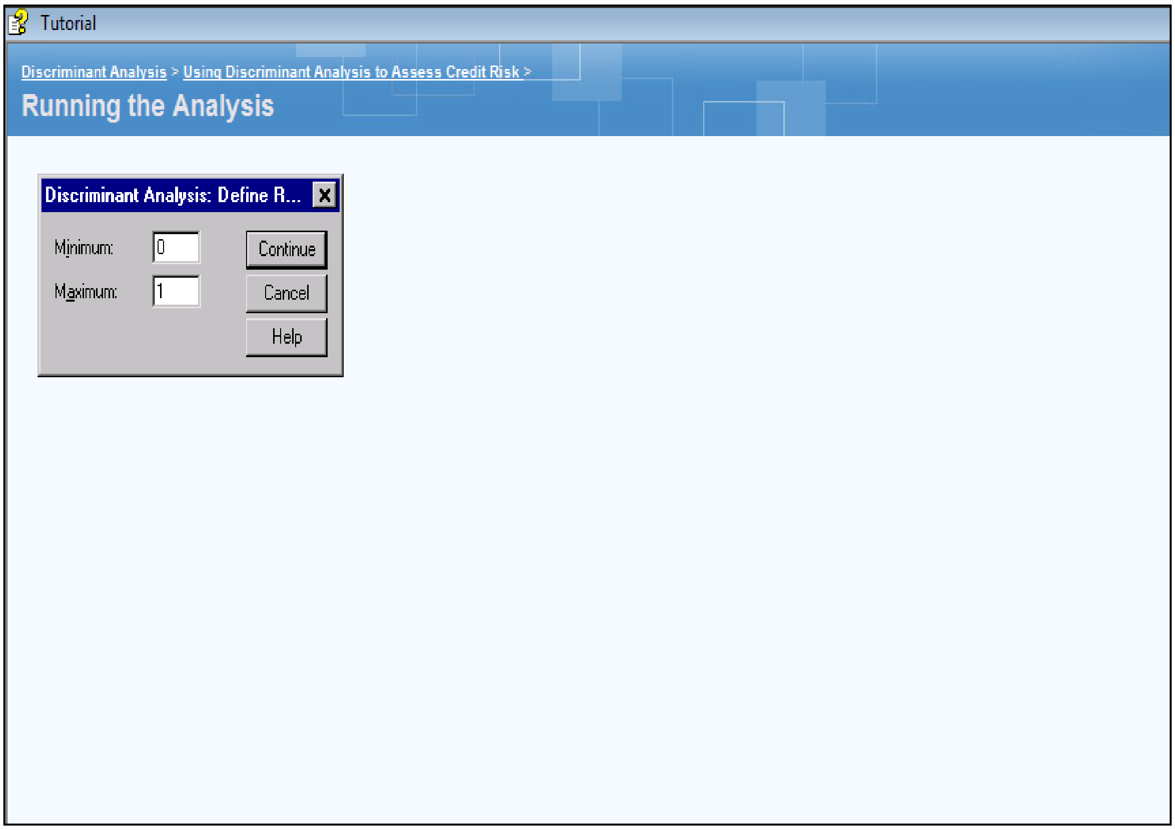






Grouping Variable: smoke(? ?)

● Enter independents together
○ Use stepwise method



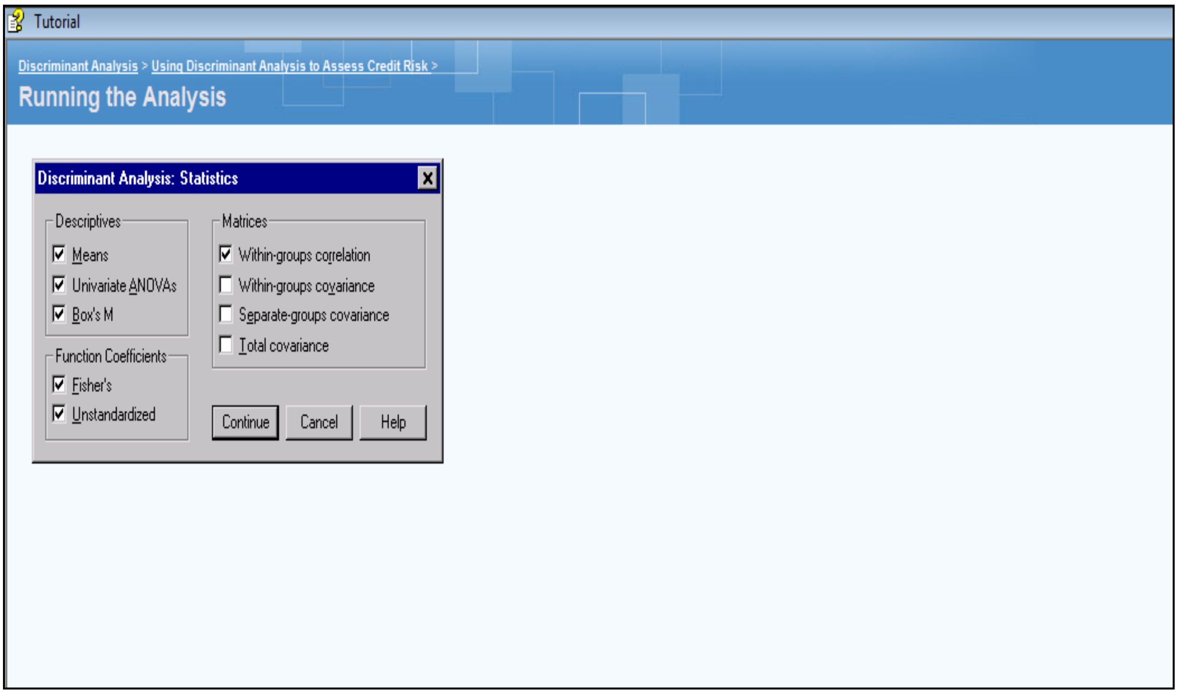
Discriminant Analysis > Using Discriminant Analysis to Assess Credit Risk >

Running the Analysis

Discriminant Analysis: Define R...

Minimum: 0

Maximum: 1



Discriminant Analysis > Using Discriminant Analysis to Assess Credit Risk >

Running the Analysis

Discriminant Analysis: Statistics

Descriptives

☑ Means

☑ Univariate ANOVAs

☑ Box's M

Function Coefficients

☑ Fisher's

☑ Unstandardized

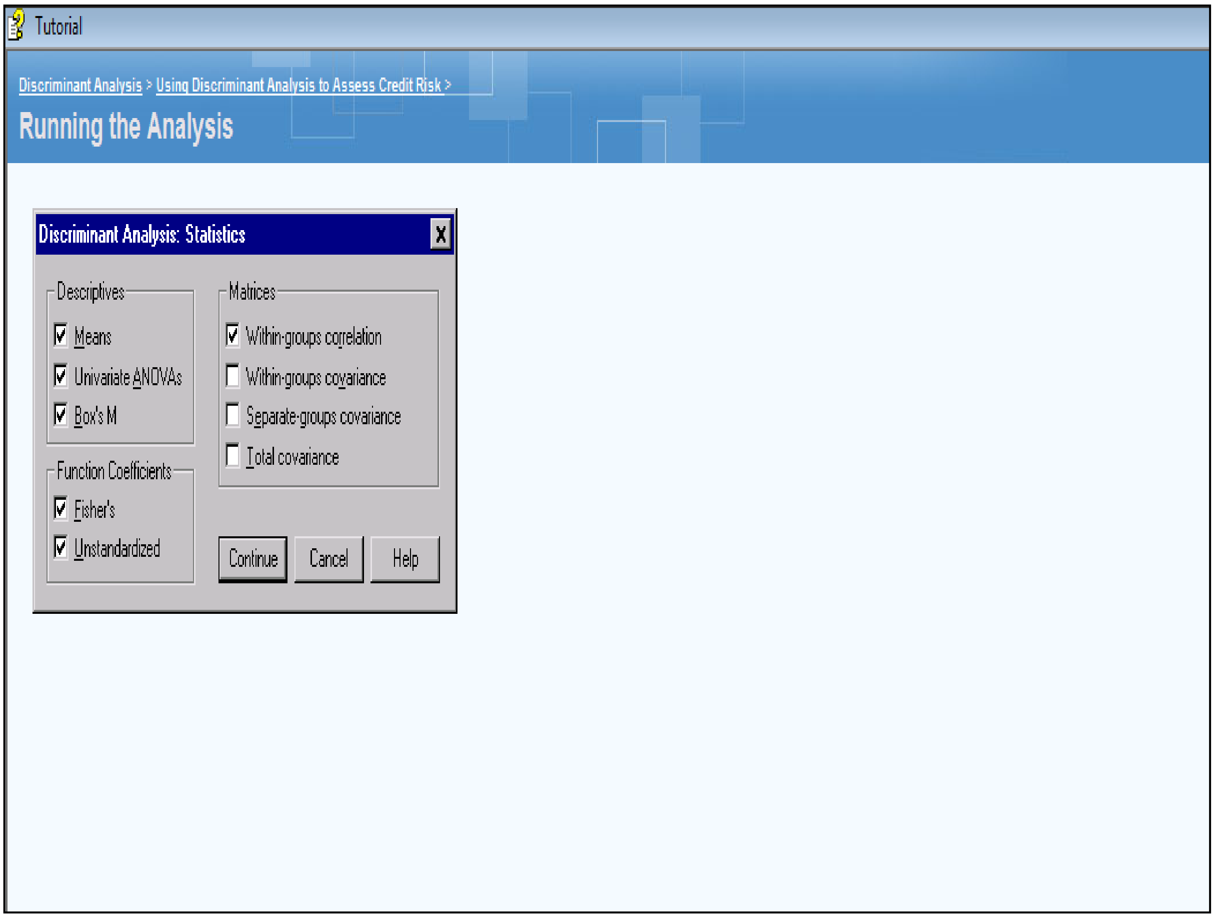
Matrices

☑ Within-groups correlation

☐ Within-groups covariance

☐ Separate-groups covariance

☐ Total covariance


Tutorial

Discriminant Analysis > Using Discriminant Analysis to Assess Credit Risk >

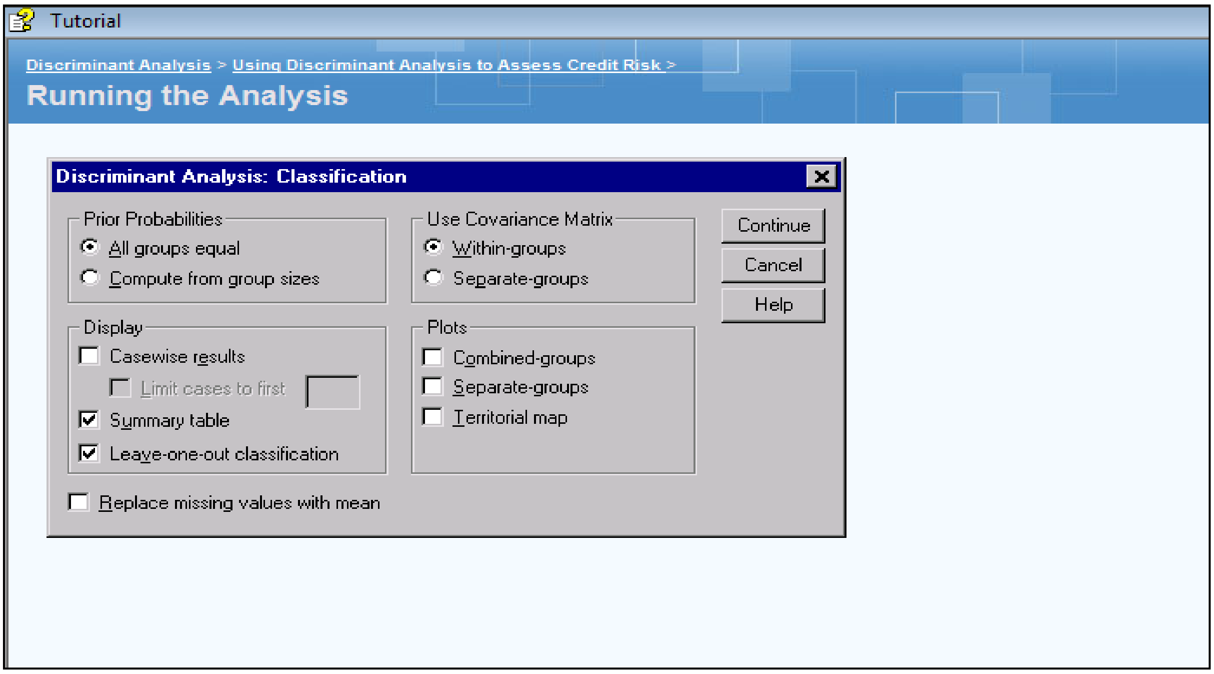
Running the Analysis

Discriminant Analysis: Statistics

Descriptives
- ☑ Means
- ☑ Univariate ANOVAs
- ☑ Box's M

Function Coefficients
- ☑ Fisher's
- ☑ Unstandardized

Matrices
- ☑ Within-groups correlation
- ☐ Within-groups covariance
- ☐ Separate-groups covariance
- ☐ Total covariance


Tutorial

Discriminant Analysis > Using Discriminant Analysis to Assess Credit Risk >

Running the Analysis

Discriminant Analysis: Classification

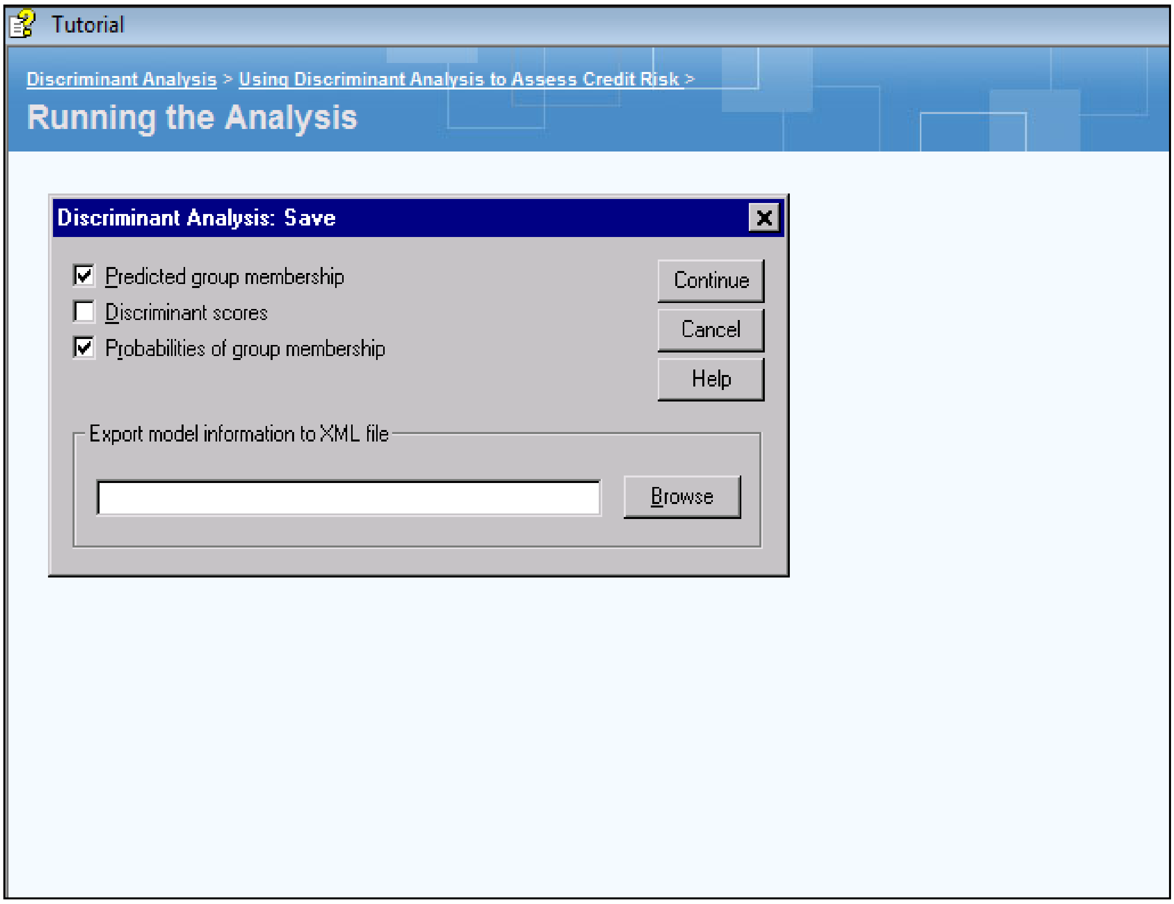
Prior Probabilities
- ● All groups equal
- ○ Compute from group sizes

Use Covariance Matrix
- ● Within-groups
- ○ Separate-groups

Display
- ☐ Casewise results
- ☐ Limit cases to first
- ☑ Summary table
- ☑ Leave-one-out classification

Plots
- ☐ Combined-groups
- ☐ Separate-groups
- ☐ Territorial map

☐ Replace missing values with mean



Tutorial

Discriminant Analysis > Using Discriminant Analysis to Assess Credit Risk >

Running the Analysis

Discriminant Analysis: Save

☑ Predicted group membership

□ Discriminant scores

☑ Probabilities of group membership

Export model information to XML file

Application of conjoint analysis for quantifying attribute preference

P. Venkatesh and Praveen K.V.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: venkatesh1998@gmail.com

Introduction

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

Conjoint Analysis

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

Examples of Conjoint Analysis

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

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

Steps in Conjoint Analysis

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

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

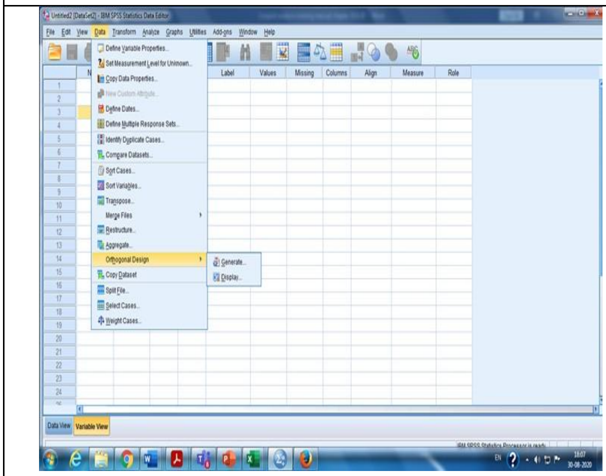
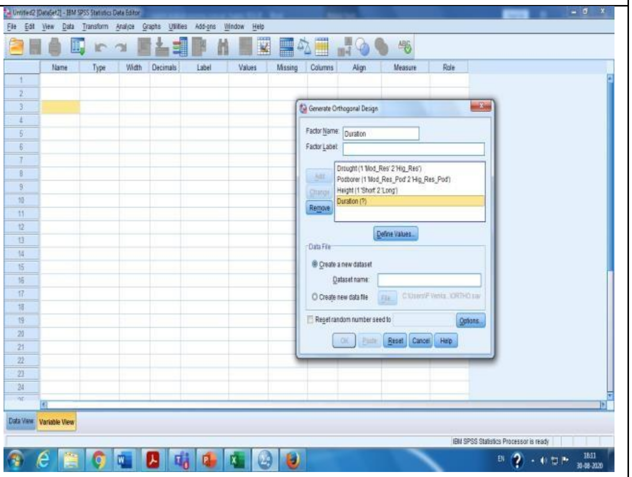
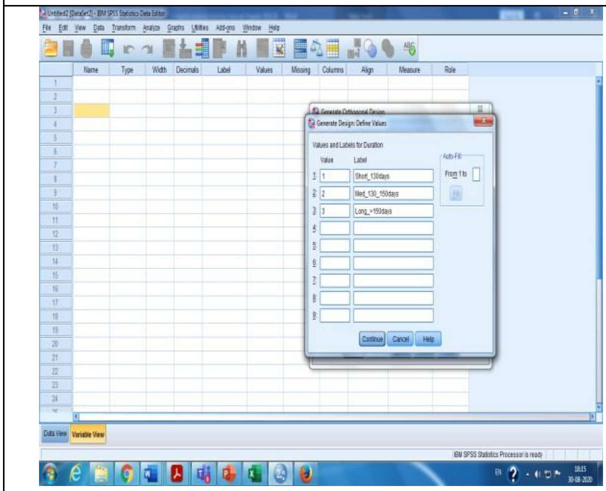
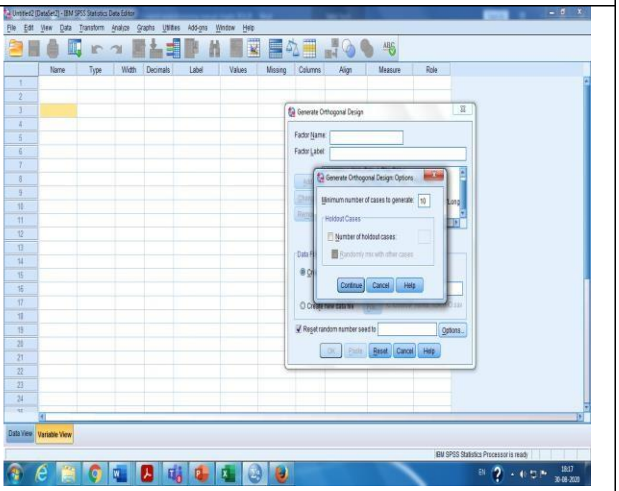
S.No.	Attributes	Levels		
1	Drought	Moderate resistant	Highly resistant	
2	Pod borer	Moderate resistant	Highly resistant	
3	Height	Short	Long	
4	Duration	Short (130 days)	Medium (13-150days)	Long (>150 days)
5	Yield (q/ha)	10-15 q	15 -20 q	> 20 q

2. **Preparation of orthogonal design:** We have considered five attributes and each attribute is having different levels. In total (2x2x2x3x3=72) all possible combinations

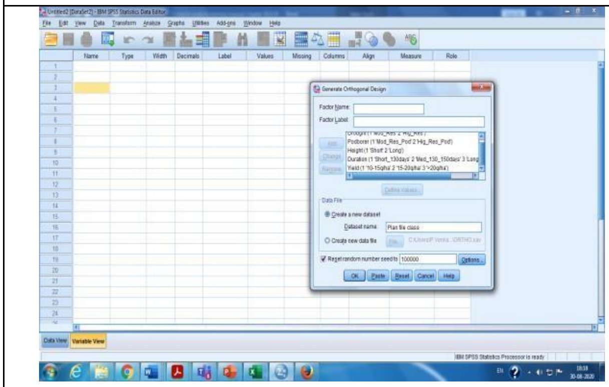
(sets) will be formed. However, it will be difficult for the respondents (farmers) to rank all these combinations. Therefore, by using orthogonal design, we can prepare a manageable number of combinations. Orthogonal Design procedure creates a reduced set of varietal combinations that is small enough to include in a survey but large enough to assess the relative importance of each attributes. By using SPSS, we can generate orthogonal design and plan file will be generated.

3. **Data collection:** A survey will be conducted among the farmers to rank the chosen level of combinations.
4. **Data analysis:** By using SPSS data can be analysed. Both plan file and data files are required for the analysis.
5. **Results:** SPSS will produce both utility files as well relative importance of the factors. By using utility values of each attributes, we can estimate the total utility value s of each combinations and we can find the most proffered combinations.

Analytical procedure in SPSS

1. Orthogonal design create	2. Defining factors
	Generate Orthogonal Design Factor Name: Duration Factor Label: Drought (1 'Mod_Res' 2 'Hig_Res') Podborer (1 'Mod_Res_Pod' 2 'Hig_Res_Pod') Height (1 'Short' 2 'Long') Duration (?) ● Create a new dataset ○ Create new data file □ Reset random number seed to
3. Defining factor levels	**4.Setting minimum number of cases**
Generate Design: Define Values Values and Labels for Duration 1. 1 Short_130days 2. 2 Med_130_150days 3. 3 Long_>150days	Generate Orthogonal Design Options Minimum number of cases to generate: 10 Holdout Cases □ Number of holdout cases: ☑ Reset random number seed to

5.Setting seed value and defining file name



Generate Orthogonal Design

Factor Name:

Factor Label:

Drought (1 'Mod_Res' 2 'Hig_Res')
Podborer (1 'Mod_Res_Pod' 2 'Hig_Res_Pod')
Height (1 'Short' 2 'Long')
Duration (1 'Short_130days' 2 'Med_130_150days' 3 'Long
Yield (1 '10-15q/ha' 2 '15-20q/ha' 3 '>20q/ha')

Data File

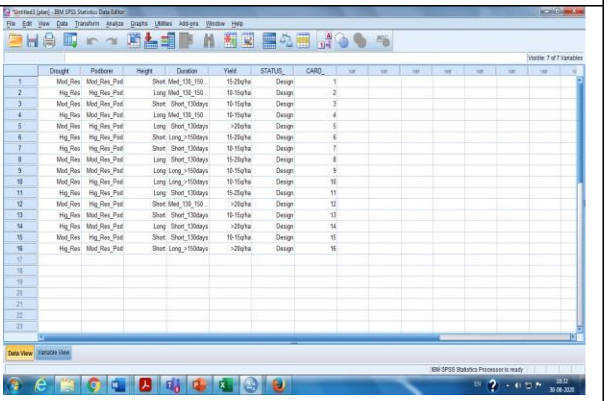
● Create a new dataset

Dataset name: Plan file class

○ Create new data file

☑ Reset random number seed to 100000

6.Plan file generated

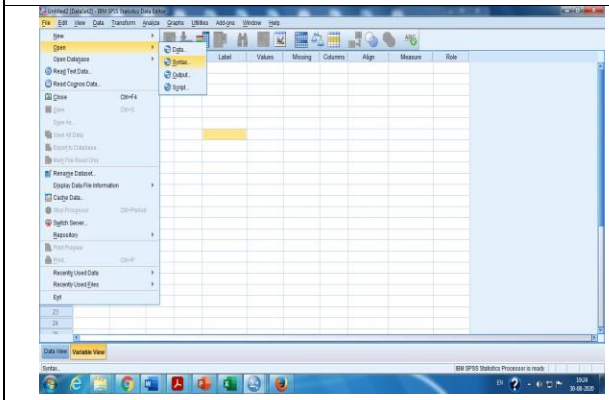


*Untitled3 [plan] - IBM SPSS Statistics Data Editor

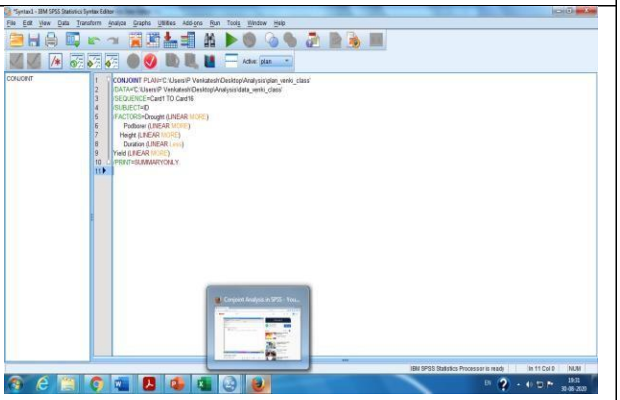
Visible: 7 of 7 Variables

	Drought	Podborer	Height	Duration	Yield	STATUS_	CARD_
1	Mod_Res	Mod_Res_Pod	Short	Med_130_150...	15-20q/ha	Design	1
2	Hig_Res	Hig_Res_Pod	Long	Med_130_150...	10-15q/ha	Design	2
3	Mod_Res	Mod_Res_Pod	Short	Short_130days	10-15q/ha	Design	3
4	Hig_Res	Mod_Res_Pod	Long	Med_130_150...	10-15q/ha	Design	4
5	Mod_Res	Hig_Res_Pod	Long	Short_130days	>20q/ha	Design	5
6	Hig_Res	Hig_Res_Pod	Short	Long_>150days	15-20q/ha	Design	6
7	Hig_Res	Hig_Res_Pod	Short	Short_130days	10-15q/ha	Design	7
8	Mod_Res	Mod_Res_Pod	Long	Short_130days	15-20q/ha	Design	8
9	Mod_Res	Mod_Res_Pod	Long	Long_>150days	10-15q/ha	Design	9
10	Mod_Res	Hig_Res_Pod	Long	Long_>150days	10-15q/ha	Design	10
11	Hig_Res	Hig_Res_Pod	Long	Short_130days	15-20q/ha	Design	11
12	Mod_Res	Hig_Res_Pod	Short	Med_130_150...	>20q/ha	Design	12
13	Hig_Res	Mod_Res_Pod	Short	Short_130days	10-15q/ha	Design	13
14	Hig_Res	Mod_Res_Pod	Long	Short_130days	>20q/ha	Design	14
15	Mod_Res	Hig_Res_Pod	Short	Short_130days	10-15q/ha	Design	15
16	Hig_Res	Mod_Res_Pod	Short	Long_>150days	>20q/ha	Design	16

7.Opening of a syntax file



8.Syntax for analysis



*Syntax1 - IBM SPSS Statistics Syntax Editor

```
CONJOINT PLAN='C:\Users\P Venkatesh\Desktop\Analysis\plan_venki_class'
/DATA='C:\Users\P Venkatesh\Desktop\Analysis\data_venki_class'
/SEQUENCE=Card1 TO Card16
/SUBJECT=ID
/FACTORS=Drought (LINEAR MORE)
    Podborer (LINEAR MORE)
   Height (LINEAR MORE)
    Duration (LINEAR Less)
Yield (LINEAR MORE)
/PRINT=SUMMARYONLY.
```

CONJOINT
Active: plan
Conjoint Analysis in SPSS - You...
IBM SPSS Statistics Processor is ready
In 11 Col 0
NUM

References

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

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

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

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

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

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

Propensity Score Matching and Coarsened Exact Matching for estimating impact[1]

Aditya K.S.[2,3]

[2]*Ph.D. Scholar, Humboldt University of Berlin*

[3]*Scientist, ICAR-Indian Agricultural Research Institute, New Delhi-110012*

Email: adityaag68@gmail.com

Introduction to Matching

In observational studies where, the researcher has no control over the variables unlike experiments, it is difficult to make causal claims. The problem is 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. Due 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.

Propensity Score Matching

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 nonparticipant samples.

[1] *The modified version of the chapter is published as a book chapter. Use appropriate citation in the following format-* ***Aditya K.S*** *and Subash S.P., 2019. Propensity Score Matching in book: "Quantitative Methods for Social Science" Eds Nikam V, Abhimanyu and Pal.S, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi*

To make the point clear, let us consider an example. Suppose we want to measure the impact of adoption of improved variety 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 don't 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/ Assumptions

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 (Figure 1 and Figure 2).

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 the outcome of interest. The aim of matching is to find the closest comparison group from a sample of nonparticipants to the sample of program participants. First, the samples of participants and nonparticipants 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.

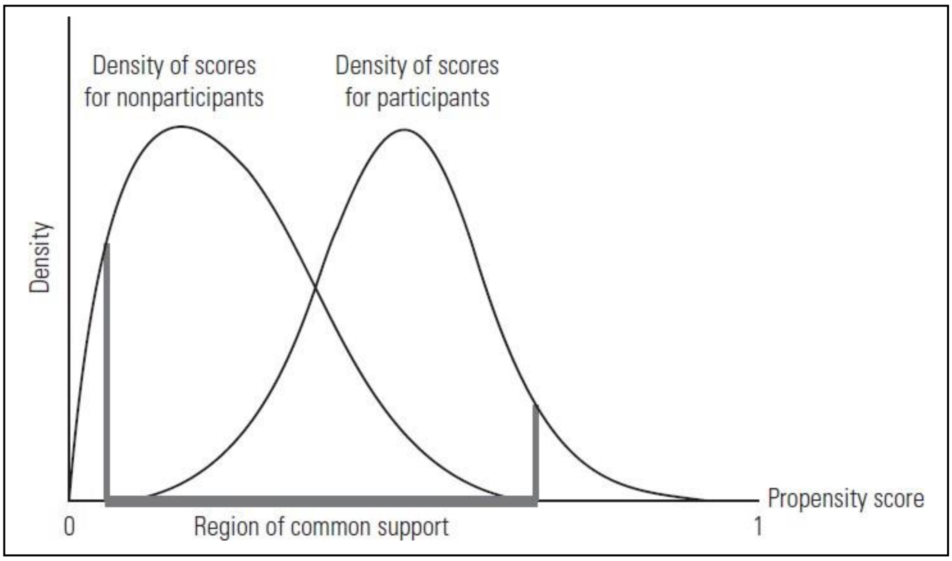


Figure 1. Example of Common Support

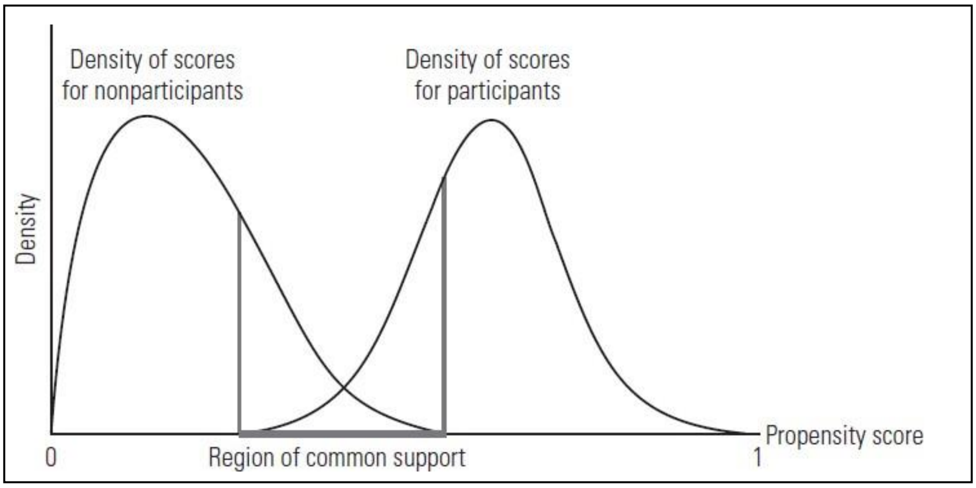


Figure 2. Weak Common Support

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 Bryson, Dorsett, and Purdon (2002); Imbens (2000); and Lechner (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 nonparticipant 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 below, 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.

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 & Imbens, 2008).

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 Neilsen, 2016)

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.

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, nonparametric matching estimators such as kernel matching and LLM use a weighted average of all nonparticipants to construct the counterfactual match for each participant.

Advantages/Criticism

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 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. (See discussion by 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 is not only depends on observables but also on variables that is not observed/ measured. We measure impact using PSM based on the assumption that to 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).
- 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.

Coarsened Exact Matching[1]

Let us denote the pre-treatment variables by a vector X. In the first step, all those treated and control units outside the common support region will be dropped. In the second step, treated units are matched with the control units with respect to some metrics. Mahalanobi's matching uses Mahalanobi's distance between the units as metrics, whereas Propensity Score Matching (PSM) uses propensity score, a scalar derived from combining the covariates to be used as a balancing score.

Univariate matching methods like PSM do not guarantee any reduction in imbalance and depend on a set of unverifiable assumptions (Conditional Independence Assumption, for example). Further, when the matching is used for pruning observation as a data pre-processing approach, commonly used methods like PSM could result in increased imbalance, due to what is known as PSM paradox (Gary and Richard, 2019). Propensity scores also depend on the specification of the model used to estimate it. PSM can sometimes lead to improvements in balance with respect to one covariate while increasing the bias with respect to some other covariates, even though the mean values are similar (QIN, 2007). Univariate balancing methods aim to obtain univariate balance on mean of covariate. But such methods may not remove imbalance due to interaction with and the nonlinear function of X (vector of confounders). CEM is one alternative to commonly used univariate balancing methods.

CEM belongs to a class of Monotonic Imbalance Bounding (MIB) methods. (Blackwell, Iacus, King, and Porro, 2009). These methods use multivariate distributions for balancing and studies

[1] This portion is a verbatim copy from the methodology section of Kumar et al., 2022. This is only for training purpose and use appropriate source while citing.

have shown them to be more effective than others in reducing data imbalance and model dependence. The method can be best described with the following set of equations following QIN (2007) and Iacus, King, and Porro (2009).

$$\left\{\begin{array}{c} D\,(f_1\,(x_{m_{T(\pi)}})\,, f_1\,(x_{m_{C(\pi)}})) \le \gamma_1(\pi_1) \\ \vdots \\ D\,(f_k\,(x_{m_{T(\pi)}})\,, f_k\,(x_{m_{C(\pi)}})) \le \gamma_k(\pi_k) \end{array}\right\}$$

In every dimension of X, Distance D between function f (.) of X in treated and f (.) X in control should be smaller than the monotonically increasing function of γ (π). This directly led us to

$$D\,(f_j\,(x_{m_{T(\pi)}})\,, f_j\,(x_{m_{C(\pi)}})) \le \gamma_j(\pi - \epsilon) < \gamma_j(\pi), j = 1, \dots, k.), \text{if } \epsilon > 0$$

Where π are the tuning parameters for each variable which researchers can specify. When the X of treated and control units meet the above set of inequalities, they can be matched. This is the fundamental principle of all Monotonic Imbalance Bounding matching techniques.

Further, let us consider X_i, one element of vector X. In CEM, X_i is divided to V_i number of classes or intervals based on researchers' understanding/intuitions.

$$\gamma_i(\pi_i) \; = \; \gamma_{i1}(\pi_{i1}), \gamma_{i2}(\pi_{i2}) \dots \gamma_{iV_i}(\pi_{iV_i})$$

The approach of CEM can be summarized in three simple steps, following (Datta, 2015) (QIN, 2007) (Blackwell et al., 2009) (Iacus et al., 2012)(Gary and Richard, 2019):

- Temporarily Coarsen the variables in X into classes, which we refer to as strata or hypercuboids
- Sort units within the hypercuboid/rectangles according to original values of X
- Keep only the matched units; Units with strata with at least one treated and control unit are retained (the original values of X can be used for adjusting remaining imbalance)

Imbalance Measure

Generally, the univariate difference in mean values of covariates across treated and control groups is considered as a measure of imbalance. But this doesn't reflect multivariate imbalance and imbalance due to other moments. Iacus et al. (2012) suggest an alternate measure;

f_1 represents the distance between the multivariate histograms of X. Let us denote $H(X_1)$ which indicates the number of bins (or unique values) chosen for the variable X_1. Multivariate histograms are constructed by Cartesian product of $H(X_1)\;X\;H\;(X_2) \dots H(X_k) = H$, which forms the cells for constructing the multivariate histograms. Denote the relative frequency of the treated and control groups by f and g. Let $l_1 \ldots l_k$ be the relative frequency corresponding to a particular cell. Then f_1 is calculated through the following formulae

$$f_1(f,g;H) = \frac{1}{2} \sum_{l_1 \ldots l_k \in H(X)} |\, f\, l_1 \ldots . l_k - g l_1 \ldots . l_k \,|$$

If the value of f_1 is 1, it indicates perfect separation and if the value is 0, it indicates perfect matching of the multivariate distributions. A good matching process should result in decreased value of f_1.

Analysis – Do files

```
**Propensity Score Matching for Impact Assessment**

**Impacts of program participation; Impact of participation of male member in SHG on the per capit
household expenditure**

**Variables of interest are- dmmfd= male member is in SHG, dfmfd- female member in SHG, and the
outcome of interest is lnexptot= Ln (Total Expenditure**

** First describe/ summarise the variable of interests including outcome**

    ttable2 sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg, by(dmmfd)

/*pscore estimation***

** If pscore user written command is not installed in your machine, please use the follwoing
command**

                              findit pscore

**browse through packages and install  "st0026" from the list**

 pscore dmmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg [pw=weight],
                                    /*

               */ pscore(ps98) blockid(blockf1) comsup level(0.001)

**Common Support Graph**

      twoway (kdensity ps98 if  dmmfd==1) (kdensity ps98 if  dmmfd==0, lpattern(dash)), */

         */ legend( label( 1 "treated") label(  2 "control" ) ) xtitle("propensity score")

**Rerun the model after dropping Egg and Land**

**We need to drop ps98 and blockf1 as they are already defined and created during first Pscore
estimation**

                           drop ps98 blockf1

  pscore dmmfd sexhead agehead educhead  vaccess pcirr rice wheat milk oil [pw=weight], /*

               */ pscore(ps98) blockid(blockf1) comsup level(0.001)
```

```
**Again install psmatch2 first**
ssc install psmatch2
**or you can use the findit command as earlier**
psmatch2  dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil, outcome(lexptot)
**Alternatively, Direct Matching using Nearest neighbor; first install nnmatch2**
findit nnmatch2
nnmatch2 lexptot dmmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg
[pw=weight], tc(att) m(1)
nnmatch2 lexptot dfmfd sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg
[pw=weight], tc(att) m(1)
**Different Methods of Matching**
psmatch2  dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil, outcome(lexptot) n(1)
psmatch2  dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil, outcome(lexptot) n(3)
psmatch2  dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil, outcome(lexptot)
cal(0.014)
**Bootstrpping standard Errors**
bootstrap r(att) :psmatch2  dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil,
outcome(lexptot) cal(0.014)
**R Bounds**
gen delta1 = lexptot - _lexptot if _treat==1 & _support==1
rbounds delta1, gamma(1 (0.1) 2)
```

```
**Coarsened Exact Matching**
**CEM** **Use the following link to get the data**
use http://gking.harvard.edu/cem/lalonde.dta, clear
**The data is on some training program aimed at increasing the earnings**
**Let's Tabulate**
table  treated
regress re78 treated
**Install the Imbalance Program**
ssc install imbalance
**Use imb command to estimate the imbalance levels**
imb  age education black nodegree re74, treatment(treated)
```

```
**Use CEM command to coarsen the data** Explore different ways of CEM**
cem age education black nodegree re74, treatment(treated)
tab cem_strata cem_matched
cem age education (0 6.5 8.5 12.5 16.5) black nodegree re74, treatment(treated)
recode q1 (12=1 "agree") (3 6 = 2 "neutral") (45=3 "disagree"),generate(cem_q1)
 cem age education black nodegree re74 cem_q1 (#0), treatment(treated)
```

```
**Causal inference**
regress re78 treated re74 re75 [iweight=cem_weights]
**USING Mahalanobi's nearest neighbour matching**
nnmatch re78 treated age education black nodegree re74
nnmatch re78 treated age education black nodegree re74, tc(att)
**with four matches**
nnmatch re78 treated age education black nodegree re74, tc(att) m(4)
**After adjusting for bias**
nnmatch re78 treated age education black nodegree re74, tc(att) m(4) bias(bias)
**Adjusting for SE and Variance**Estimate conditional variance function to allow for
heteroscedastic variance**
nnmatch re78 treated age education black nodegree re74, tc(att) m(4) bias(bias) robust(4)
```

References

Abadie, A. and Imbens, G.W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6): 1537-1557.

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.

Aditya, K.S., Khan, T. and Kishore, A. (2018). Adoption of crop insurance and impact: insights from India. *Agricultural Economics Research Review*, 31(347-2019-565): 163-174.

Becker, Sascha and Andrea Ichino (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *Stata Journal*, 2(4): 358–77.

Bryson, Alex, Richard Dorsett and Susan 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, Marco and Sabine Kopeinig (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1): 31–72.

Fischer, Elisabeth and Matin Qaim (2012). Linking Smallholders to Markets: Determinants and Impacts of Farmer Collective Action in Kenya. *World Development*, 40(6): 1255–68.

Heckman, James J., Hidehiko Ichimura and Petra Todd (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*, 64(4): 605–654.

Heckman, James J., Hidehiko Ichimura and Petra Todd (1998). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies*, 65(2): 261–294.

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): 141–159.

Imbens, Guido (2000). The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*, 87(3): 706–710.

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

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. Political Analysis, 1-20.

Kumar, A., Sonkar, V.K. and Aditya, K.S. (2022). Assessing the Impact of Lending Through Kisan Credit Cards in Rural India: Evidence from Eastern India. *The European Journal of Development Research*, 1-21.

Lechner, Michael (2001). Identifi cation 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, pp 43–58. Heidelberg and New York: Physica-Verlag.

Mendola, Mariapia (2007). Agricultural Technology Adoption and Poverty Reduction: A Propensity-Score Matching Analysis for Rural Bangladesh. *Food Policy*, 32(3): 372–393.

Ravallion, Martin (2008). Evaluating Anti-Poverty Programs. In: *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, pp 3787–846. Amsterdam: North-Holland.

Shiferaw, Bekele, Tewodros, 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: 309–323.

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): 243–248.

Synthetic Control Method for Assessing Impact of Policy Change

Prabhat Kishore
ICAR- National Institute of Agricultural Economics and Policy Research (NIAP), New Delhi-110012
Email: kishore.prabhat89@gmail.com

Theoretical Concept

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

However, in some cases intervention takes place at state or country and policy maker are interested to know impact of the intervention at that macro level. With traditional approach it seems to be difficult as the unit of intervention itself is single or may be few some time. So there is lack of sufficient number of treated and control unit for inferences. Major policy intervention occurred at macro level like Government of Bihar has repealed APMCs Act in 2006 with motif to remove barriers in agricultural marketing or Punjab Government enacted Punjab Preservation of subsoil water act 2009 to regulate groundwater depletion in the state. To know the impact of these interventions at aggregate level, there is need to have a robust method which can provide suitable comparison unit for the inference. In this context, Synthetic

Control Method provides new insight to tackle stated problems in impact assessment methodology.

Analytical Method

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

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

Econometric Model

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

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

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

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

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

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

by selecting the optimal values of *W* and *Vm* where *X₁m* is the value of the *m*th attribute of the treated unit; X_0m is a 1 x *j* vector containing the values of the *m*th predictor attribute of each of the S potential comparison or control units; *W* is a vector of weights on control units; and *Vm* is a vector of weights on attributes of the control units such that they maximize the ability to predict the outcome variable of interest (Abadie et al. 2010). This optimization process minimizes prediction error between the actual and the synthetic in the pre-treatment period.

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

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

Illustration

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

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

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

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

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

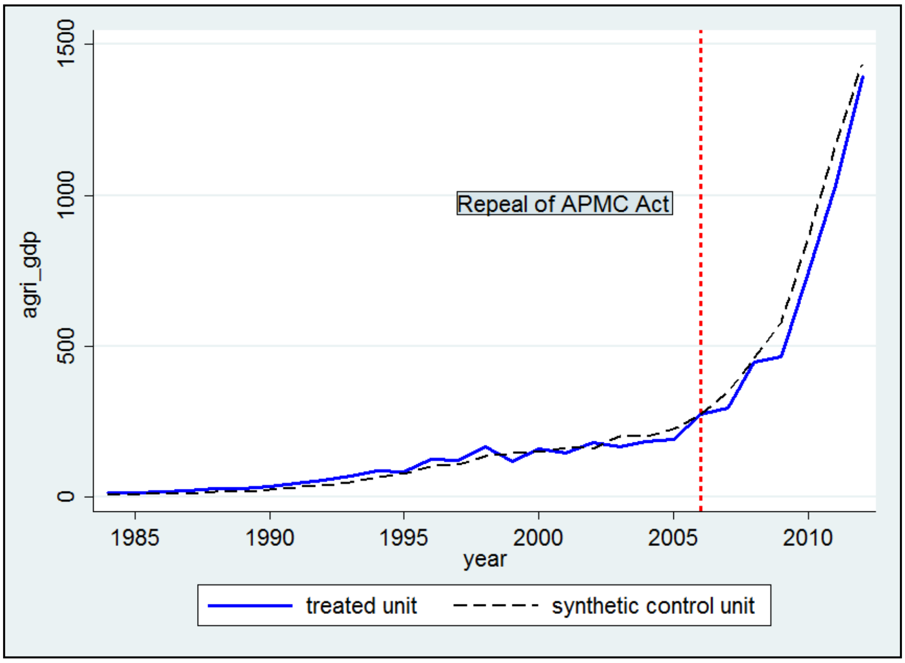


Figure 1. Trends in Agri. GDP: Bihar vs. synthetic Bihar

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

Table 2: State weight in Synthetic Bihar

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

Post Estimation

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

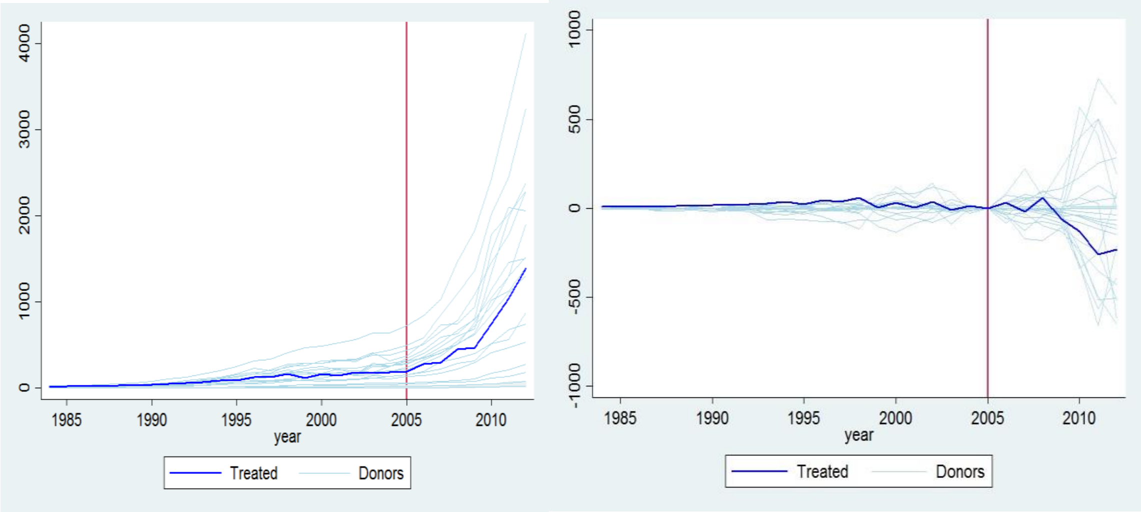


Figure 2. Placebo test: Counterfactual for agricultural GDP and gap between actual and synthetic of Bihar

Acknowledgment

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

References

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

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

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

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

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

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

Draft paper:

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

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

An Introduction to Systematic Reviews and Meta-Analysis

Praveen K.V. and Asha Devi S.S.

Division of Agriculture Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: veenkv@gmail.com

Introduction

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

Research Questions

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

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

Table 1. PICO formulation

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

Defining Inclusion Criteria

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

Literature Search and Study Selection

Systematic reviews require a careful, objective, and reproducible search of a variety of sources to extract as many studies (eligible) as possible. The quest for studies should be as broad as possible to diminish the danger of reporting bias and to identify maximum evidence as possible. Database determination ought to be guided by the survey theme. 'Grey literature' should also be considered. Authors ought to search for dissertations and conference abstracts also. They should also think about looking through the web, hand searching of journals and looking through full texts of journals electronically where accessible. They ought to inspect past reviews on a similar theme and check reference lists of included studies. Suitable search strategy should be formulated for searching in different databases. Choices about which studies to include for a review is among the most compelling choices that are made in the review procedure and they include judgment. Involvements of at least two individuals, working independently, are required to decide if each study meets the qualification standards. A PRISMA flow chart mentioning the selection of studies at each stage should be included in the report.

Table 2. List of databases to search

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

Coding and Data Collection

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

- Study identification
- Intervention discriptives
- Process and implementation
- Context
- Popultion characteristics
- Research methods
- Effect size data
- Outcomes
- Subgroups

Effect Measures

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

Meta-analysis

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

- ***Fixed effect model***
  - Under the fixed-effect model we assume that all studies in the meta-analysis share a common (true) effect size.
  - Put another way, all factors that could influence the effect size are the same in all the studies, and therefore the true effect size is the same in all the studies.

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

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

Setting the question

The effects of biofertilizer use in crop yields in India

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

Search strategy for meta-analysis

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

Screening, coding and data extraction

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

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

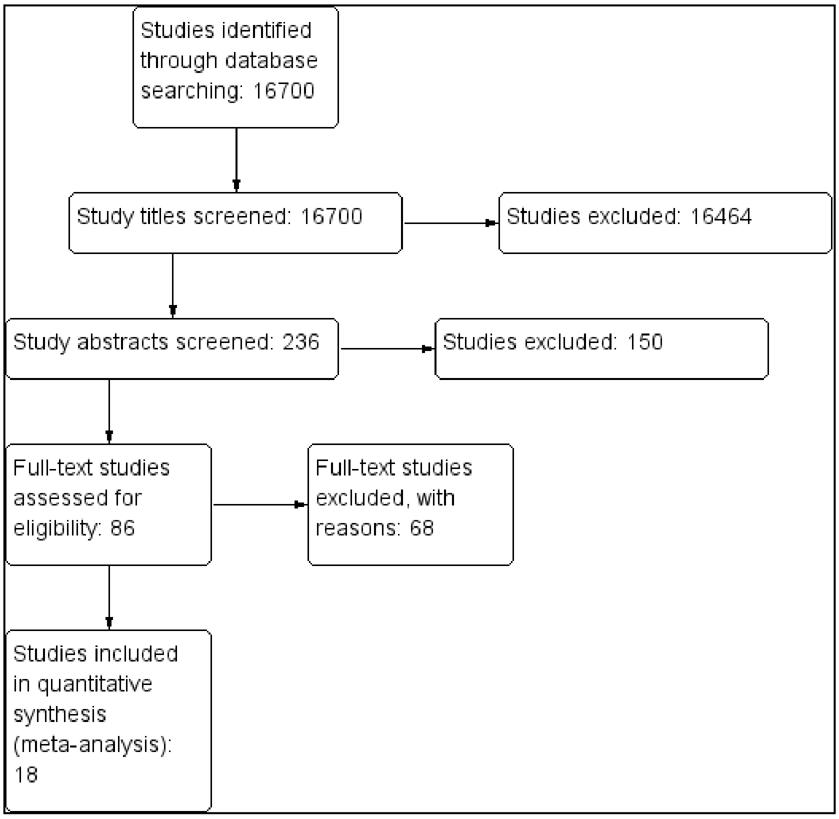
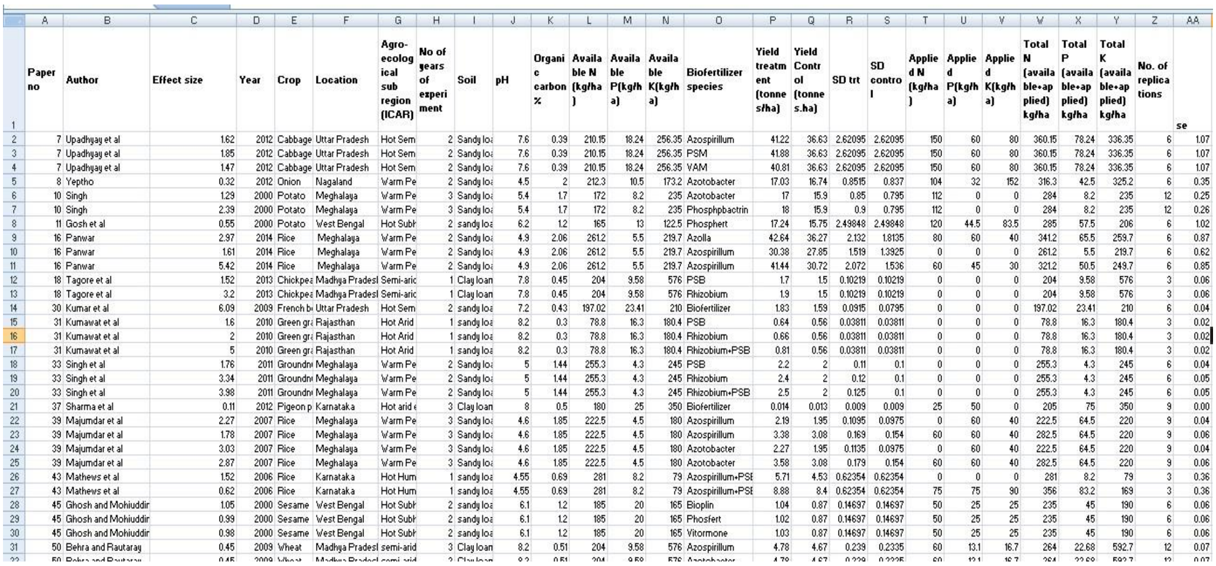

Studies identified through database searching: 16700
Study titles screened: 16700
Studies excluded: 16464
Study abstracts screened: 236
Studies excluded: 150
Full-text studies assessed for eligibility: 86
Full-text studies excluded, with reasons: 68
Studies included in quantitative synthesis (meta-analysis): 18

Figure 1. PRISMA flow chart

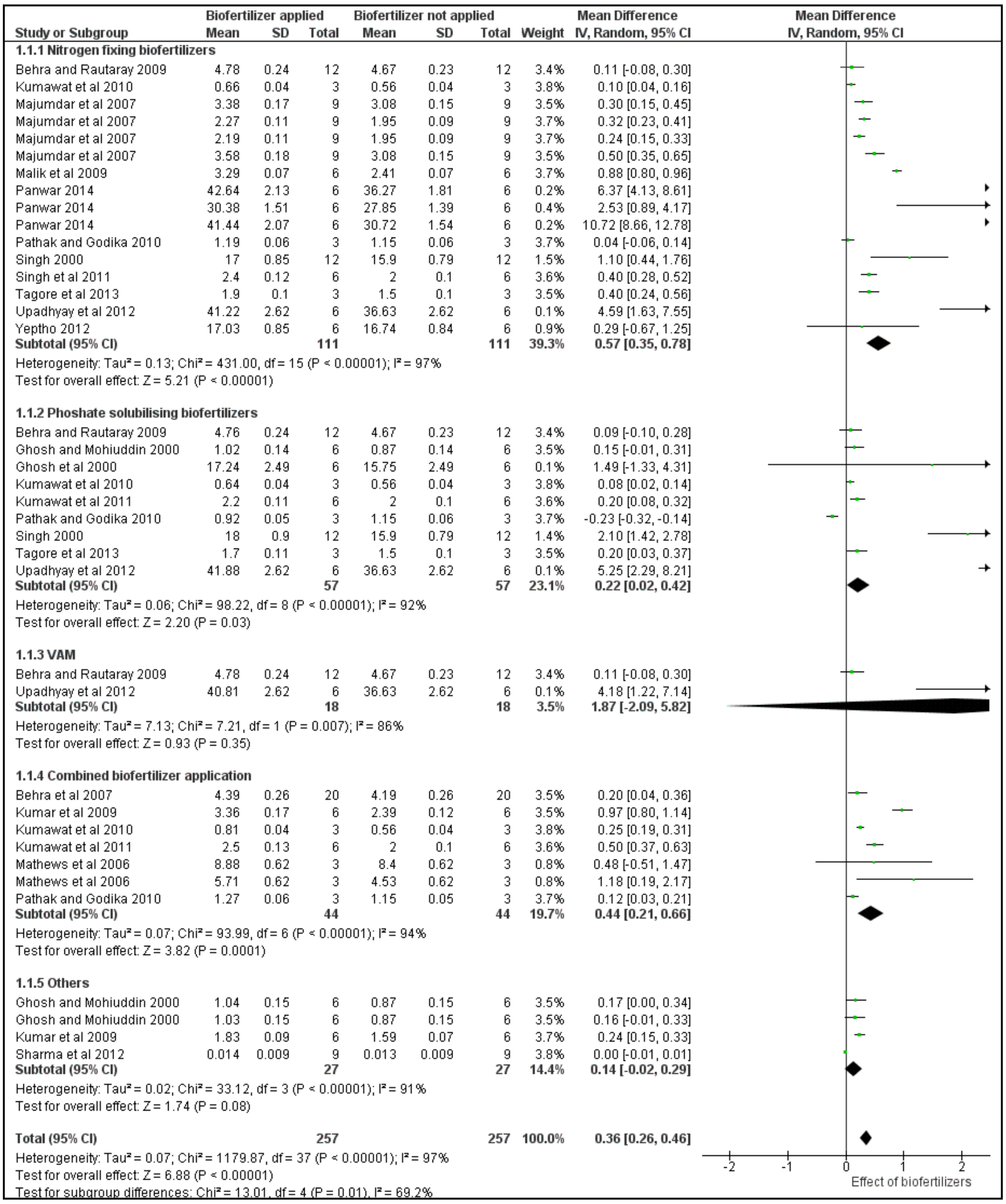


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Paper no	Author	Effect size	Year	Crop	Location	Agro-ecological sub region (ICAR)	No of years of experiment	Soil	pH	Organic carbon %	Available N (kg/ha)	Available P(kg/ha)	Available K(kg/ha)	Biofertilizer species	Yield treatment (tonnes/ha)	Yield Control (tonnes.ha)	SD trt	SD control	Applied N (kg/ha)	Applied P(kg/ha)	Applied K(kg/ha)	Total N (available+applied) kg/ha	Total P (available+applied) kg/ha	Total K (available+applied) kg/ha	No. of replications	se
2	7	Upadhyay et al	1.62	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	Azospirillum	41.22	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
3	7	Upadhyay et al	1.85	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	PSM	41.88	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
4	7	Upadhyay et al	1.47	2012	Cabbage	Uttar Pradesh	Hot Sem	2	Sandy loa	7.6	0.39	210.15	18.24	256.35	VAM	40.81	36.63	2.62095	2.62095	150	60	80	360.15	78.24	336.35	6	1.07
5	8	Yeptho	0.32	2012	Onion	Nagaland	Warm Pe	2	Sandy loa	4.5	2	212.3	10.5	173.2	Azotobacter	17.03	16.74	0.8515	0.837	104	32	152	316.3	42.5	325.2	6	0.35
6	10	Singh	1.29	2000	Potato	Meghalaya	Warm Pe	3	Sandy loa	5.4	1.7	172	8.2	235	Azotobacter	17	15.9	0.85	0.795	112	0	0	284	8.2	235	12	0.25
7	10	Singh	2.39	2000	Potato	Meghalaya	Warm Pe	3	Sandy loa	5.4	1.7	172	8.2	235	Phosphpbactrin	18	15.9	0.9	0.795	112	0	0	284	8.2	235	12	0.26
8	11	Gosh et al	0.55	2000	Potato	West Bengal	Hot Subh	2	sandy loa	6.2	1.2	165	13	122.5	Phosphert	17.24	15.75	2.49848	2.49848	120	44.5	83.5	285	57.5	206	6	1.02
9	16	Panwar	2.97	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azolla	42.64	36.27	2.132	1.8135	80	60	40	341.2	65.5	259.7	6	0.87
10	16	Panwar	1.61	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azospirillum	30.38	27.85	1.519	1.3925	0	0	0	261.2	5.5	219.7	6	0.62
11	16	Panwar	5.42	2014	Rice	Meghalaya	Warm Pe	2	Sandy loa	4.9	2.06	261.2	5.5	219.7	Azospirillum	41.44	30.72	2.072	1.536	60	45	30	321.2	50.5	249.7	6	0.85
12	18	Tagore et al	1.52	2013	Chickpea	Madhya Pradesl	Semi-arid	1	Clay loan	7.8	0.45	204	9.58	576	PSB	1.7	1.5	0.10219	0.10219	0	0	0	204	9.58	576	3	0.06
13	18	Tagore et al	3.2	2013	Chickpea	Madhya Pradesl	Semi-arid	1	Clay loan	7.8	0.45	204	9.58	576	Rhizobium	1.9	1.5	0.10219	0.10219	0	0	0	204	9.58	576	3	0.06
14	30	Kumar et al	6.09	2009	French b	Uttar Pradesh	Hot Sem	2	sandy loa	7.2	0.43	197.02	23.41	210	Biofertilizer	1.83	1.59	0.0915	0.0795	0	0	0	197.02	23.41	210	6	0.04
15	31	Kumawat et al	1.6	2010	Green gra	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	PSB	0.64	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
16	31	Kumawat et al	2	2010	Green gra	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	Rhizobium	0.66	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
17	31	Kumawat et al	5	2010	Green gra	Rajasthan	Hot Arid	1	sandy loa	8.2	0.3	78.8	16.3	180.4	Rhizobium+PSB	0.81	0.56	0.03811	0.03811	0	0	0	78.8	16.3	180.4	3	0.02
18	33	Singh et al	1.76	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	PSB	2.2	2	0.11	0.1	0	0	0	255.3	4.3	245	6	0.04
19	33	Singh et al	3.34	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	Rhizobium	2.4	2	0.12	0.1	0	0	0	255.3	4.3	245	6	0.05
20	33	Singh et al	3.98	2011	Groundn	Meghalaya	Warm Pe	2	Sandy loa	5	1.44	255.3	4.3	245	Rhizobium+PSB	2.5	2	0.125	0.1	0	0	0	255.3	4.3	245	6	0.05
21	37	Sharma et al	0.11	2012	Pigeon p	Karnataka	Hot arid e	3	Clay loan	8	0.5	180	25	350	Biofertilizer	0.014	0.013	0.009	0.009	25	50	0	205	75	350	9	0.00
22	39	Majumdar et al	2.27	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azospirillum	2.19	1.95	0.1095	0.0975	0	60	40	222.5	64.5	220	9	0.04
23	39	Majumdar et al	1.78	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azospirillum	3.38	3.08	0.169	0.154	60	60	40	282.5	64.5	220	9	0.06
24	39	Majumdar et al	3.03	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azotobacter	2.27	1.95	0.1135	0.0975	0	60	40	222.5	64.5	220	9	0.04
25	39	Majumdar et al	2.87	2007	Rice	Meghalaya	Warm Pe	3	Sandy loa	4.6	1.85	222.5	4.5	180	Azotobacter	3.58	3.08	0.179	0.154	60	60	40	282.5	64.5	220	9	0.06
26	43	Mathews et al	1.52	2006	Rice	Karnataka	Hot Hum	1	sandy loa	4.55	0.69	281	8.2	79	Azospirillum+PSE	5.71	4.53	0.62354	0.62354	0	0	0	281	8.2	79	3	0.36
27	43	Mathews et al	0.62	2006	Rice	Karnataka	Hot Hum	1	sandy loa	4.55	0.69	281	8.2	79	Azospirillum+PSE	8.88	8.4	0.62354	0.62354	75	75	90	356	83.2	169	3	0.36
28	45	Ghosh and Mohiuddir	1.05	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Bioplin	1.04	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
29	45	Ghosh and Mohiuddir	0.99	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Phosfert	1.02	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
30	45	Ghosh and Mohiuddir	0.98	2000	Sesame	West Bengal	Hot Subh	2	sandy loa	6.1	1.2	185	20	165	Vitormone	1.03	0.87	0.14697	0.14697	50	25	25	235	45	190	6	0.06
31	50	Behra and Rautaray	0.45	2009	Wheat	Madhya Pradesl	semi-arid	3	Clay loan	8.2	0.51	204	9.58	576	Azospirillum	4.78	4.67	0.239	0.2335	60	13.1	16.7	264	22.68	592.7	12	0.07
32	50	Behra and Rautaray	0.45	2009	Wheat	Madhya Pradesl	semi-arid	3	Clay loan	8.2	0.51	204	9.58	576	Azotobacter	4.78	4.67	0.239	0.2335	60	13.1	16.7	264	22.68	592.7	12	0.07

Figure 2. Coding

Meta-analysis

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



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

Figure 3. Forest plot

Table 3. Meta-regression

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

References

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

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

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

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

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

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

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

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

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

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

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

Application of Analytic Hierarchy Process (AHP) in Social Science Research

Misha Madhavan M. and Bhagya Vijayan

Division of Agricultural extension, ICAR-Indian Agricultural Research Institute, New Delhi-110012

Email: mishamadhavanmsy4@gmail.com

Introduction

Prioritization of different available options is a basic necessity in social science research. The options can be of different kinds like research projects, constraints, strengths, weaknesses, opportunities, threats etc. The interactive decision making among choices using a valid scientific instrument or procedure is having immense significance in field-oriented research works. The Analytic Hierarchy Process (AHP) was introduced by T.L. Saaty as discussed by Saaty (1977) & Saaty (1980) and used for systematic priority setting. This method is based on relative importance of alternatives as perceived by various stakeholders. AHP considers the tangible and intangible values (Dyer and Forman,1992), it creates interest among stakeholders through interactive group discussion (Hartwich and Oppen, 2000) and it is structured and has discriminating potential (Soam, 2004). It is widely used in prioritization studies under different areas like career choice (Canada et al., 1985), information system project selection, performance evaluation (Chan and Lynn, 1991), finance, marketing (Anderson et al., 2000) and industrial project selection (Dey,2002). But in agriculture sector particularly in India, still its use is very limited (Soam, 2004). AHP can be used as a multi-criteria multi-level decision making tool.

Procedure for Analytic Hierarchy Process (AHP)

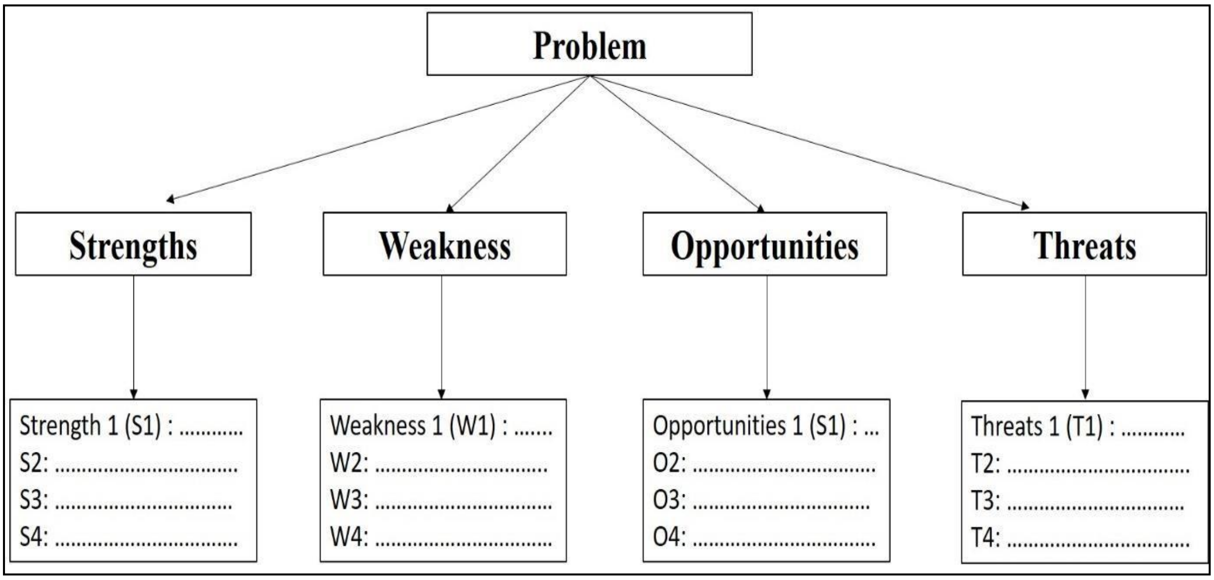
The first requirement of AHP is breaking up of decision problem into elements in hierarchial fashion. This method is based on pair-wise comparison but associated with hierarchic formulation of multi-criteria. This method has significant advantage of providing ‘objective decision’, based on subjective and personal preference of an individual or group of individuals. This method has the ability to make quantitative and qualitative decision attributes commensurable and has flexibility with regard to the setting objectives (Kangas, 1992). In the same decision analysis we can include expert knowledge, subjective preferences and objective information (Kurttila *et al*, 2000).

Step-wise Description of Application of Analytic Hierarchy Process

Step I: Problem Modelling

For doing AHP, a better understanding of the decision problem is necessary. The targeted goal should be clear along with various factors associated with it. After identifying the factors, different criteria have to be derived. These are the essential features of ‘problem modelling’ in AHP. The problem structure should include less number of criteria yet covering a vast aspect.

Example for problem modelling is shown below:


Problem
Strengths
Weakness
Opportunities
Threats
Strength 1 (S1) : ....
S2: ....
S3: ....
S4: ....
Weakness 1 (W1) : ....
W2: ....
W3: ....
W4: ....
Opportunities 1 (S1) : ...
O2: ....
O3: ....
O4: ....
Threats 1 (T1) : ....
T2: ....
T3: ....
T4: ....

Step II: Pair-wise comparison

During the next step, different factors are compared, pair-wise; and, in the next level, criteria are also compared, pair-wise, among themselves. Thus, based on comparison, in each case (both factor and criteria) the ‘comparison matrix’ can be formed.

$$A = \begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ 1/a_{12} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1/a_{1n} & 1/a_{2n} & \cdots & 1 \end{bmatrix}.$$

A= Pair-wise comparison matrix

Step III: Judgemental scale

The strength of AHP lies in the possibility to evaluate quantitative as well as qualitative criteria and alternatives on the same preference scale. Here, the response may be in numerical, verbal or in graphical form, but it can be converted into cardinal measurement. Saaty (2008) proposed one nine-point scale to have a pair-wise comparison of different factors and criteria. In Saaty’s AHP, the verbal statements are converted into integers from one to nine, based on the intensity of importance of one over other.

The fundamental scale of absolute numbers

Intensity of importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
2	Weak or Slight	
3	Moderate importance	Experience and judgement slightly favour one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgement strongly favour one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	An activity is favoured very strongly over another; its dominance demonstrated in practice
8	Very Very strong	
9	Extreme importance	The evidence favouring one activity over another is of the highest possible order of affirmation

(Adopted from Saaty, 2008)

Step IV: Aggregation of judgement

In AHP, several processes are used to aggregate the decision-makers opinions, with the two most popular being: (1) aggregating individual judgements regarding each set of pair-wise comparisons to produce an aggregate hierarchy; (2) synthesizing each of the individual hierarchies and aggregating the resulting priorities (Forman and Peniwati, 1998). These two processes are also termed the ‘Aggregation of individual judgements’ (AIJ), and the ‘Aggregation of Individual Priorities’ (AIP). Forman and Peniwati (1998) explained that the optimal mathematical procedure for aggregation depends on whether the group is assumed to be a synergistic unit or merely a collection of individuals. ‘Aggregating individual Judgement’ (AIJ) with geometric mean in case of former and ‘Aggregating Individual Priorities’(AIP) with either geometric mean or arithmetic mean should be used in the latter case, respectively. However Wu et al. (2008) compared different aggregation methods and categorically stated that methods of aggregation did not influence the final results. However, the judgements of every expert and then getting the arithmetic mean is inefficient. Here, AIP method was used by taking geometric mean, while calculating the priorities of SWOT factors, and arithmetic mean was calculated to aggregate the priorities among the criteria. By taking the derived value from AIP method, comparison matrix was developed for the group of individuals.

Step V: Determination of consistency ratio

As priorities only would be usable and valid, if derived from matrices that are consistent; so a consistency check must be applied. Saaty (1977) has proposed a Consistency Index (CI), which is related to the eigen value method. The 'Eigen value'(λmax) can be obtained by summing of products of each element of 'Eigen vector' multiplied by the total of columns of reciprocal matrix. He also proved that biggest 'Eigen value' is equal to the number of comparisons (λmax = n). Therefore 'Consistency Index' can be calculated by the following formula.

$$CI=\frac{(\lambda max-n)}{(n-1)}$$

Where n = Dimension of the matrix

λ_{max} = Maximal eigen value

The consistency ratio, the ratio of CI and RI, is given by:

CR=CI/RI

Where RI is the random index

Table of Random Indices

N	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

(Adopted from Saaty and Forman, 1992; actual calculation made by Saaty, 1977)

The thumb rule for consistency of the matrix is, consistency ratio should be less than 0.1

Step VI: Calculation of priorities

Identification of scaling factor is of utmost importance, as the scaling factor or priority would show how much importance a particular factor has, in terms of overall goal. Various criteria had local priority (priority/scaling factor within a particular factor) which told about the cardinal importance of each factor. Priority or local weights are calculated by dividing each element of row by the sum of each column in 'comparison matrix. Then normalize the 'Eigen vectors' by averaging the value of the factors/ criteria across the new rows to identify scaling factors or priority vector. The global or overall priority (priority/ scaling factor in relation to overall goal) is also worked out. To get global priority of criteria to overall goal, priority vector of factors was multiplied with local priorities of respective criteria within the particular factor. In this way, calculation of priority of factors to overall goal and local priority of criteria are achieved.

The final results after doing SWOT analysis with AHP can be presented in the following format:

SWOT Group	Priority of the Group (Scaling factor)	SWOT factors	Consistency Ratio (CR)	Priority of SWOT factors within SWOT Group	Global Priority of Factor
		S1:			
		S2:			
		S3:			
		S4:			
λ_{max} =………….. , CI =…………..					

(Source: Bhatt *et al*., 2019)

References

Anderson, D. D., Dennis, J., Sweeney, T. and Williams, A. (2000). An Introduction to Management Science: quantitative approach to decision making. South Western College Publishing, New York, pp. 715-735

Bhatt, A., Meena, B. S. and Paul, P. (2019). Draught Animal Power: Opportunities and challenges in mountain agriculture. *International Journal of Livestock Research*, 9(7): 127-134.

Canada, J. R., Frazelle, E.H., Koger R.K., and Mac Cormac, E. (1985). How to make a career choice: the use of Analytic Hierarchy Process, *Industrial Management*, 27(5): 16-22.

Chan, Y.L. and Lynn, B.E. (1991). Performance evaluation and the Analytic Hierarchy Process. *The Journal of Management Accounting Research*, 3: 57-87.

Dey, P.K. (2002). Application of analytic hierarchy process to benchmarking of project management performance: an application in the Caribbean public sector, *Vikalpa*, 27(2): 29-48.

Dyer, R. F. and Forman, E.H. (1992). Group decision support with the analytic hierarchy process. *Decision Support System*, 8: 99-124.

Forman, E. and Peniwati, K. (1998). Aggregating individual judgments and priorities with analytic hierarchy process. *European Journal of Operational Research*, 108: 165-169

Hartwich, F. and Oppen, M. von, (2000). The use of DEA in performance evaluation of agricultural research systems in Sub-Saharan Africa. International DEA Symposium: measurement and improvement of productivity in the 21st century, 3-5 July, Brisbane, Australia.

Madhavan, M.M. (2018). A Study on management of dairy animal waste and its effect on environment in urban and peri-urbanareas of National Capital Region (NCR), India. PhD Thesis *(Unpub.)*, National DairyResearch Institute (Deemed University) Karnal (Haryana), India**.**

Saaty, T.L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15: 234-281

Saaty, T. L. (1980). The Analytic Hierarchy Process. New York: McGraw-Hill.

Saaty, T. L. (2008). Decision-making with the analytic hierarchy process. *International Journal Service Sciences*, 1(1): 83-98.

Soam, S.K. (2004). Research project prioritization through training in analytic hierarchy process: case study of a village in semi-arid region of Central India. *Uganda Journal of Agricultural Sciences,* 9: 157-162.

Wu, C.R., Chang, C.W. and Lin, H.L. (2008). Comparing the aggregation methods in the analytic hierarchy process when uniform distribution. *Wseas Transactions on Business and Economics,* 5(3): 82-87.

Structural Break Analysis and its Application

P. Anbukkani and Haritha K.
Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: anbueconomic@gmail.com

Introduction

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

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

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

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

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

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

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

Testing for Structural Change of Unknown Timing

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

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

Chow Tests

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

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

Step 1: Define the null and alternative hypotheses.

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

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

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

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

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

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

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

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

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

Step 2: Calculate the test statistic.

If we define the following terms:

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

Then we can say that the Chow test statistic is:

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

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

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

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

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

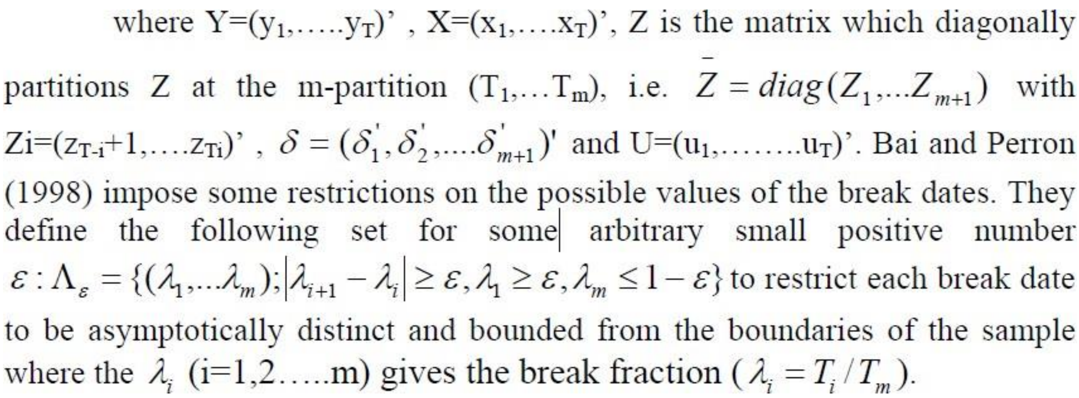
Bai-perron test

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

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

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

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



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

R code for Bai-Perron

```
library(strucchange)

maize_area_kr <- breakpoints (maize_area_kr~year, h=8, data=tsdata)
maize_area_kr
summary(maize_area_kr)
plot(maize_area_kr)

maize_area_mp <- breakpoints (maize_area_mp~year, h=8, data=tsdata)
maize_area_mp
summary(maize_area_mp)
plot(maize_area_mp)

maize_area_tn <- breakpoints (maize_area_tn~year, h=8, data=tsdata)
```

```
maize_area_tn
summary(maize_area_tn)
plot(maize_area_tn)
```

References

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

Andrews, Donald W.K. and Ray C. Fair. (1988). Inference in Nonlinear Econometric Models with Structural Change. *Review of Economic Studies*, 55(4): 615–623.

Bai, Jushan (1994). Least Squares Estimation of a Shift in Linear Processes. *Journal of Time Series Analysis*, 15(5): 453–472.

Bai, Jushan (1997a). Estimation of a Change Point in Multiple Regression Models. *Review of Economics and Statisticsm* 79(4): 551–563.

Bai, Jushan (1997b). Estimating Multiple Breaks One at a Time. *Econometric Theory*, 13(3): 315–352.

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

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

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

Breaks in Multivariate Time Series. *Review of Economic Studies.* 64(3): 395–432.

Chow, Gregory C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica.* 28(3): 591–605.

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

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

Hansen, Bruce E. (2000). Testing for Structural Change in Conditional Models. *Journal of Econometrics*, 97(1): 93–115.

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

Perron, Pierre (1997). Further Evidence on Breaking Trend Functions in Macroeconomic Variables. *Journal of Econometrics*, 80(2): 355–385.

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

Quandt, Richard (1960). Tests of the Hypothesis that a Linear Regression Obeys Two Separate Regimes. *Journal of the American Statistical Association*, 55: 324–330.

Training Manual

Advanced Research Methods and Essential Skills for Social Sciences

December 12- 22, 2022

Division of Agriculture Economics, ICAR-IARI, New Delhi

Course Director

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

Coordinators

R R Burman
Principal Scientist
Division of Agricultural Extension
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail: burman_extn@hotmail.com

Praveen K V
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail: veenkv@gmail.com

Asha Devi S S
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi-110012
E-mail:ash.nibha@gmail.com

Published by

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