Training Manual

Production and Market Trends Forecasting Using Statistical Modeling

Compiled and Edited by:

Dr. Md. Mosharraf Uddin Molla, MD (AERS), BARC
Dr. Md. Shofiqul Islam, PSO (AERS), BARC

January 2025

Agricultural Economics and Rural Sociology Division
Bangladesh Agricultural Research Council (BARC)

Training Manual

Production and Market Trends Forecasting Using Statistical Modeling

05-09 January, 2025
Venue: Computer Training Lab, AIC Building, BARC

Dr. Md. Mosharraf Uddin Molla
Course Director

Dr. Md. Shofiqul Islam
Course Coordinator

The Training Module is Designed for Scientists of NARS Institutes

Organized by

Agricultural Economics and Rural Sociology Division
Bangladesh Agricultural Research Council (BARC)

Program Schedule

Day/Date	Time	Topic/Event	Resource person
05-01-25 Sunday	9:00-9:20	Registration	--
	9:20-9:40	Pre-Evaluation	--
	9:40-10:15	Opening	--
	10:15-10.30	Tea Break	--
	10:30-11:20	Importance of forecasting in production and market trends in agriculture	Dr. Nazmun Nahar Karim, EC, BARC
	11:20-12:10	Fundamental of forecasting in agricultural research	Dr. S.M. Sayem
	12:10-1:00	Data management using STATA software	Dr. S.M. Sayem
	1:00-2:00	Scenario and pattern analysis of agricultural production and product market	Dr. S.M. Sayem
	2:00-3:00	Lunch	
	3:00-4:00	Scenario and pattern analysis of production and agricultural product market using STATA software	Dr. S.M. Sayem
	4.00-5.00	Trend analysis for future forecast	Dr. S.M. Sayem
06-01-25 Monday	9.00-10.00	Socio-economic research priorities in agriculture	Dr. Md. Mosharraf Uddin Molla
	10.00-10:30	Tea break	
	10:30-11:30	Hands-on exercise of trend analysis for agricultural production and price accuracy	Dr. S.M. Sayem
	11.30-12.30	Methods of model selection and forecasting	Dr. S.M. Sayem
	12.30-2.00	Lunch	
	2.00-3.00	Different techniques of stationarity checking	Dr. S.M. Sayem
	3.00-4.00	Hands-on exercise of stationarity checking using STATA software	Dr. S.M. Sayem
	4.00-5.00	Box-Jenkins methodology for forecasting crop and livestock production	Dr. S.M. Sayem
07-01-25 Tuesday	9.00-10.00	ARIMA model for forecasting crop/livestock production and price	Dr. S.M. Sayem
	10.00-10:30	Tea break	
	10:30-11:30	Practicing ARIMA model for forecasting crop production using STATA software	Dr. S.M. Sayem
	11.30-12.30	Use of SARIMA model for forecasting production and price	Dr. S.M. Sayem
	12.30-2.00	Lunch	
	2.00-3.00	Practicing SARIMA model for forecasting production and price using STATA software	Dr. S.M. Sayem
	3.00-4.00	Complex survey technique for agricultural production and market information	Dr. S.M. Sayem
	4.00-5.00	Maximizing efficiency: Use of forecasting methods in supply chain management	Dr. Md. Mosharraf Uddin Molla
08-01-25 Wednesday	9.00-10.00	Multiple and dynamic regression model for long term forecasting	Dr. Md. Akhtarul Alam
	10.00-10:30	Tea break	

Day/Date	Time	Topic/Event	Resource person
	10:30-11:30	Practicing dynamic regression model with real life data	Dr. Md. Akhtarul Alam
	11.30-12.30	Basics of autoregressive distributed lag model (ARDL)	Dr. Md. Akhtarul Alam
	12.30-2.00	Lunch	
	2.00-3.00	Practicing autoregressive distributed lag model (ARDL)	Dr. Md. Akhtarul Alam
	3.00-4.00	Hands-on exercise of VAR model using STATA software	Dr. Md. Akhtarul Alam
	4.00-5.00	Future research scope in socio-economic perspectives: Advancements and challenges in forecasting	Dr. Md. Shofiqul Islam
09-01-25 Thursday	9:00-10:00	Panel regression for fixed and random effect model	Dr. S.M. Sayem
	10:00-10.30	Tea break	
	10:30-11:30	Panel regression for mixed effect model	Dr. S.M. Sayem
	11:30-12:30	Application of panel regression analysis for forecasting agricultural product and price using STATA software	Dr. S.M. Sayem
	12:30-2:00	Lunch	
	2:00-3:00	Development of adaptive expectation model for prediction	Dr. Md. Abdus Salam
	3.00-4.00	Post evaluation	--
	4.00-5.00	Closing & certificate awarding ceremony	--

Resource Persons

1. Professor Dr. Mohammad Akhtarul Alam, Dept. of Agricultural and Applied Statistics, BAU, Mymensingh
2. Dr. Sheikh Mohammad Sayem, Associate Professor, Dept. of Agricultural and Applied Statistics, BAU, Mymensingh
3. Dr. Md. Fuad Hassan, Associate Professor, Dept. of Agricultural and Applied Statistics, BAU, Mymensingh
4. Dr. Md. Abdus Salam, Principal Scientific Officer, AERS Division, BARC
5. Dr. Md. Shofiqul Islam, Principal Scientific Officer, AERS Division, BARC

Contents

Fundamental of forecasting in agricultural research: Scenario and Pattern Analysis

Dr. Sheikh Mohammad Sayem
Associate Professor
Department of Agricultural and Applied Statistics
Bangladesh Agricultural University, Mymensingh

Lecture: Fundamental of forecasting in agricultural research

🕮 Forecasting

Forecasting is the prediction of values of a variable based on known past values of that variable or other related variables. Forecasting is the process of making informed estimations about future events based on the analysis of historical and current data.

In agricultural research, forecasting is crucial for anticipating future conditions that affect crop production, market trends, weather patterns, and resource management. It is very essential for the policy makers of a country.

🕮 Forecasting approaches

- **Qualitative methods:** Little or no quantitative information is available, but sufficient qualitative knowledge exists.
- **Quantitative methods:** Sufficient quantitative information is available.

🕮 Three conditions for applying quantitative forecasting

1. Information about past is available.
2. This information can be quantified in the form of numerical data.
3. It can be assumed that some aspects of the past pattern will continue in the future.

🕮 Time series data

A time series is a set of observations on the values that a variable takes at different times. Such data need to be collected at regular time interval. The time intervals can be annually, quarterly, daily, hourly, etc. For example, yearly potato production.

🕮 Cross-section data

Cross-section data are data on one or more variables collected at the same point in time. For example, amount of rice stock and whole sale price of rice in different district in September, 2023.

🕮 Panel data

Panel data set has both a cross-sectional and a time series dimension, where all cross section units are observed during the whole time period.

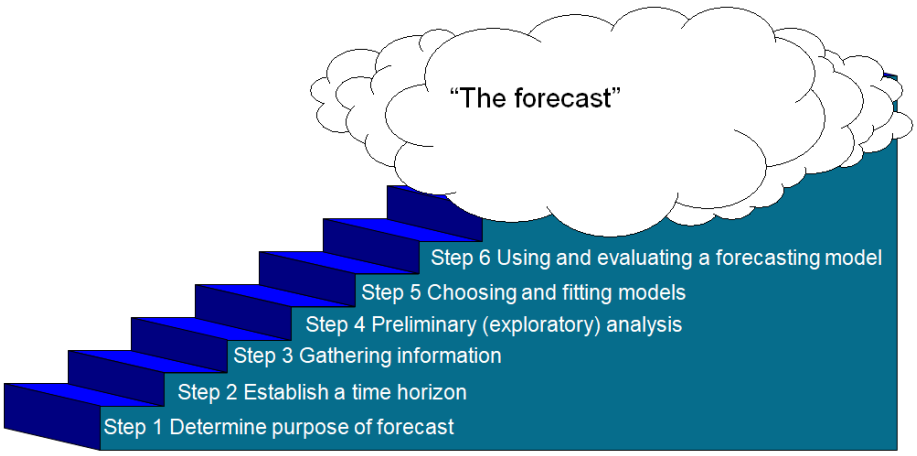
Time Series Data	
Year	Global Food Security Index in Bangladesh
2012	34.7
2013	35
2014	35.4
2015	36
2016	36.8

Cross-section Data		
Country	Global Food Security Index	Human Development Index
Bangladesh	34.7	0.515
India	48.7	0.554
Pakistan	43.7	0.515
Nepal	39.5	0.463
Srilanka	52.3	0.715

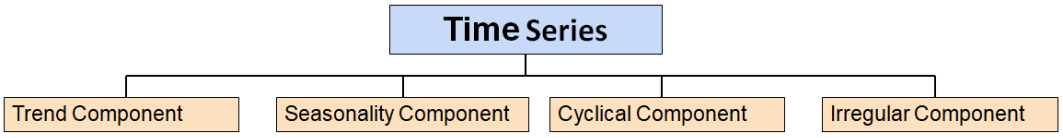
Example of Time Series, Cross-section, and Panel Data

Panel Data			
Year	Country	Global Food Security Index	Human Development Index
2012	Bangladesh	34.7	0.515
2013	Bangladesh	35	0.558
2014	Bangladesh	35.4	0.570
2012	India	48.7	0.554
2013	India	48.2	0.586
2014	India	47.9	0.609
2012	Nepal	39.5	0.463
2013	Nepal	38.2	0.540
2014	Nepal	40.7	0.548

🕮 Basic steps of forecasting

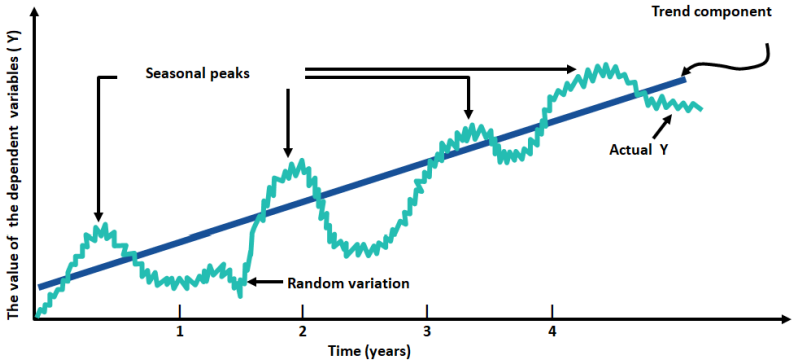


🕮 Time series components



- **Trend Component:** Long-run increase or decrease over time. Trend can be upward or downward, linear or non-linear.
- **Seasonal Component:** Short-term regular wave-like patterns. Need to observed within 1 year after monthly or quarterly.
- **Cyclical Component:** Long-term wave-link pattern.

- **Irregular Component:** Unsystematic, unpredictable and random.



🕮 Importance of forecasting in the agricultural sector

- **Enhanced Decision Making:** Forecasting provides data-driven insights that help farmers and stakeholders make informed decisions about planting, harvesting, and marketing.
- **Risk Management:** By predicting potential risks such as extreme weather events, pest outbreaks, and market fluctuations, forecasting enables proactive measures to mitigate adverse impacts.
- **Efficiency Improvement:** Forecasting helps optimize the use of resources such as seeds, fertilizers, and water, leading to cost savings and increased productivity.
- **Economic Stability:** Accurate market forecasts help farmers and traders plan their activities, stabilize income, and reduce the risks associated with price volatility.
- **Sustainable Agriculture:** Forecasting supports sustainable practices by anticipating the impact of environmental changes and guiding resource management strategies.

🕮 Scope of forecasting in agricultural sector

- **Crop Yield Prediction:** By analyzing historical yield data and weather patterns, forecasting models can predict future crop yields, helping farmers plan their production and manage supply chains.
- **Pest and Disease Outbreak Prediction:** Forecasting can identify the likelihood of pest invasions or disease outbreaks based on environmental conditions and historical occurrences, allowing for timely interventions.
- **Market Price Forecasting:** Predicting future prices of agricultural commodities helps farmers and traders make strategic decisions about planting, harvesting, and marketing.
- **Resource Management:** Forecasting helps optimize the use of resources like water, fertilizers, and pesticides, ensuring efficient and sustainable agricultural practices.

- **Climate Impact Assessment:** By forecasting the effects of climate change on agricultural systems, researchers can develop adaptation strategies to minimize negative impacts and enhance resilience.

🕮 Ethics of forecasting in agricultural research

- **Transparency:** Be open about the data sources, methodologies, and assumptions used in forecasting. This builds trust and allows stakeholders to understand and evaluate the forecasts.
- **Accuracy:** Strive for high accuracy in forecasts to avoid misleading stakeholders and causing unintended harm. Regularly update and refine models to improve accuracy.
- **Bias Reduction:** Ensure that data collection and model selection processes are free from biases that could skew results and disproportionately affect certain groups or regions.
- **Privacy:** Protect the privacy of data sources, especially when dealing with sensitive information from individual farmers or organizations. Implement robust data protection measures to prevent misuse.
- **Sustainability:** Promote sustainable agricultural practices by using forecasts to guide resource management and environmental conservation efforts.

🕮 Limitations of forecasting in agricultural research

- **Data Quality:** Incomplete, inaccurate, or inconsistent data can lead to unreliable forecasts. Data quality issues are common in regions with limited data collection infrastructure.
- **Model Limitations:** No single model can capture all the variables influencing agricultural outcomes. Combining multiple models or developing hybrid models can help mitigate this limitation.
- **External Factors:** Unpredictable events such as natural disasters, political changes, and market disruptions can impact forecast accuracy. Incorporating scenario analysis can help account for such uncertainties.
- **Resource Intensity:** Forecasting can be resource-intensive, requiring significant time, expertise, and computational power. This can be a barrier for small-scale farmers or research institutions with limited resources.

- **Adaptability:** Models need to be regularly updated and validated to adapt to changing conditions and new data sources. This ongoing process can be complex and time-consuming.

Lecture: Data Management using STATA
Dr. Sheikh Mohammad Sayem, BAU

Stata is a complete, integrated software package that provides all your data science needs—data manipulation, visualization, statistics, and automated reporting.

History of STATA

- The development of Stata began in 1984, initially by William (Bill) Gould and later by Sean Becketti. Stata was initially developed by Computing Resource Center in California and the first version was released in 1985.

- In 1993, the company moved to College Station, TX and was renamed Stata Corporation, now known as StataCorp.

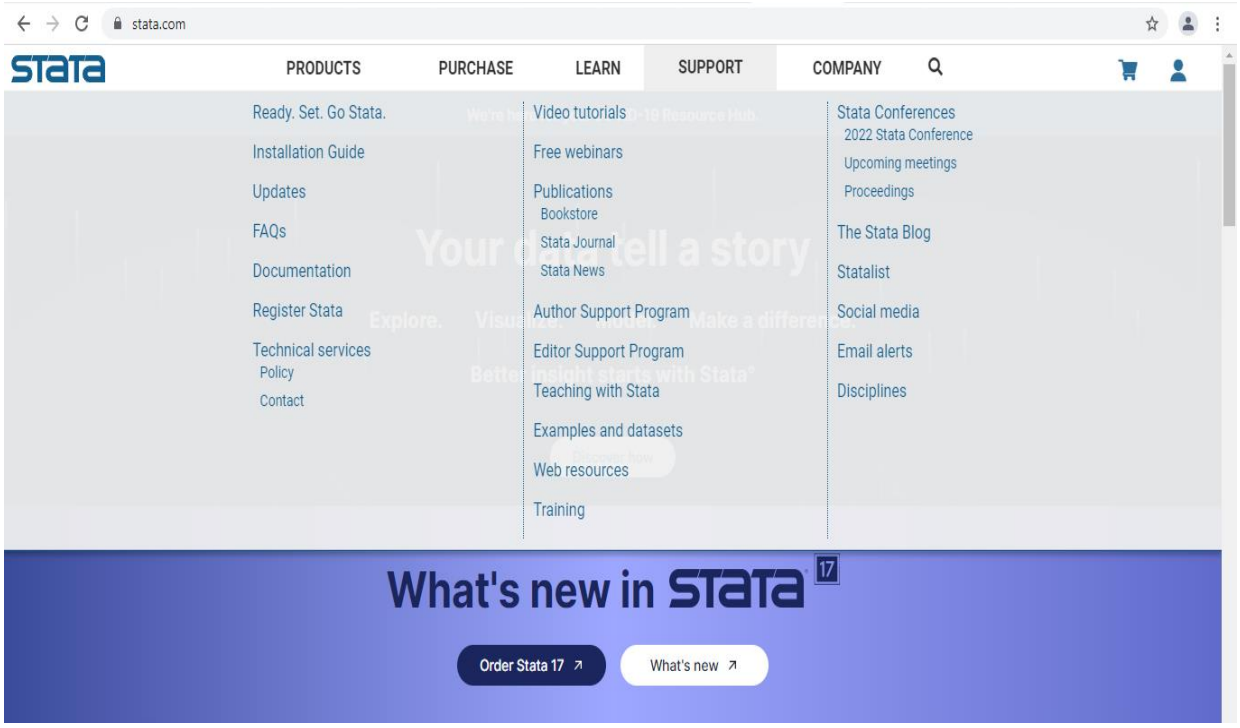
- There have been 17 major releases of Stata between 1985 and 2021. The first version was released in 1985 with 44 commands. Hundreds of commands have been added to Stata in its 36-year history.

Why STATA: Fast. Accurate. Easy to use.

Stata is a powerful statistical package with

- smart data-management facilities
- a wide array of up-to-date statistical techniques
- and an excellent system for producing publication-quality graphs.

Website: www.stata.com



🕮 Stata Interface

Once Stata is installed, it can be started from the Program Menu of your PC or by double-clicking the Stata shortcut that may be available on your desktop. The STATA interface has 5 windows-

✓ **Command Window**

– Where we will tell STATA what to do by typing commands

✓ **Review Window**

– Lists all the commands that you have already used.

– Allows us to easily repeat command by clicking on the right one

✓ **Results Window**

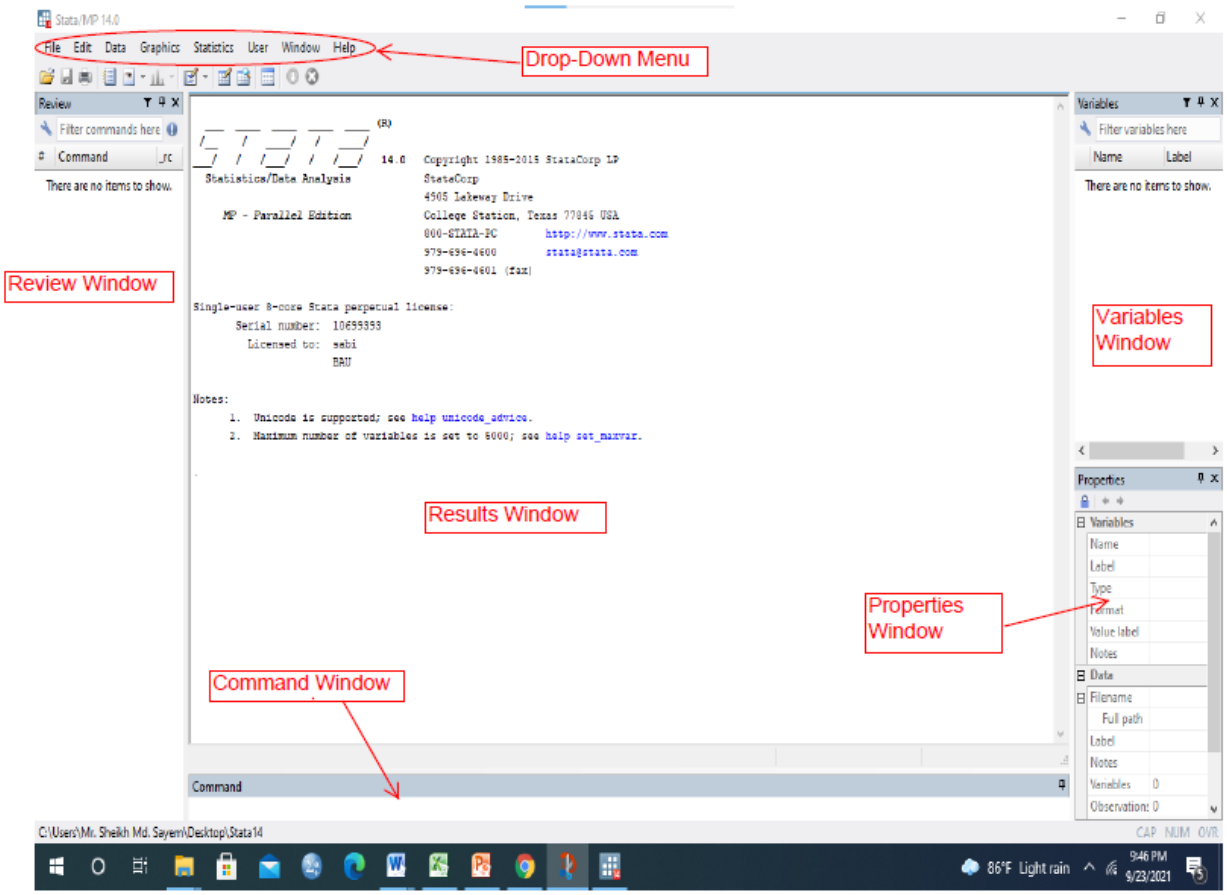
– Where all the output from our commands will appear

✓ **Variables Window**

– Lists the variables that are in your active data set. Any time a variable is selected, a small arrow will show up, and the user can click on the arrow to make the name of the variable appear in the command window.

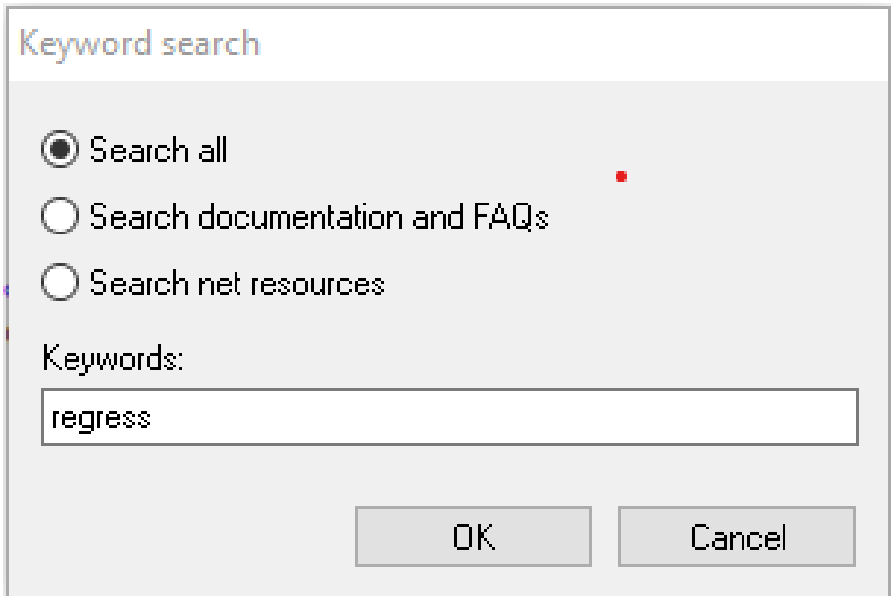
✓ **Properties Window**

– where details about the active data file (including information about each variable) are provided



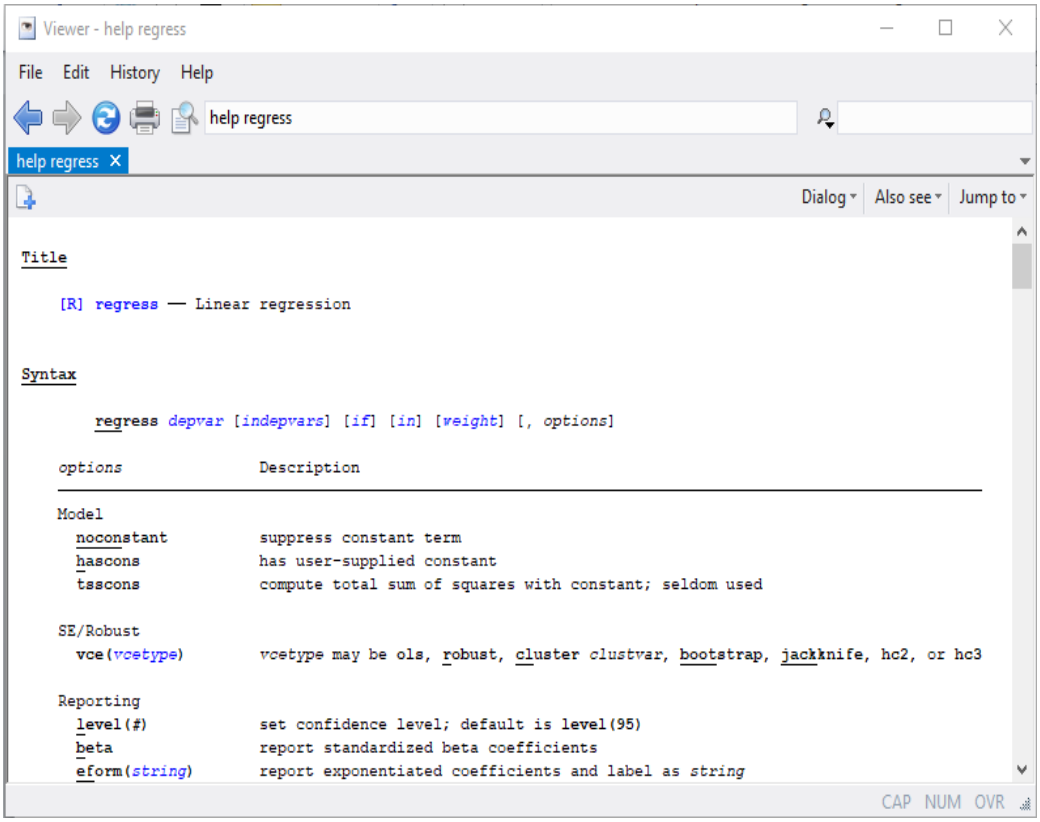
🕮 Stata Help

➢ ***Click Help → Click Search*** $\xrightarrow{\textit{after type keywords}}$ ***Click Ok***



[StataCommand "help [command name]": ***help regress***]

➢ The Viewer dialog box will be



➢ *These help files provide information about a specific command, guide you to general resources, and help you "search" for specific terms. The "search" is performed over STATA official documents and FAQs.*

???? Use Help command to search about graph command.

➢ The command *findit* is useful for broader searches and permits users to download user-written commands that are not part of the official STATA version.

[Stata Command "findit [command name]": ***findit regress***]

🕮 Measurement Scale

Measurement Scale	Properties	Example
Nominal	Name/ identification	Household Identification, Division Code
Ordinal	Data classification are ranked or ordered	Military rank, Professional Position

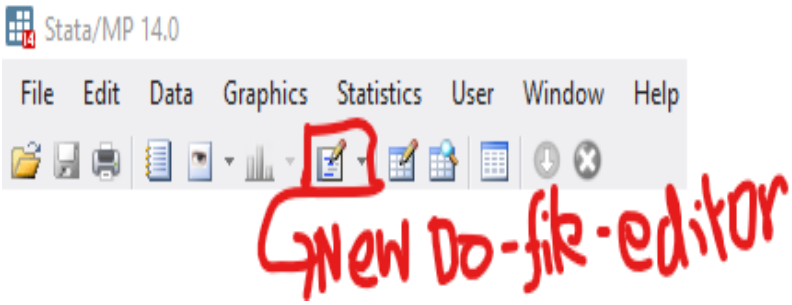
Interval	Has distance, no true zero, ratio is meaningful	Temperature, IQ score
Ratio	Ratio is meaningful, has true zero	Weight, Expenditure, GDP, Family Size

Do File

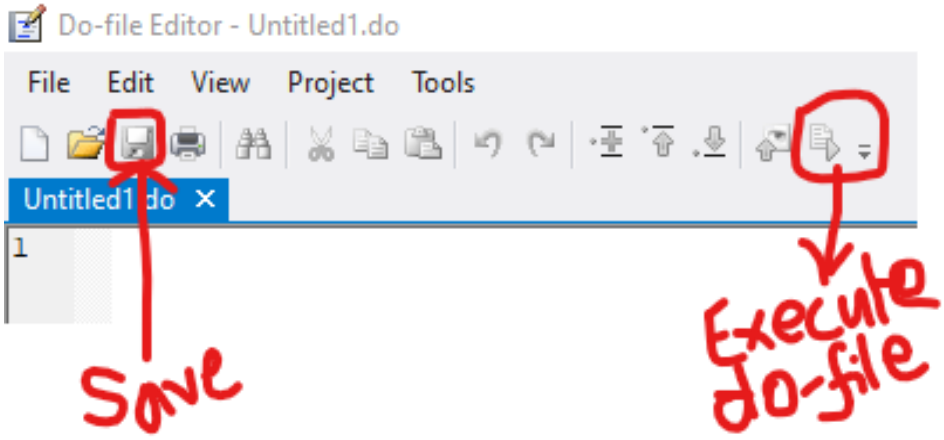
Do-file editor is an integrated text editor where do file contains a list of Stata commands. Stata will save and execute all the commands in do file.

🕮 Creating do files

Step 1: Open the Stata do-file editor. Click on the button shown below



The do-file editor should open in a new window, with a clean page looking something like this:



Step 2: Start typing your commands. I suggest starting with:

clear
set mem xxm
use file name
log using file name, text replace
This clears your workspace, frees up memory to speed the calculations, opens your data file, and opens a Stata log. (Be sure to end your do-file with log close.)

- Click on the execute do-file button in the do-file editor. Stata will execute all the commands in the do file. You will see all of these commands executed in the results window. If you create a log file, you can review all of your results at your leisure.

Important Issues:

1) You can continually update your do-file with additional commands. You can try your commands interactively, and if they seem to work, cut and paste them into the do file.
2) If you can put a * before a line in the do-file, Stata will not execute that line. This serves two different purposes. First, you can rerun your do-file while leaving out certain commands. (Just * the commands you want to skip.) Second, you can annotate your do file.
3) You can have Stata skip over several lines by using /* and */.If you have a long command that you need on separate lines, add /// at the end of each line. That tells Stata that the next line is part of the same command.
4) If there is a syntax error in your do-file, Stata will stop execution at the point of the error. You can go back to the do-file editor, correct the syntax error, and rerun your program.
5) You may want to create two do-files for any project. The first manipulates the data and creates new variables. At the end of this do-file, be sure to save the resulting data set in a new data file. The second file uses the data set you created in the first file to perform all of your analyses. This will save a lot of time, especially if you have a large data set that requires some time to get into shape prior to analysis.

Step 3: To save your do-file, you can either use the icon on the toolbar or use the "File"→"Save As" menu while the do-file editor is active.Let's suppose you name the file BBS Do File.do.

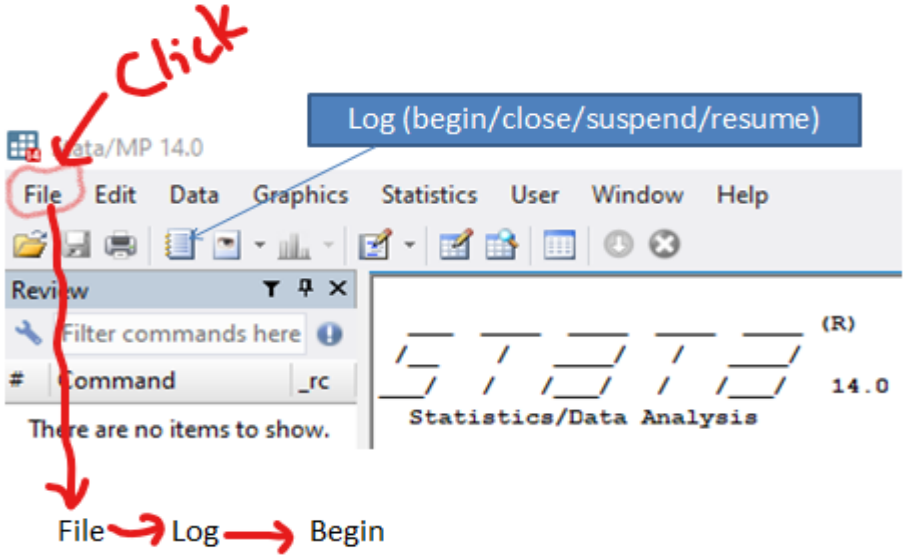
🕮 Common problems with do-files

- Forget to clear at the start of the file (far and away the biggest mistake)
- Forget to close the log file (a strong number two)
- Data sets and do-files are not in the same folder (relatively rare)
- Save a variable into a data set that already has that variable. Be sure to use replace rather than save
- Try to merge data sets but do not drop _merge before doing so.

🕮 Creating and Printing a Stata Log File

A Stata log file is an electronic record of all input (commands) and output (results with the exception of graph) presented in Stat's result window that can be printed or saved to a file.

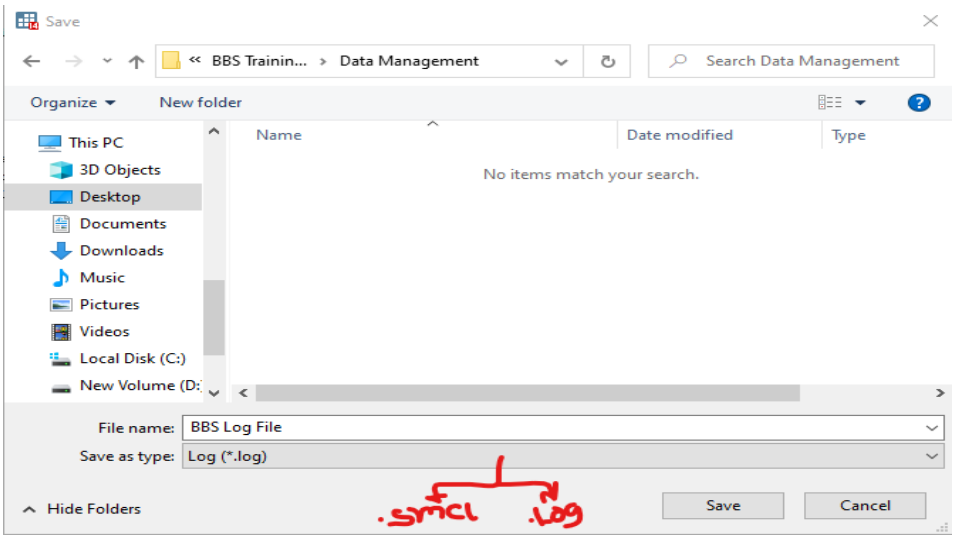
Step 1: Click on the Log button



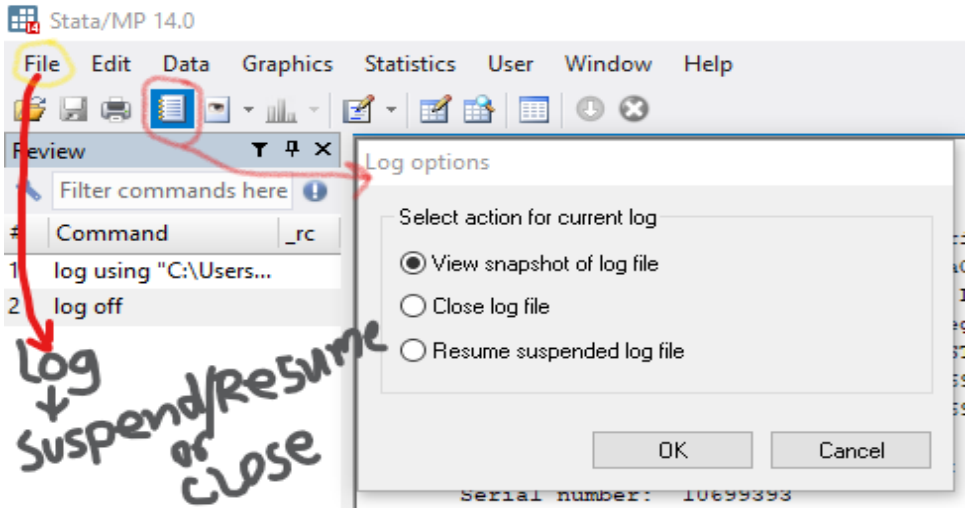
Step 2: If you do not have an active log (i.e. you have not already opened a log file), then Stata gives you the option to open one.

The default in Stata is to save the file with the extension smcl ("Stata Markup and Control Language file"). This will allow you to open the log file in Stata, but other programs will not read this type of file.

The other extension available is .log. This file format will allow you to open your log file in other programs and may be easier to manage than the smcl files.

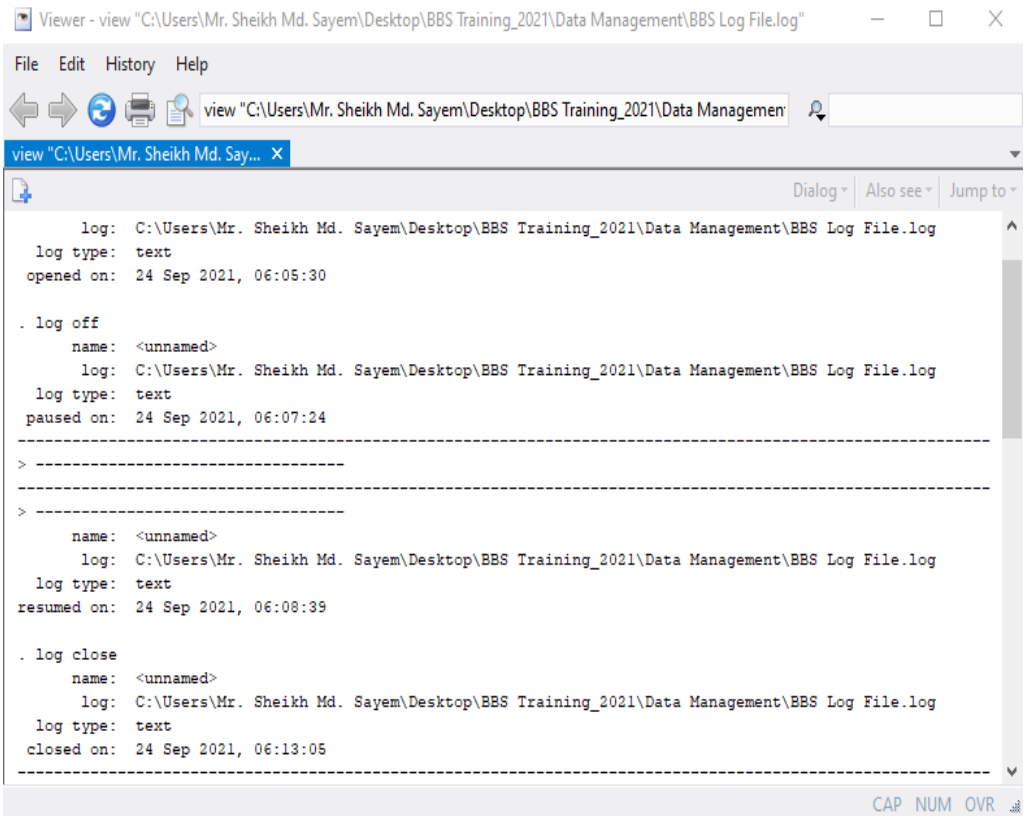


Step 3: If there is a currently active log, then when you click on the log button, it gives the option to view the log, close it, or suspend it. If you suspend it, the log remains open, but results will no longer be added to it until you choose to continue it.



You can suspend entries into the log by typing log off. You can resume entries by typing log on. You can close the log completely by typing log close or using the menus, click on File, Log, Close.

Step 4: When you want to view your log, you can click on the log button and select "view snapshot of log file."



To append material to a previously-written log file:

Way 1: you can type (Note: PC file path) ***log using"C:\Users\Mr. Sheikh Md. Sayem\Desktop\BBS Training_2021\Data Management\BBS Log File.log", append*** in the Stata Command window.

Way 2: Using the Stata menus, click on File, Log, Append, select the log file that you would like to add material to, and click OK.

Your new material will appear at the end of the previously-written log file.

- **Replace Existing Log File**

If you would like to replace an existing log file with a newer version add "replace" after the file name (Note: PC file path)

log using"C:\Users\Mr. Sheikh Md. Sayem\Desktop\BBS Training_2021\Data Management\tt Log File.log", replace

Alternately, you can use the menu: click on File, then on Log, then on Begin. Navigate to the folder where you want to save your log file, enter the file name, and specify whether you want to save the log as a .log or smcl file. Then click Save.

Either way, the previously-written file will be erased and overwritten with a log of your current Stata session.

- **Print Log File:**

First way: open the log file using MS Word and print it.
Second way: click on the File command, and select "print viewer". Note that after you do this, Stata will give you a series of options to select your printer, label the output, and even to adjust the font.

Import Data in Stata

Algorithm 1: Copy/Paste Data from Excel Data into Stata

Step 1: Open target Excel → Select and Copy Excel Data

Step 2: *Open Stata → Data → Data Editor → Edit → Paste*

Click (yes) Variables Name (to treat the first row as variables name)

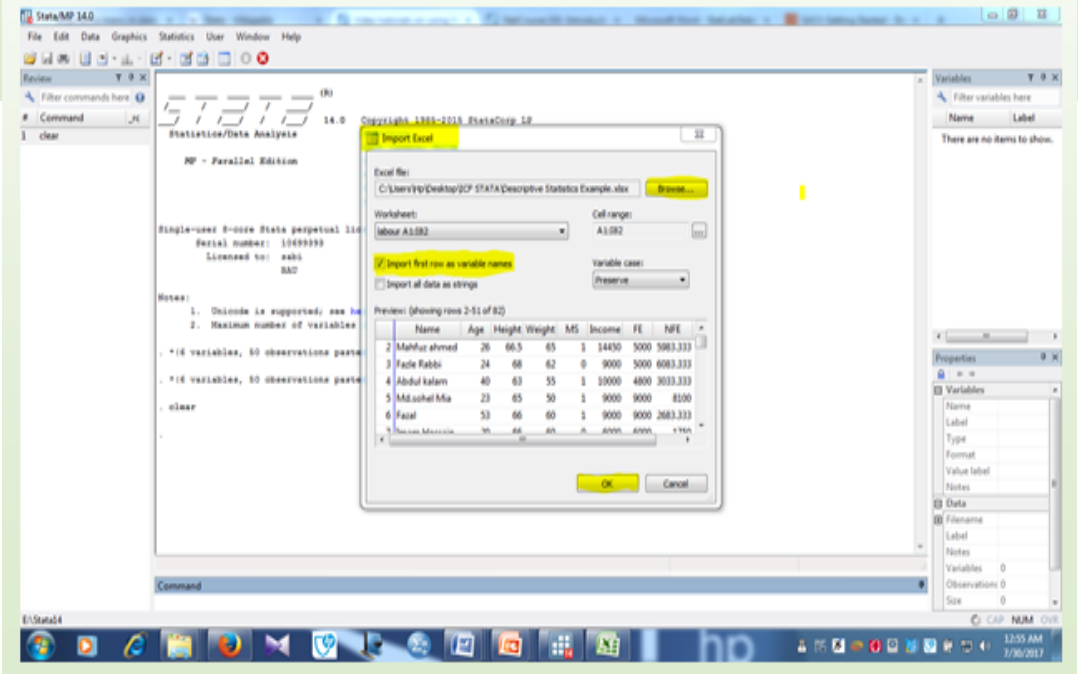
Algorithm 3: Import Excel Data into Stata

Step 1: *Open Stata → Click File → Import → Excel Spreadsheet → Import Excel Dialog Box*

Step 2:

Browse Target Excel File ⟹ Tick "Import first row as a variable"

Click OK



Algorithm 4: Export Stata data to Excel

Step 1: $File \rightarrow Export \rightarrow Data\ to\ Excel\ Spreadsheet$

Step 2: In “exportexcel-Export to Excel file” dialog box-

- Choose Excel filename
- Click Save variable names to first row in Excel file

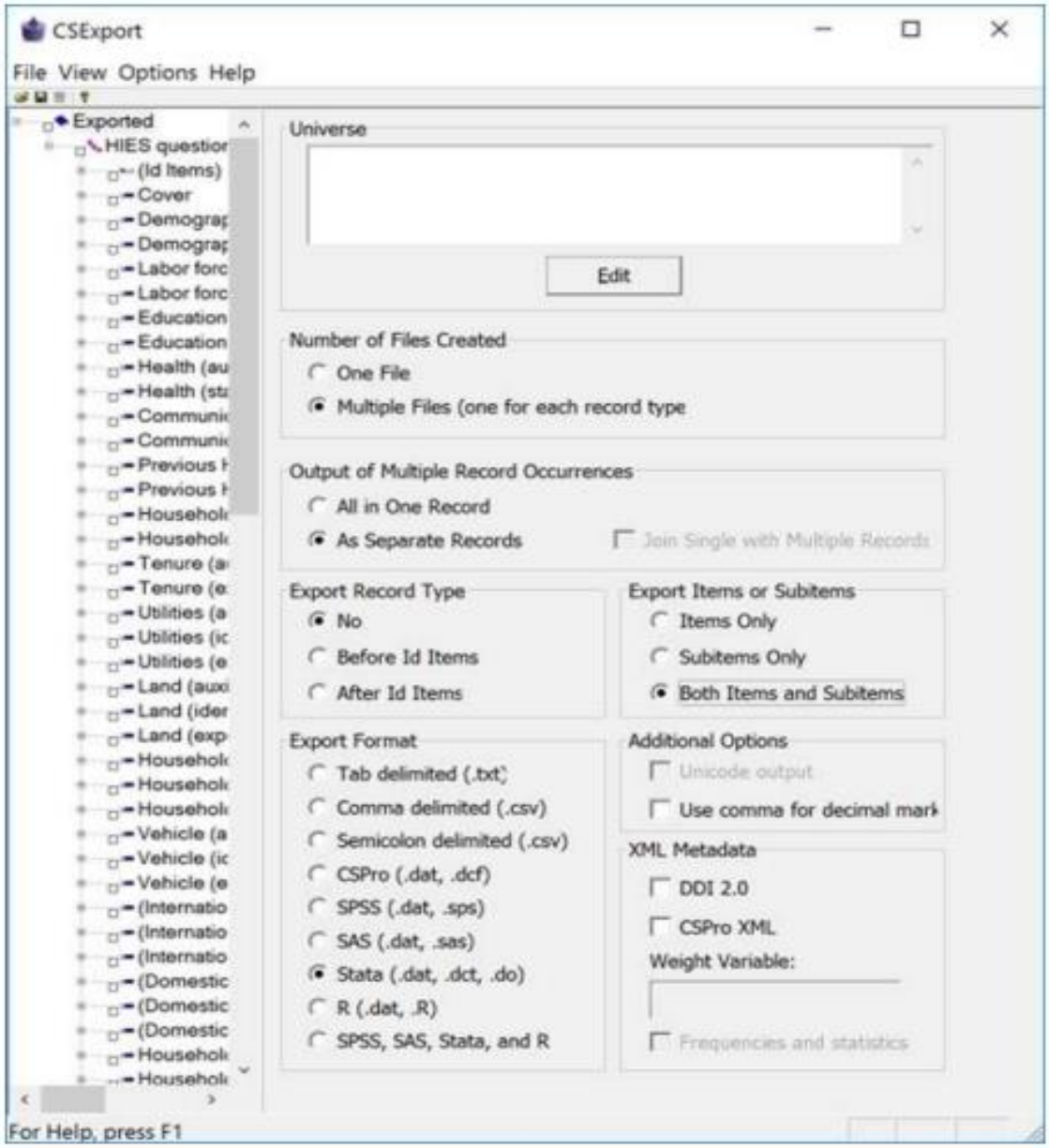
Click Ok

📖 **Importing Data from CsPro**

Export procedure of a CsPro dataset to Stata-

Step 1: Create a new folder in which you will save the exported materials.

Step 2: Open the CsPro data dictionary corresponding to the file to be exported, then select Tools > Export Data. The CsPro Export dialog box will be opened. Enter the options as shown in the screenshots below.



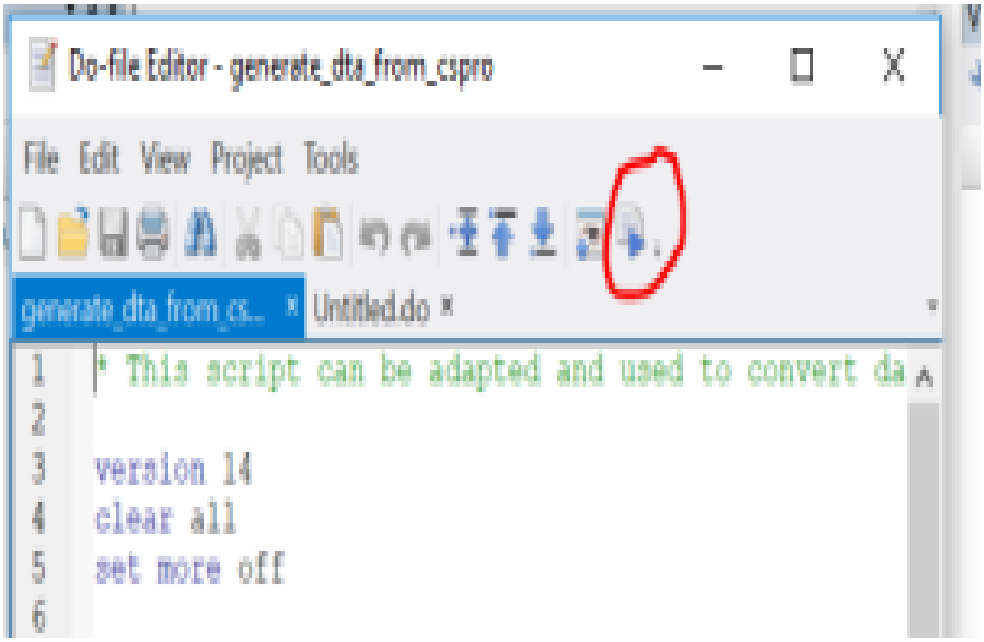
Step 3: CsPro will generate a collection of files (to be saved in the new folder). These files contain the materials needed to produce the Stata data files (not yet the Stata data files themselves).

You will have to run all [.do] files to produce the data files in Stata format, and save them.

CsPro export to Stata will generate, for each record type in the CsPro dictionary: One do file (extension DO) One dictionary file (extension DCT) One data file (extension DAT)

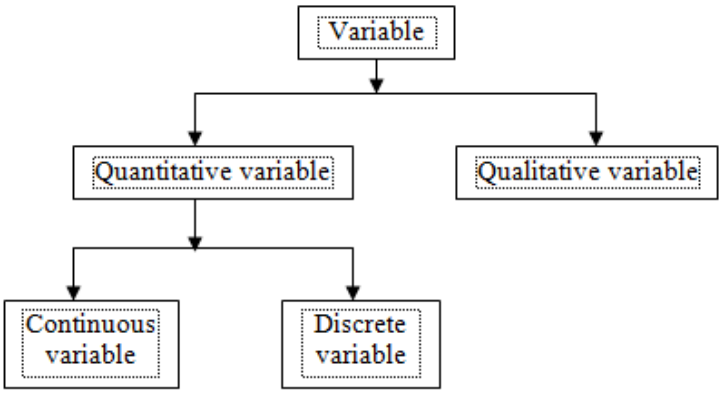
CsPro does not generate the Stata data files; it generates the materials needed to produce the Stata data files. This can involve executing many do files (one per record type). They can be run one by one, or a do file can be produced to run them in one batch.

Converting files one by one For each record type in the CsPro data dictionary, CsPro will have produced a DAT file (an ASCII fixed-format file containing the data for each specific record type), a DCT file that contains the information on the position of each variable in the DAT file and the variable and value labels, and a DO file that applies the DCT information to the DAT file. For each do file, you will have to execute (in STATA) the do-file:



🕮 Variable

Measurable characteristics of a population that may vary from element to element either in magnitude or in quality are called *variables*.

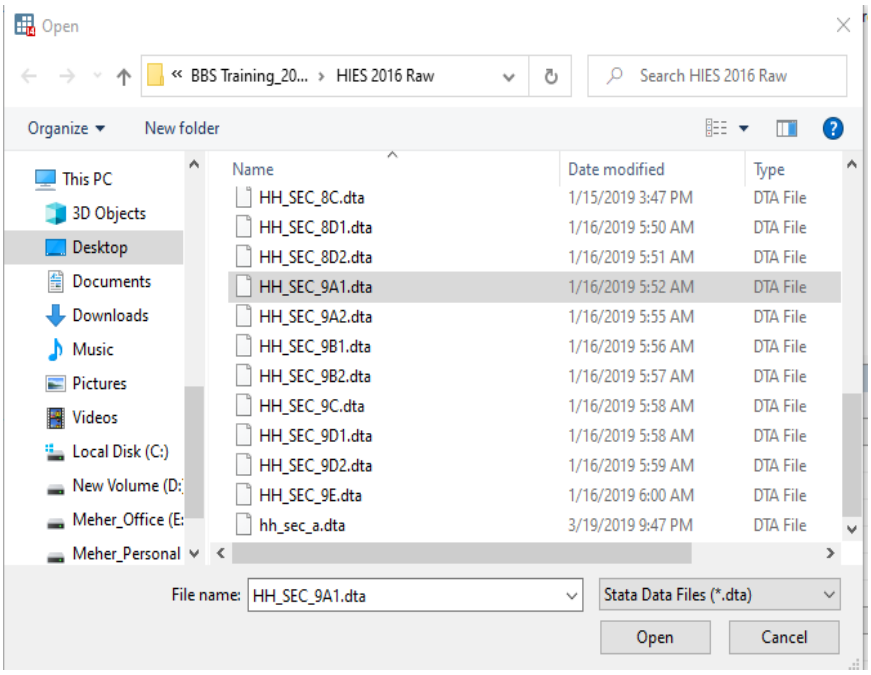


🕮 Generate New Variable

Algorithm 1: (Generating new variable by Stata)

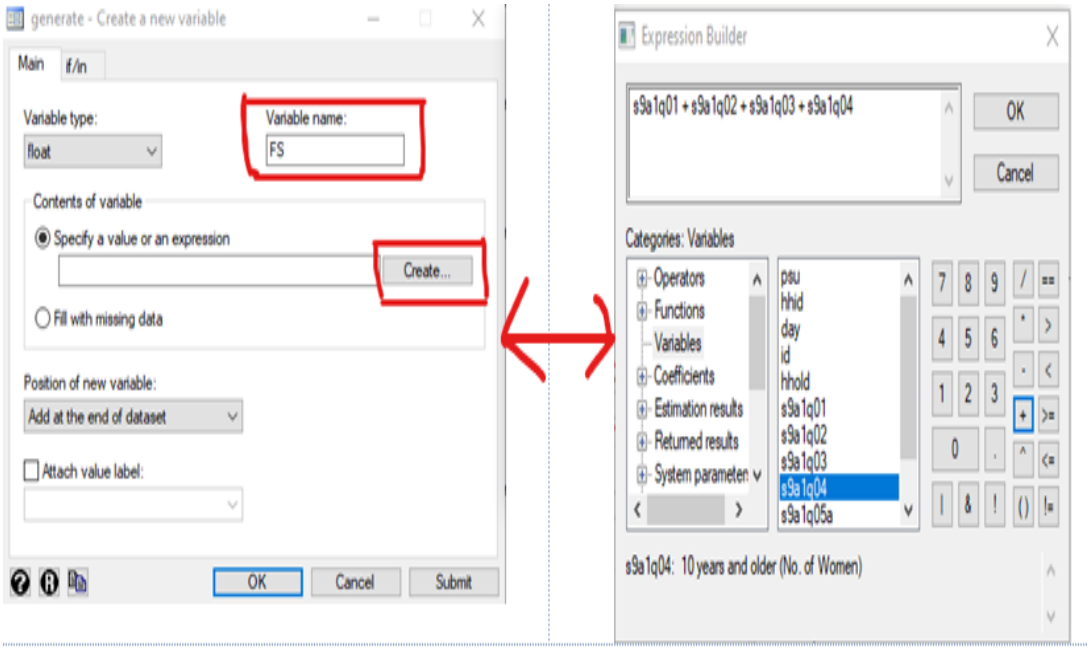
Step 1:

$$Open\ Stata \rightarrow File \rightarrow Open \rightarrow Select\ Stata\ dataset \rightarrow click\ Open$$



Step 2:

$$Data \xrightarrow{Click} Create\ or\ change\ data \rightarrow Create\ New\ Variable \rightarrow Click$$



In "generate…Create a new variable" dialog box-

- Fill up the variable name and click create in generate dialog box

- Build the expression and click Ok

  Click Ok for generating new variable

Dropping Variables

Select target variable from variable window → Click Drop selected variables → Click Yes

🕮 Drop or Keep Observation

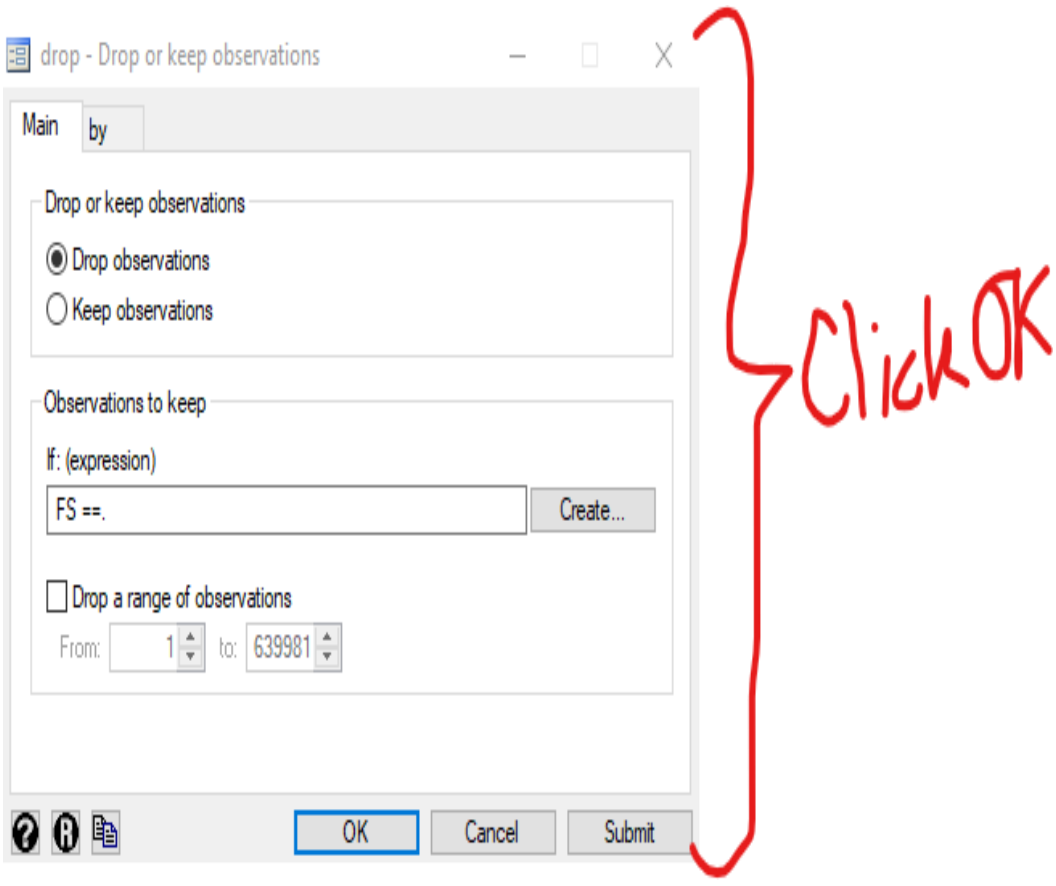
The **drop command** is used to remove variables or observations from the dataset in memory.

- If you want to drop variables, use drop varlist.
- If you want to drop observations, use drop with and if or an in qualifier or both.

The **keep command** is used to drop all variables except those specified explicitly or through the use of and if or inexpression.

- Just like drop, keep can be used with varlist or with qualifiers but not with both at once.

$$Data \xrightarrow{Click} Create\ or\ change\ variable \rightarrow Drop\ or\ Keep\ Obsevation \rightarrow Click$$



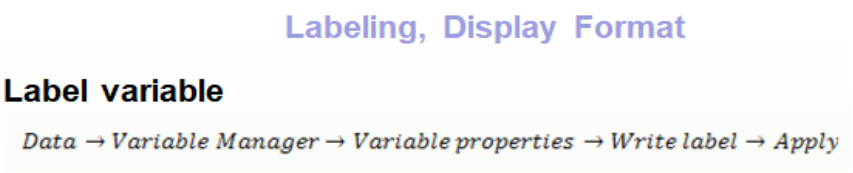
🕮 Typical Operator

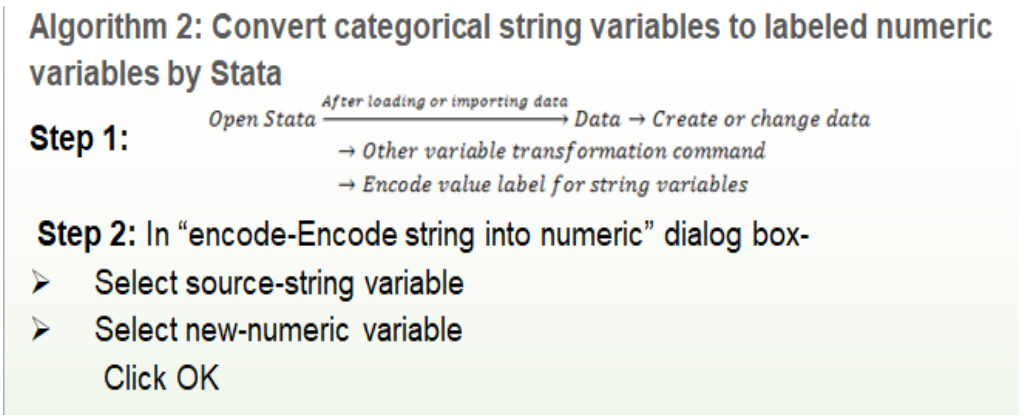
Arithmetic		Logical		Relational	
+	addition	&	and	>	greater than
-	subtraction	|	or	<	less than
*	multiplication	!	not	>=	> or equal
/	division	~	not	<=	< or equal
^	power			==	equal
-	negation			!=	not equal
+	string concatenation; e.g. "this" + "that" = "thisthat"			~=	not equal
*	string multiplication; e.g. 2*"this" = "thisthis"				

Missing values for a numeric variable are denoted by a dot (.), and string variables by two quotation marks ("").

🕮 Variables Manager

The Variables Manager is a tool for managing properties of variables both individually and in groups. It can be used to create variable and value labels, rename variables, change display formats, and manage notes. It has the ability to filter and group variables as well as to create variable lists. Users will find these features useful for managing large datasets.





Algorithm 3: Convert categorical string variables (Observational values are string) to labeled numeric variables by Stata

Step 1: *Open Stata* $\xrightarrow{\text{After loading or importing data}}$ *Data → Create or change data*
→ Other variable transformation command
→ Convert variables from string to numeric

Step 2:

In "destring-converting string to numeric" dialog box-
Select string variable
Create new-numeric variable name
Click OK

Algorithm 4: Create a categorical variable from a continuous variable by Stata

Step 1:

Data → Create or change data → Other variable transformation commands
→ Recode categorical variable

Step 2:

In "recode-Recode Categorical Variables" dialog box-

- Select variables
- Select and write required expression
- Select and write optional expression
- Click options
- Tick "Generate new variables"

Click OK

rule	Example	Meaning
(# = #)	(3 = 1)	3 recoded to 1
(# # = #)	(2 . = 9)	2 and . recoded to 9
(#/# = #)	(1/5 = 4)	1 through 5 recoded to 4
(nonmissing = #)	(nonmiss = 8)	all other nonmissing to 8
(missing = #)	(miss = 9)	all other missings to 9

Annex:

```
/* Create Categorical Variable from Continuous Variable*/
clear
set mem 400m
import excel "C:\Users\HP\Desktop\BBS Final Training Class_2021\Final BBS Training Data_26_9_21\Descriptive statistics 1.xls", sheet("Sheet1") firstrow
recode Age (15/20 = 1) (15/20 = 1) (21/25 = 2) (26/30 = 3) (31/35 = 4) (36/40 = 5), generate(newage)


/* Create string Obsevation to numeric*/

Clear
use "C:\Users\HP\Desktop\BBS Final Training Class_2021\Final BBS Training Data_26_9_21\string observation.dta", clear
destring age, generate(ds_age)
encode HealthStatus, generate(new_health) /*conver string variable to lebel numeric variable*/
/* Generate new variable*/

generate year_wage = Wage *12
drop if Wage >=60 /* Drop command for observation*/
```

keep if Wage >=55 /* keep command for observation*/

🕮 Collapse Data

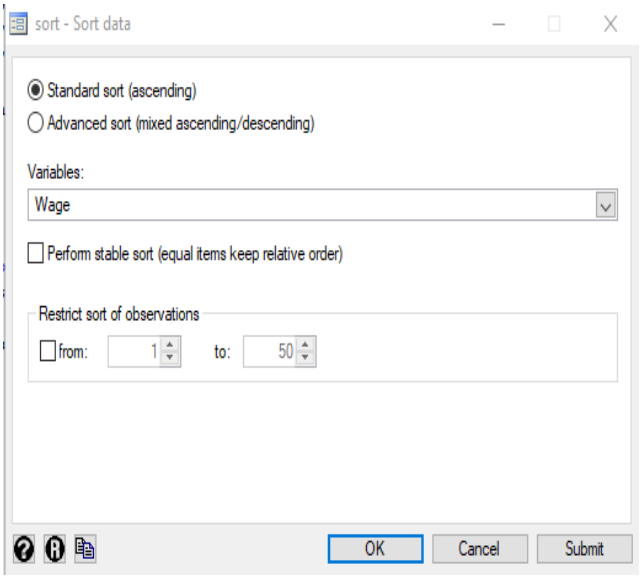
- Collapsing your data means to combine several cases into single lines.
- This command takes an open (or master) dataset and creates a new dataset by summarizing statistics on a selected variable.

Command: ***collapse (statistic) var1, by (var2)***

🕮 Sorting Data

Case 1:

$$Data \rightarrow Sort \xrightarrow{Click} Fill\ up \text{ "sort-Sort data" } dialog\ box$$

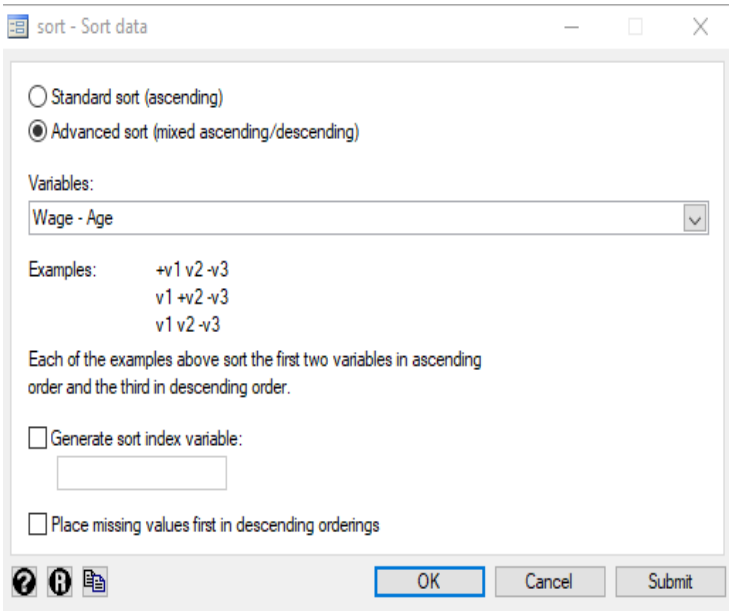


- The command [sort Wage] will sort your data according to the values in variable Wage, in ascending order.

You may insert more than one variable name; data will then be sorted first according to the first variable, and wherever there are several cases with the same value in this variable, these will be sorted according to the second variable, and so on.

Case 2:

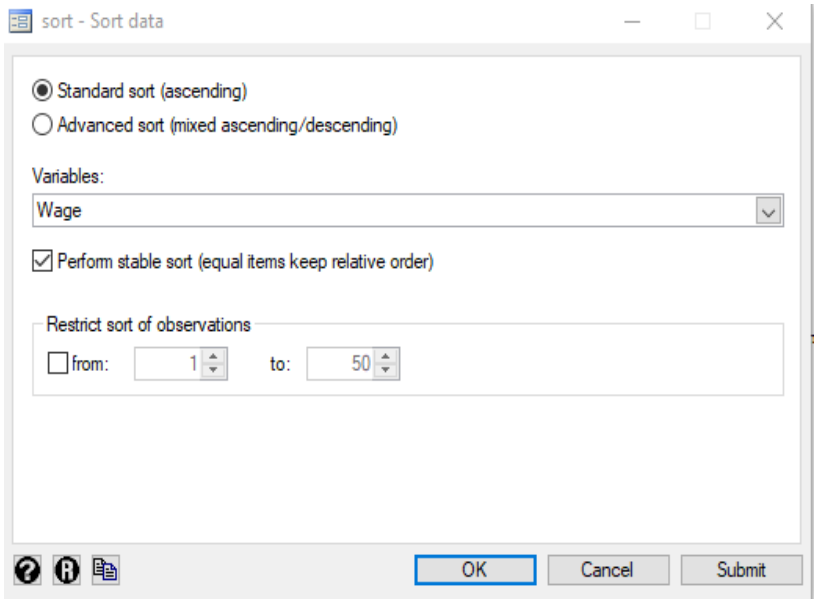
$$Data \rightarrow Sort \xrightarrow{Click} Fill\ up \text{ "sort-Sort data" } dialog\ box$$



- The command [gsort Wage - Age] will sort your data according to the values in variable wage, in ascending order, and within all same values of wage according to age, in descending order. That is, only gsort allows to sort data in descending order. Note that the plus sign may be omitted.

Case 3:

$$Data \rightarrow Sort \xrightarrow{Click} Fill\ up \text{ "sort-Sort data" } dialog\ box$$



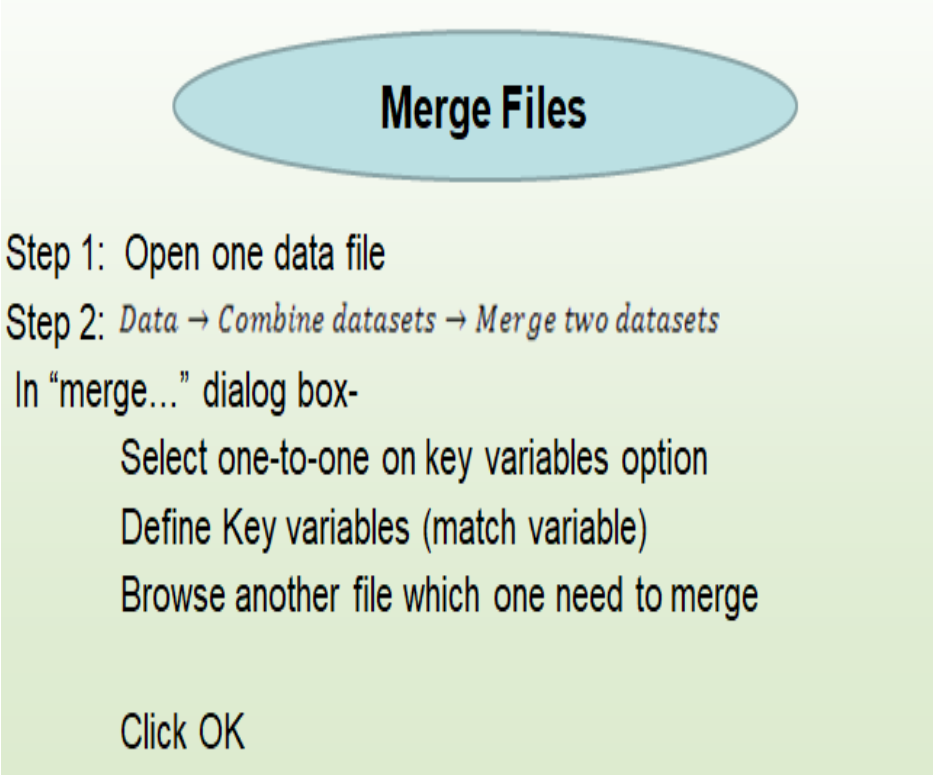
The command [sort Wage, stable]: Assume there are several cases with the same value in wage. Using the option stable will make Stata keep the order of cases within the same value of wage

after sorting (that is, the first value with a given wage in the original data will also be the first case in the sorted data, and so on). Otherwise, the order in which cases with the same value in the sorting variable appear is subject to chance.

🕮 Combining Datasets

When merging datasets, you will try to match different information about the same cases; information that for some reason or other is stored in more than one data set (e.g., because one part of the information was collected earlier on and additional information has been obtained later).

The challenge, of course, is to merge the datasets in such a way that information about individual A in dataset X is matched to the information about the same individual in dataset Y, and so on. The prerequisite for accomplishing this is to have in each of the datasets involved one or several variables that uniquely identifies or identify each case; such variables are called "key variable".



🕮 One-to-one merging

Command:

merge 1:1 varlist using filename

Here, "using dataset" is merged to the data in memory (called the "master dataset"), assuming that each value of key variable is present only once in each of the data sets. Not each caseid in

one dataset has to have an equivalent in the other dataset, but unless there is a certain amount of overlap you will not be inclined to merge data.

🕮 Many-to-one & one-to-many merge

Command:

merge m:1 varlist using filename

One assumes that for each value of *key variable* in the using dataset there may be more than one case with this *key variable* in the master dataset. The information from a case in the using dataset will be merged to *each* case in the master dataset with the same *key variable*. This is called a "many-to-one merge".

merge 1:m varlist using filename

This command will perform a "one-to-many merge", assuming that a *key variable* may occur several times in the using dataset, but only once in the master dataset. This is the same pattern as before, the only difference being that the currently active dataset (the master dataset) contains information about the higher level units and the lower level information (e.g., individuals) is found in the "using dataset".

merge m:m varlist using filename

This command will perform a "many-to-many merge", but there are few occasions in which this could be useful.

🕮 Retaining only selected variables from the "using" file

If you wish to add only a subset of the variables that can be found in "using" file, you can tell Stata via option `keepusing`, as in the following example-

merge 1:1 varlist using filename, keepusing (v17 v22 v25)

🕮 Variable "_merge"

With each merge, a variable called `_merge` is added to the dataset, indicating the status of each case. Possible values of `_merge` in the standard case are:

- _merge = 1: A case that was present in the master dataset only.
- _merge = 2: A case that was present in the using dataset only.
- _merge = 3: A case that was present in both datasets.

Note that the variable `_merge` has to be deleted before a new merge is performed; otherwise, an error message will result.

🕮 Append Command

Adding cases / observations

The append command combines the dataset in memory, known as the master dataset, with a dataset on disk, known as the using dataset. Typically, a user would implement the append command when they would like to add observations to an existing dataset with the same or similar variables.

Command: *append using dataset.dta*

Exercise: *Consider the HIES dataset 2018 for creating, cleaning, sorting, merge with other variables.*

Requirement:

1. How to drop the variable id_02_name from HIES module "hh_sec_a.dta"?

2. How to get and sort total main food expenditure of HH from HIES module "HH_SEC_9A2.dta"?

3. How to merge the file "hh_sec_a.dta" with "HH_SEC_9A2.dta".

4. How to keep only the Household of Dhaka division for "hh_sec_a.dta" dataset.

Lecture: Scenario and pattern analysis of agricultural production and product market

Dr. Sheikh Mohammad Sayem, BAU

🕮 Descriptive Statistics

Descriptive statistics are the methods of collecting, organizing, summarizing and presenting data in an informative way.

🕮 Line Graph

The line graph is particularly useful for numerical data where a series of data plotted at various time intervals.So, line graph are particularly effective for time series data because we can show the change or trends in a variable over time.

Problem (1): The following data represent the minimum temperature (in Celsius) recorded at Dhaka Meteorological station in different years:

Year	:	1994	1995	1996	1997	1998
Temperature	:	12.5	11.3	21.5	11.5	22.0

Algorithm 1: (Time series plot/Line Graph by Stata)

Step 1:

$$Open\ Stata \rightarrow File \rightarrow \text{Im}\,port\ Data \rightarrow Excel\ Spreadsheet \rightarrow click\,or\,click\,Enter$$

Step 2: In Import Excel dialog box

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step3:

$$Statistics \rightarrow Time\ series \rightarrow Declare\ dataset\ to\ be\ time\ series\ data \rightarrow click$$

In "tsset-declare dataset to be time series data" dialog box-

- Select Time variable

Click OK

Step 4: (Time Series Plot)

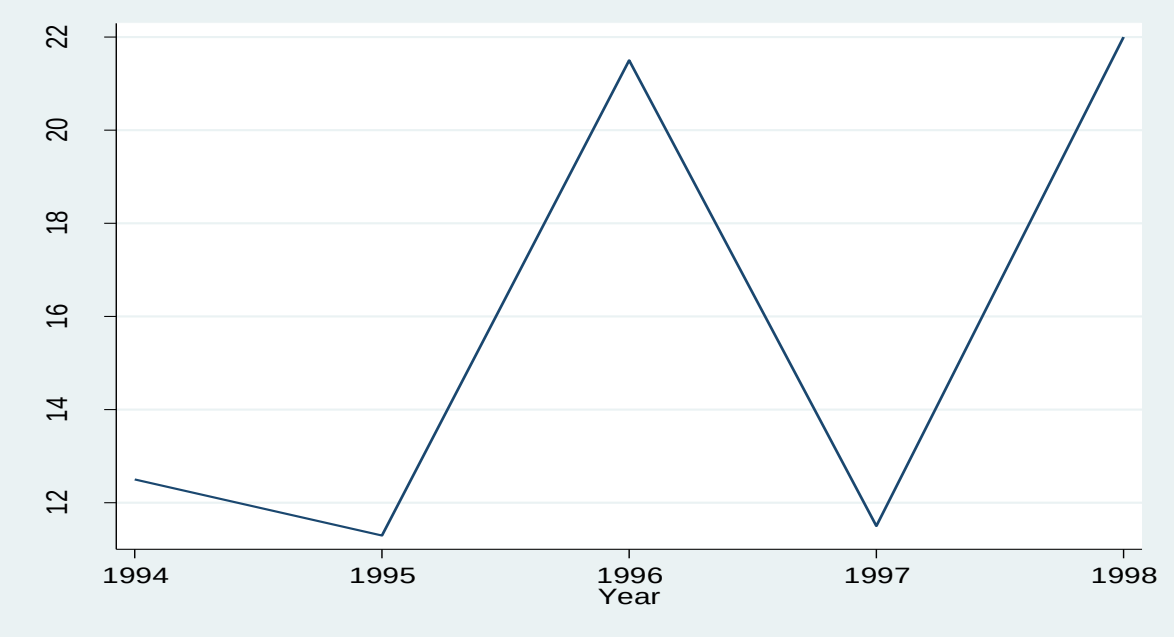
In plot-

$$Statistics \rightarrow Time\ series \rightarrow Graphs \rightarrow Line\ plots \rightarrow Click$$

- Click create
- Select Y variable
- Click Accept

Click OK

Figure 1: The line graph of temperature



Task 4: The line graph representing the minimum temperature in different years recorded in Dhaka Meteorological Station. We observed that the minimum temperature exist in 1995.

🕮 Measures of Center Point

A measure of center point is a single value that attempts to describe a set of data by identifying the central value or central position within that set of data.

Different measures of central point

- ✓ Arithmetic mean
- ✓ Geometric mean
- ✓ Harmonic mean
- ✓ Median
- ✓ Mode

Arithmetic Mean is the best average among them.

🕮 Measures of Dispersion

The variation of observations from their average is called the dispersion.

Different measures of dispersion

- ✓ Range
- ✓ Quartile deviation
- ✓ Mean deviation

- ✓ Standard deviation

Standard deviation is the most widely used measure of dispersion.

🕮 Coefficient of Variation

The coefficient of variation, CV is a measured of relative dispersion that expresses the standard deviation as a percentage of the mean.

The population coefficient of variation is-

$$CV = \frac{\sigma}{\mu} \times 100$$

The sample coefficient of variation is-

$$CV = \frac{SD}{\bar{X}} \times 100$$

The coefficient of variation is helpful in comparing the relative variation in several data sets that have different means and different standard deviations.

A distribution with smaller CV is said to be more homogeneous or uniform or consistent than the other.

🕮 Standard Error

The standard deviation of the sampling distribution of a statistic is known as its standard error (SE). The standard error of the sample mean can be computed by

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

And it is estimated by $\frac{s}{\sqrt{n}}$, where *s* represents sample standard deviation.

🕮 Skewness

Skewness means lack of symmetry. A distribution is said to be symmetrical when the values are uniformly distributed around the mean.

🕮 Kurtosis

A measure of the peakedness or convexity of a curve is known as **kurtosis**.

Problem 1: The maximum temperature for Dhaka, Rajshahi, Khulna and Chittagong stations during the period 1st January, 2016 to 31st January, 2016 are collected from Bangladesh Meteorological Division.

The data are given below-

Day	Dhaka	Rajshahi	Khulna	Chittagong	Day	Dhaka	Rajshahi	Khulna	Chittagong
1	27.6	26.6	27.6	26.2	17	26.7	25.7	26	27.3
2	27	26.8	26.8	26.4	18	25.7	24.4	25.3	27.5
3	25.8	26	26.5	26	19	25.4	22.3	26.8	27.5
4	26.5	25.4	26.8	25.5	20	22.8	17.8	25	27.4
5	25.8	24.5	26	25.6	21	22.7	20.8	23	21
6	26	25.7	26.5	26.6	22	20	18.3	21	23.8
7	26.6	25.8	26.7	26.2	23	20.8	21.5	21.8	22
8	26.8	28.3	28.5	26.1	24	22	19	22.8	23.8
9	25.2	26.8	29	27.5	25	23.1	21	23.5	23.6
10	24.4	25.5	26.5	28	26	22.1	21.8	24.2	24
11	25.1	24.3	26.6	25.4	27	25	24	24.5	24.7
12	24.5	24	25.2	26.3	28	25.8	25.3	26.5	24.9
13	24.1	24.7	25	25	29	25.2	24	26	25.4
14	25.4	25.8	25.5	24.8	30	24.4	24.1	27.6	25
15	26	26.2	26.8	25.2	31	26	26.4	29	27.5
16	26.6	25	26.8	27.5					

📖 **Algorithm 1: (Summary statistics by different categories using Stata)**

Step 1:

Open Stata → File → Import Data → Excel Spread sheet → click or click Enter

Step 2: In Import Excel dialog box

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step 3:

Statistics → Summarizes, tables, and tests → Other Tables → Compact table of summary statistics

In "tabstat-compact table of summary statistics" dialog box-

- ✓ Select variables
- ✓ Tick and select group statistics by variable
- ✓ Tick statistics to display (say, mean, standard deviation, variance, coefficient of variation)

Click OK

The Stata results are given below:

Table: Comparative analysis among the temperature of two stations

Station	**Average**	**Standard Deviation (SD)**	**Coefficient of Variation (CV)**
Khulna	25.81	1.922	7.446726
Chittagong	25.68	1.72	6.697819

Comment: A distribution with smaller CV is said to be more homogeneous or uniform or consistent than the other. So, the temperature value of Khulna station fluctuates more from period to period than does that of Chittagong station

Lecture: Trend analysis for future forecast
Dr. Sheikh Mohammad Sayem, BAU

🕮 Polynomial Regression Analysis

Polynomial regression is a form of linear regression in which the relationship between the independent variable (X) and the dependent variable is modeled as an n^{th} order polynomial in X.

The n^{th} order polynomial regression model is-

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_n X_i^n + \varepsilon_i \qquad i = 1,2,\ldots,N$$

The simplest form of second order polynomial regression model is-

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \qquad i = 1,2,\ldots,N$$

Which is also called quadratic regression analysis.

🕮 Main Features of the model:

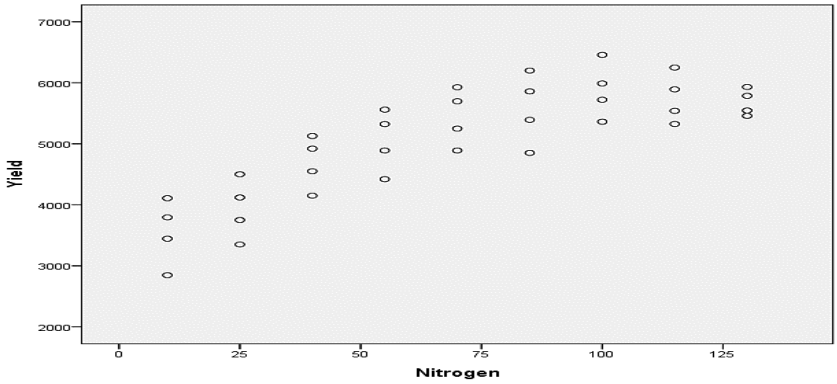
1. It is linear in the parameters but nonlinear in the variables.
2. There is only one independent variable on the right-hand side but it appears with various powers.
3. The rate of change in the dependent variable for a unit change in the independent variable is $\hat{\beta}_1 + 2\hat{\beta}_2 X$ which are dependent on X i.e. on the values of the independent variable.
4. The maximum value of Y occurs at

$$X = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}.$$

Problem (3): Let us consider the following data:

Yield (kg/ha) of a rice variety with different dozes of Nitrogen						
Nitrogen level	Replication				Total	Mean yield
	R1	R2	R3	R4		
10	2846	3794	4108	3444	14192	3548.0
25	3350	4500	4120	3750	15720	3930.0
40	4550	5128	4150	4920	18748	4687.0
55	4420	5560	5323	4890	20193	5048.3
70	5928	5698	5250	4890	21766	5441.5
85	4850	6200	5860	5393	22303	5575.8
100	5990	5362	6458	5722	23532	5883.0
115	5325	5893	6250	5540	23008	5752.0
130	5458	5546	5786	5932	22722	5680.5
Grand total and grand mean					182184	5060.7

The scatter plot of the data shows that the overall trend of yield increases up to a certain level of nitrogen but after that point it decreases. This type of relationship is better represented by a second-degree polynomial.



Polynomial Regression Analysis

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.886	.785	.772	430.470
The independent variable is Nitrogen.			

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	22376224.467	2	11188112.233	60.377	.000
Residual	6115047.533	33	185304.471		
Total	28491272.000	35			
The independent variable is Nitrogen.					

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Nitrogen	51.001	7.853	2.220	6.495	.000
Nitrogen ** 2	-.230	.055	-1.443	-4.220	.000
(Constant)	2962.869	237.292		12.486	.000

Explanation of the output:

1. R^2 is very high (0.785). The model and also the individual coefficients are highly significant. So the model fitness is good.
2. The estimated model is $Y = 2962.869 + 51.001X - 0.23X^2$.

3. The rate of change in Y for a unit change in X is

$$\hat{\beta}_1 + 2\hat{\beta}_2 X = 51.001 - 2 \times 0.23X = 51.001 - 0.46X$$

The rate of change in yield for a unit change in nitrogen level varies with the levels of nitrogen. The following table shows the different rates at the given levels of nitrogen:

Table: Changes in yield for a unit change in nitrogen at various levels of nitrogen.

Nitrogen level (kg/ha)	Change in yield (kg/ha)
10	46.40
25	39.50
40	32.60
55	25.70
70	18.80
85	11.90
100	5.00
115	-1.90
130	-8.80

4. The level of nitrogen at which yield is maximum is

$$X = \frac{-\hat{\beta}_1}{2\hat{\beta}_2} = \frac{-51}{2 \times (-0.23)} = \frac{-51}{-0.46} = 110.87\, kg/ha$$

🕮 Trend Analysis

Trend analysis is a special form of simple regression in which time is the explanatory variable. Trend analysis is a statistical technique used to identify and examine patterns and trends in a data set over time. Analyze time series data to understand the direction and magnitude of time-varying variables. Trend analysis helps identify long-term trends, estimate growth or decline, and make predictions based on past trends.

Problem (4): The yearly secondary data of milk production (in Lakh tones) from the financial year 2007-08 to 2015-16 have been collected from Bangladesh Economic Review 2017. Now our task is to forecast the production for 2017-18 to 2020-21 by using trend analysis.

Financial Year	**Milk Production (in Lakh tones)**
2008-09	22.86
2009-10	23.65
2010-11	29.47
2011-12	34.63
2012-13	50.67
2013-14	60.90
2014-15	69.69
2015-16	72.75
2016-17	92.83

Linear Trend Model

The linear trend model is given below–

$$Y_t = \beta_0 + \beta_1 T + \varepsilon_t$$

Where Y_t is the value of the dependent variable at time$t$.

Output 1: Linear Trend Analysis

Model Summary					
R	R Square	Adjusted R Square	Std. Error of the Estimate		
.981	.961	.956	5.220		
ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	4750.642	1	4750.642	174.347	.000
Residual	190.737	7	27.248		
Total	4941.379	8			
Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	8.898	.674	.981	13.204	.000
(Constant)	6.337	3.792		1.671	.139

Quadratic Model

The quadratic model is given below–

$$Y_t = \beta_0 + \beta_1 T + \beta_2 T^2 + \varepsilon_t$$

Where Y_t is the value of the dependent variable at time t.

Output 2: Results of quadratic model

Model Summary					
R	R Square	Adjusted R Square		Std. Error of the Estimate	
.989	.978	.971		4.212	
ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	4834.951	2	2417.476	136.288	.000
Residual	106.428	6	17.738		
Total	4941.379	8			

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	3.666	2.461	.404	1.490	.187
Case Sequence ** 2	.523	.240	.591	2.180	.072
(Constant)	15.929	5.359		2.972	.025

Exponential Growth Curve Model

The model for exponential growth are given below-

$$Y_t = \beta_0 e^{\beta_1 T + \varepsilon_t}$$

Where Y_t is the value of the dependent variable at time$t$.

Output 3: Exponential growth model

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.987	.973	.970	.091

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	2.114	1	2.114	255.329	.000
Residual	.058	7	.008		
Total	2.172	8			

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	.188	.012	.987	15.979	.000
(Constant)	17.716	1.171		15.128	.000
The dependent variable is ln(Milk Production).					

Comment: *The R^2 and adjusted R^2 for quadratic model are better than linear and exponential trend. Hence, quadratic model can be used to forecast about future.*

Output 4: Forecasted values by using quadratic variable is as follows-

Financial Year	Forecasted Value	Lower Confidence Interval	Upper Confidence Interval
2017-18	104.91048	88.23253	121.58843
2018-19	119.56376	97.69084	141.43669
2019-20	135.26343	106.58082	163.94604
2020-21	152.00949	115.06700	188.95198

Reference:

- Damodar N. Gujarati (2003), Basic Econometrics, 4th edition, McGraw Hill.
- Spyros Makridakis, Steven C. Wheelwright and Rob J. Hyndman (1998*), Forecasting Methods and Applications*, 3rd edition, John Wiley & Sons. Inc.

Trend Analysis for Agricultural Production and Price Accuracy: Model Selection Strategies

Dr. Sheikh Mohammad Sayem
Associate Professor
Department of Agricultural and Applied Statistics
Bangladesh Agricultural University, Mymensingh

Lecture: Methods of model selection and forecasting

🕮 Forecasting accuracy

Forecasting accuracy refers to the degree of closeness between predicted values and actual outcomes in a forecasting model. It measures how well a model can predict future values based on historical data or other relevant information.

🕮 Importance of forecasting accuracy in agricultural research

- **Crop Yield Prediction**: Accurate forecasts help farmers and agricultural researchers predict crop yields. This information is vital for planning planting schedules, managing resources like water and fertilizer, and making decisions about storage and distribution.
- **Resource Management**: Farmers need to manage resources efficiently, such as water, pesticides, and labor. Accurate forecasts can help optimize the allocation of these resources by providing insights into expected crop production levels and potential pest or disease outbreaks.
- **Market Planning**: Agricultural forecasts influence market planning and price setting. They help farmers and other stakeholders anticipate supply and demand dynamics, enabling them to make informed decisions about when to sell their products and at what price.
- **Risk Management**: Agriculture is inherently vulnerable to various risks, including weather fluctuations, pest outbreaks, and market volatility. Accurate forecasts allow farmers to identify and mitigate risks more effectively by implementing appropriate risk management strategies.
- **Research and Development**: Agricultural researchers rely on forecasting accuracy to assess the effectiveness of new technologies, farming practices, and crop varieties. Accurate forecasts provide feedback on the performance of these innovations under different environmental conditions, guiding further research and development efforts.
- **Policy Formulation**: Governments and policymakers use agricultural forecasts to develop policies and programs aimed at supporting farmers, ensuring food security, and promoting sustainable agricultural practices. Accurate forecasts provide the necessary data to inform policy decisions and assess their impact over time.

- **Climate Change Adaptation**: With climate change affecting weather patterns and agricultural productivity, accurate forecasts become even more critical. They help farmers and researchers anticipate and adapt to changing conditions, such as shifts in temperature, rainfall patterns, and the prevalence of extreme weather events.

Overall, forecasting accuracy in agricultural research plays a fundamental role in improving productivity, sustainability, and resilience within the agricultural sector, ultimately contributing to global food security and economic stability.

🕮 Methods of measuring forecasting accuracy

MAPE (Mean Absolute Percentage Error): The MAPE is a measure used to identify the accuracy of a forecast model as a percentage. The average percentage difference between predicted values and actual observed values is quantified through MAPE.

$$\text{MAPE} = \frac{1}{n}\sum \left|\frac{e_t}{Y_t}\right| \times 100$$

where $Y_t is$observed value for the time period t and e_t is the corresponding error of time$t$.

MAD (Mean Absolute Deviation): The MAD measures the average magnitude of the prediction errors, regardless of the direction of the errors (i.e. whether the model is over or under predicted).

$$\text{MAD} = \sum \frac{(Y_t - f_t)}{n}$$

Where $Y_t is$observed value for the time period t and f_t is the corresponding predicted or fitted value at time$t$ and n is the total number of observation.

MSD (Mean Squared Deviation): The MSD also measures the magnitude of prediction errors MAD measures, but by squaring the difference it gives more weight to large errors. In certain applications, such as time series forecasting using Gaussian error distributions MSD is often used.

$$\text{MSD} = \sum \frac{(Y_t - f_t)^2}{n-1}$$

Where $Y_t is$observed value for the time period t and f_t is the corresponding predicted or fitted value at time$t$ and n is the total number of observation.

RMSE (Root Mean Square Error):The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals.

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

Where y_i is the actual value for the i^{th} observation, $\hat{y}_i$ is the predicted value for the i^{th} observation, N is the number of observations and P is the number of parameter estimates, including the constant.

Bayesian Information Criterion (BIC): The Bayesian Information Criterion (AIC) is a way of selecting a model from a set of models for a given set of data.

The function of BIC is as follows-

$$AIC = -2\{Log(Likelihood)\} + d \times \log(N)$$

Where *N* is the sample size and *d* is the total number of parameters.

The R^2 Criterion: The coefficient of determination (R^2) is a statistical model selection criterion whose functional form is as follows:

$$R^2 = 1 - \frac{\sum_{n=1}^{t} \hat{U}_t^2}{\sum_{n=1}^{t}(Y_t - \bar{Y})^2}$$

Where n is the number of observations, Y_t is the observed value of the dependent variable and $u_t$is the difference between the observed and estimated values.

Decision: *The best model can be found out on the basis of maximum value of R^2, and minimum value of root mean squared error (RMSE), mean absolute percent error (MAPE) mean absolute deviation (MAD), mean squared deviation (MSD) and Bayesian information criterion (BIC).*

Ljung-Box Test: The Ljung-Box testis a test for autocorrelated errors. The functional form of the test statistic is as follows-

$$Q = n(n+2)\sum_{k=1}^{h}(n-k)^{-1}\rho_k^2 \quad which \ \ follows \ \chi^2 \ with \deg ree \ of \ \ freedom \ h-m.$$

Where n, k, h, m, and ρ_k refer the number of observations, the number of lag, the maximum lag, the number of parameters and the autocorrelation function respectively. Ljung-Box test can be used to test the hypothesis that all of the autocorrelations are zero; that is, that the series is white noise.

Lecture: Techniques of stationarity
Dr. Sheikh Mohammad Sayem, BAU

🕮 Stationary

If the underlying generating process for a time series is based on a constant mean and constant variance, then the time series is stationary otherwise, non-stationary.

A time series $\{Y_t,\ t=0,\pm1,\pm2,\cdots\}$ is said to be stationary if it has similar statistical properties to the "time shifted" series $\{Y_{t+h},\ t=0,\pm1,\pm2,\cdots\}$ for each integer h. Simply, a time series $\{Y_t,\ t=0,\pm1,\pm2,\cdots\}$ is said to be stationary time series if it is independent of time t.

🕮 Methods of checking stationarity

- **The Autocorrelation Function (ACF):** The functional form of autocorrelation is given below-

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

- **Partial Autocorrelation:** This measure of correlation is used to identify the extent of relationship between current values of a variable with earlier values of that same variable while holding the effect of all other time lags constant.

- **Correlogram:** Correlogramof a time series is a graph of autocorrelation at various lags. Stationarity can be tested by using correlogram.

- **Unit Root Test of Stationary**

An alternative test of stationary, which recently has become popular, is known as the unit root test. The easiest way is to introduce this test is to consider the following model

$$y_t = y_{t-1} + u_t \quad \cdots \quad \cdots \quad \cdots \quad (1)$$

where u_t is the stochastic error term which has zero mean and constant variance σ^2 .

Now if the coefficient of y_{t-1} is in fact equal to 1 , we face that the unit root problem i.e. a non-stationary situation. Therefore, if we run the following regression

$$y_t = \rho y_{t-1} + u_t \quad \cdots \quad \cdots \quad \cdots \quad (2)$$

where $-1 \le \rho \le 1$ and u_t is a White Noise error term.We know that if $\rho = 1$, that is in the case of the unit root y_t becomes a Random Walk model without drift , which is known as non-stationary stochastic process.

If the estimated ρ is statistically equal to 1, then y_t is non-stationary. Equation (2) is express as $\Delta y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$ where $\delta = (\rho - 1)$ and Δ represents the first difference operator. The hypotheses are

$$H_0 : \delta = 0$$
$$H_1 : \delta < 0$$

If we accept H_0 then $\rho = 1$, that is we have a unit root, i.e. the time series is non-stationary.

➢ Dickey Fuller Test

Dickey and fuller have shown that under the $H_0 : \rho = 0$ the estimated coefficient of y_{t-1} in the model $\Delta y_t = \delta y_{t-1} + u_t$ following τ statistics is know as Dickey Fuller test.

If the computed absolute value of the τ statistic exceeds the DF then we accept the hypothesis that the given time series is stationary. If it is less than the critical value, the time series is non-stationary.

➢ Augmented Dickey Fuller (ADF) Test

For theoretical and practical reasons, the dickey fuller test is applied to regression model of the following form

$$\Delta y_t = \delta y_{t-1} + u_t \quad \cdots \quad \cdots \quad \cdots \quad (1)$$
$$\Delta y_t = \beta_1 + \delta y_{t-1} + u_t \quad \cdots \quad \cdots \quad \cdots \quad (2)$$
$$\Delta y_t = \beta_1 + \beta_2 t + \delta y_{t-1} + u_t \quad \cdots \quad \cdots \quad \cdots \quad (3)$$

where t is the time or trend variable.If each case the null hypothesis that $H_0 : \delta = 0$ i.e. there is a unit root.

If the error term u_t is auto-correlated, one modifies (3) as follows

$$\Delta y_t = \beta_1 + \beta_2 t + \delta y_{t-1} + \alpha \sum_{i=1}^{m} \Delta y_{t-i} + \varepsilon_t \quad \cdots \quad \cdots \quad \cdots \quad (4)$$

The null hypothesis still that $\delta = 0$ or $\rho = 1$, i.e. a unit root exit in y .

When the Dickey Fuller test is applied to models like (4) , it is called Augmented Dickey Fuller (ADF) test. The Augmented Dickey Fuller test statistic has the same asymptotic distribution as the Dickey Fuller statistic.

Lecture: Box-Jenkins methodology for forecasting crop and livestock production

People have always wanted to predict the future to reduce their fear and anxiety about the unknown and an uncertain tomorrow. So, forecasting is an important aid in effective and efficient planning. A stochastic approach of time series model is a function that relates the value of time series to previous value of that time series, it's uncertain errors and other related time series. Autoregressive integrated moving average (ARIMA) is one of the popular stochastic approaches of forecasting time series data. The Box-Jenkins methodology can be applied to fit best autoregressive integrated moving average model for time series forecasting.

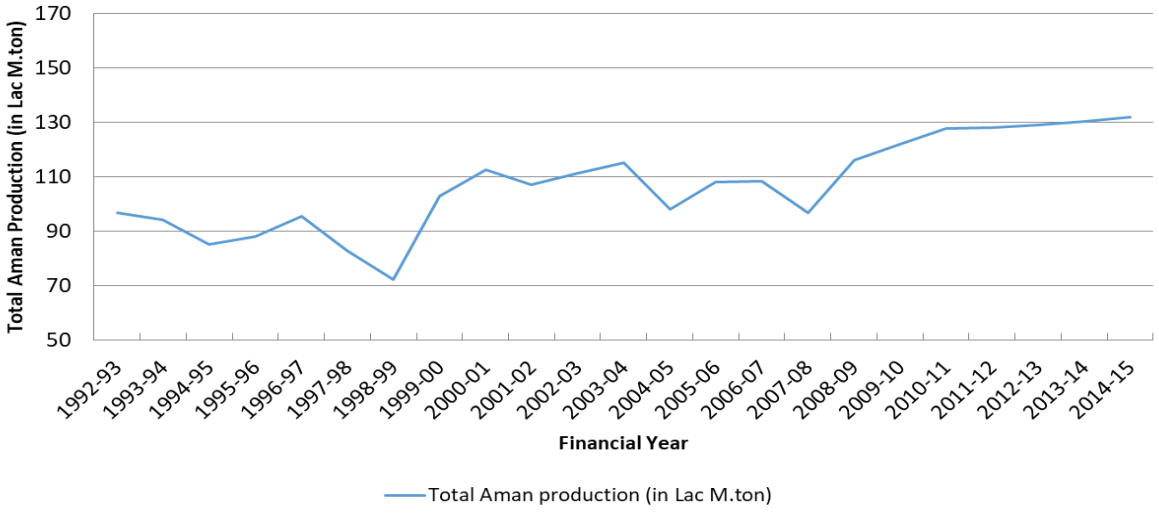
The method consists of three phases.

Phase 1: Identification

The aim of identification phase is to find out tentatively appropriate model.

- Plot the data to identify any unusual observations
- For achieving stabilize variance among the time series data, possible transformation is needed for data and again, plot the data

Example (1): The secondary data of yearly Aman production (in Lac M.ton) from the financial year 1992-93 to 2014-15 have been collected from BBS. Now our challenge is to forecast the production from 2015-16 to 2020-21. The time series plot represents the basic idea about the time series data.



Aman rice production of Bangladesh over the time is not stationary. So, the question is "What is the meaning of stationary?"

Phase 1: Identification (Cont…)

Stationary: If the underlying generating process for a time series is based on a constant mean and constant variance, then the time series is stationary otherwise, non-stationary.

- But, the accuracy of forecasting depends on the time series data which must be stationary. If the time series is not stationary, the stationary of the series need to be made by means of seasonal and non-seasonal differencing.

Differencing: When a time series is non-stationary, it can be often made stationary by taking first differences of a series-that is creating a new time series of successive differences (X_t - Xt_{-1}). If the first differences do not convert, the series to stationary for, then first differences of first differences can be created.

- When stationary has been achieved, examine the autocorrelation to see if any pattern remains.

Then the order of autoregressive and moving average are needed to be identified using autocorrelation function (ACF) and partial autocorrelation function (PACF) of the original or differenced series.

Table: Theoretical patterns of ACF and PACF

Type of model	Typical pattern of ACF	Typical pattern of PACF
AR(p)	Decays exponentially	Significant spikes through lags q
MA(q)	Significant spikes through lags q	Declines exponentially
ARMA(p,q)	Exponentially decay	Exponentially decay

The Autocorrelation Function (ACF): The functional form of autocorrelation is given below-

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

Partial Autocorrelation: This measure of correlation is used to identify the extent of relationship between current values of a variable with earlier values of that same variable while holding the effect of all other time lags constant.

Correlogram of a time series is a graph of autocorrelation at various lags.

Stationarity can be tested by using correlogram.

Algorithm of checking stationarity of time series data by SPSS:

Step 1: Input/Open Data file

$$Start \rightarrow IBM\ SPSS\ statistics\ 20 \rightarrow File \rightarrow Open \rightarrow Data$$

Step 2: Open Data dialog box. Input necessary information.

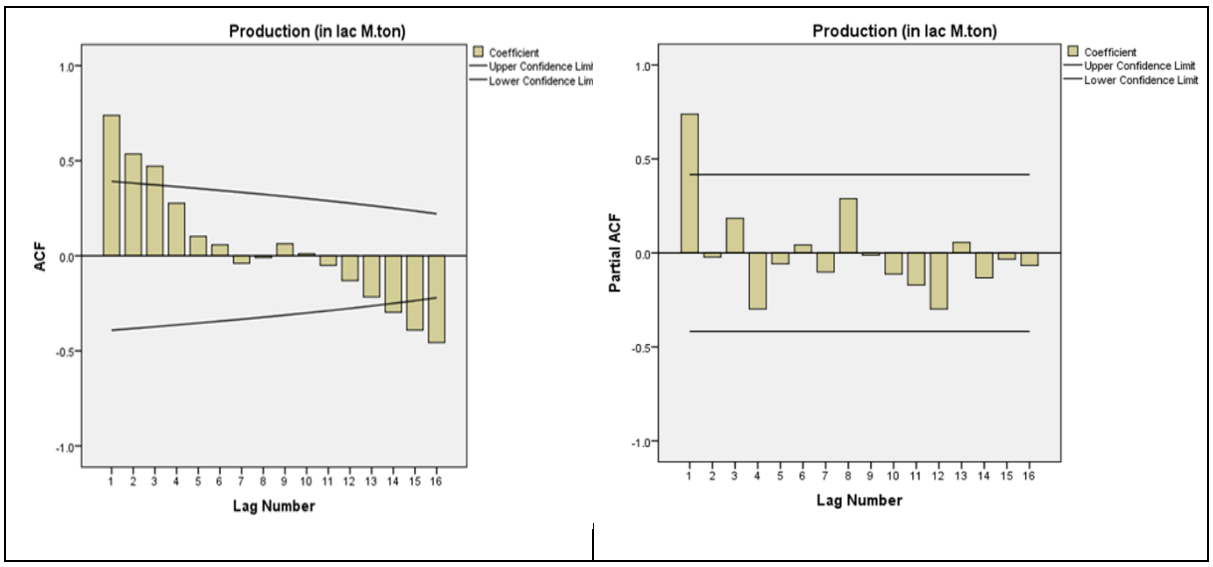
Click Ok.

Step 3: $Analyze \rightarrow Forecasting \rightarrow Autocorrelation \rightarrow Click\ Ok$

Step 4: Select Target Variable

Click "Display Autocorrelation and Partial autocorrelation"

Click Ok



Checking Stationarity by using Correlogram

The correlogram is a visual presentation of autocorrelations at various lag. Autocorrelation of the time series fall outside the boundaries of the correlogram suggest there may be some additional information in the series which is not being captured by the forecasted method. So, the time series is not stationary.

Aman rice production of Bangladesh over the time is not stationary because autocorrelation and partial autocorrelation of the time series fall outside the boundaries of the correlogram.

Algorithm of stochastic time series forecasting by SPSS:

Step 1: Input/Open Data file

$$Start \rightarrow IBM\ SPSS\ statistics\ 20 \rightarrow File \rightarrow Open \rightarrow Data$$

Step 2: Open Data dialog box.

Input necessary information.

Click Ok.

Step 3: Click Data

Define Date

Step 4: $Analyze \rightarrow Forecasting \rightarrow Creat\ Models \rightarrow Click\ Ok$

Step 5: Select Dependent Variables

Select methods

Define criteria (Order of autoregressive, difference, moving average) Select appropriate statistic, plots for checking forecasting accuracy

Select options (target time of forecasting)

Click Ok

At first difference, Aman rice production in different year's data are stationary which is important for next step of forecasting.

Phase 2: Estimation and Testing

Autoregressive integrated moving average model ARIMA (p, d, q) is the general model for forecasting purpose where p is the order of autoregression, d is the order of integration, q is the order of moving average.

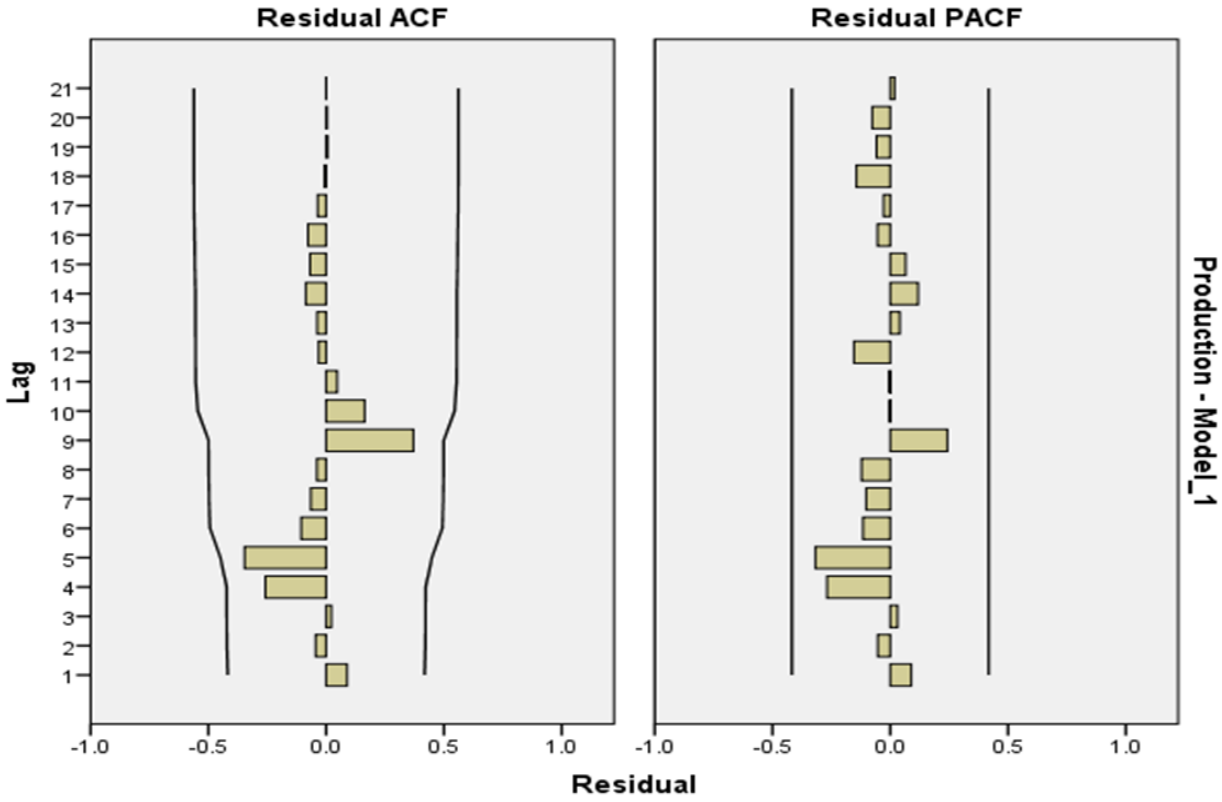
The method of least squares or the method of maximum likelihood estimation can be used to estimate the parameter of the models. The best model can be find out on the basis of maximum value of R2 and minimum value of root mean squared error (RMSE), mean absolute percent error (MAPE), Bayesian information criterion (BIC) where Ljung-Box test reveals that the residuals follows white noise.

Table 2: Model selection criteria for checking the best fitted model for AMAN rice

Model	**R^2**	**RMSE**	**MAPE**	**BIC**	**Ljung-Box Q**	
					Statistic	P-value
ARIMA(2,1,0)	**0.686**	**10.026**	**7.375**	**5.032**	**15.010**	**0.524**
ARIMA(1,1,0)	0.607	10.942	7.737	5.066	17.550	0.418
ARIMA(1,0,0)	0.602	10.838	8.635	5.039	17.409	0.427
ARIMA(0,0,1)	0.500	12.145	10.149	5.266	43.570	0.000

Among these models, ARIMA (2,1,0) has both lowest RMSE, normalised BIC and MAPE values and appeared to be the best model. Moreover, the Ljung-Box test suggested that the ACF of residuals for the model at different lag times was not significantly different (no significant autocorrelation) from zero (Q18=15.010) and p>.20).

Correlogram for the residual to check the fitted model



Lecture: ARIMA Model for forecasting crop/livestock production and price

Phase 3: Forecasting

The finally selected model can be used for the purpose of forecasting.

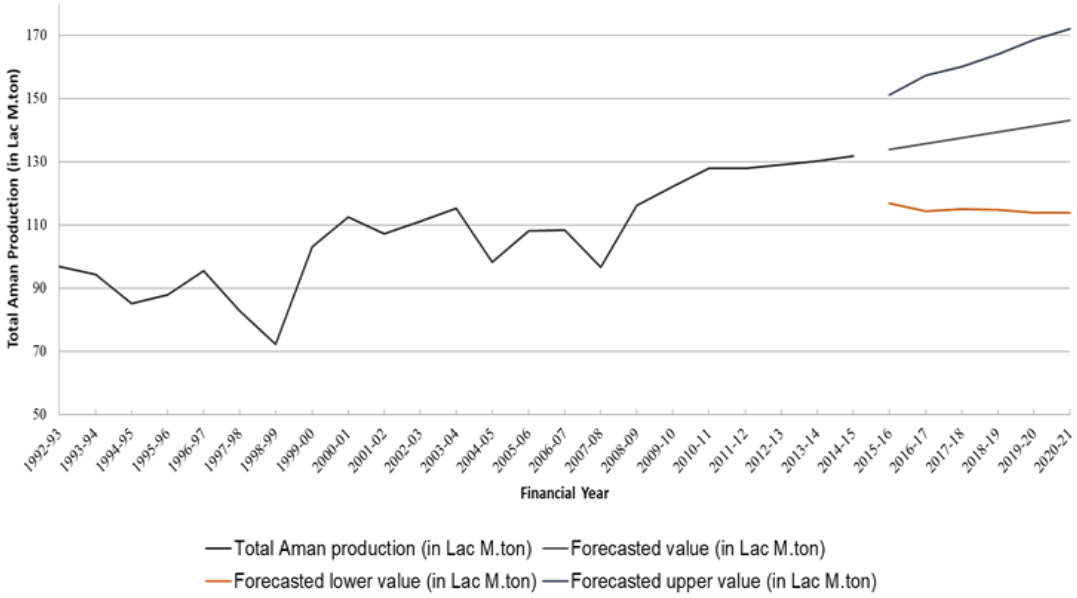
Example 1 (Cont…) Forecasting: The following table represents the Forecast production of AMAN rice (in Lac M.ton) from the financial year 2015-16 to 2020-21.

Financial year	ARIMA(2,1,0)		
	Forecast	Lower limit	Upper limit
2015-16	133.998	123.972	144.024
2016-17	135.810	125.784	145.836
2017-18	137.503	127.477	147.529
2018-19	139.350	129.324	149.376
2019-20	141.212	131.186	151.238
2020-21	143.003	132.977	153.029

- The equations to construct forecast limit are-

  Upper Limit= Forecast value + (Mean Square Error)$^{1/2}$

  Lower Limit= Forecast value - (Mean Square Error)$^{1/2}$



Forecast Aman Rice Production by ARIMA (2,1,0)

ARIMA is a famous stochastic approach for long term forecasting. But the policy maker need to forecast both short term and long term.

Hence, the combination of regression and ARIMA will be best!!!

Reference

- Damodar N. Gujarati (2003), Basic Econometrics, 4th edition, McGraw Hill.
- Spyros Makridakis, Steven C. Wheelwright and Rob J. Hyndman (1998), Forecasting Methods and Applications, 3rd edition, John Wiley & Sons. Inc.

Application of ARIMA and SARIMA Models in Agricultural Research: Techniques in Complex Surveys

Dr. Sheikh Mohammad Sayem
Associate Professor
Department of Agricultural and Applied Statistics
Bangladesh Agricultural University, Mymensingh

Lecture: SARIMA Model for forecasting production and price

🕮 Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is a sophisticated extension of the ARIMA model that is particularly useful for time series data exhibiting seasonal patterns. Here's a breakdown of the components:

1. **Seasonal Component (S)**: This refers to the repeating pattern that occurs over fixed intervals of time, such as monthly, quarterly, or annually. SARIMA accounts for this seasonality by introducing additional autoregressive and moving average terms that capture the seasonal patterns in the data.
2. **Autoregressive Component (AR)**: This component models the relationship between an observation and a number of lagged observations (auto-correlations) in the time series. The order of the autoregressive component (p) indicates how many lagged observations are included in the model.
3. **Integrated Component (I)**: This component deals with the differencing of the time series data to make it stationary. Differencing involves taking the difference between consecutive observations in the time series. The order of differencing (d) represents how many times differencing is required to achieve stationarity.
4. **Moving Average Component (MA)**: This component models the relationship between an observation and a residual error from a moving average model applied to lagged observations of the time series. The order of the moving average component (q) indicates how many lagged residuals are included in the model.

The notation for SARIMA is SARIMA (p, d, q) (P, D, Q) s, where:

- p, d, q are the non-seasonal ARIMA parameters.
- P, D, Q are the seasonal ARIMA parameters.
- s is the seasonal period (e.g., 12 for monthly data, 4 for quarterly data).

SARIMA models are estimated using historical time series data, and once fitted, they can be used to forecast future values of the time series.

Overall, SARIMA models are powerful tools for analyzing and forecasting time series data with seasonal patterns, making them widely used in various fields such as agriculture, economics, finance, and environmental science.

🕮 **Basic outline of how you can use SARIMA in Stata:**

1. **Data Preparation**: Ensure your time series data is loaded into Stata.
2. **Identify Seasonality and Trends**: Examine your time series data to identify any seasonality, trends, or patterns.
3. **Model Specification**: Determine the appropriate SARIMA model for your data based on its characteristics, including the order of autoregressive (AR), differencing (I), and moving average (MA) terms, as well as the seasonal order.
4. **Estimation**: Use Stata's built-in commands to estimate the parameters of the SARIMA model.
5. **Diagnostic Checking**: Evaluate the adequacy of the model using diagnostic checks such as residual analysis, AIC/BIC values, and other goodness-of-fit measures.
6. **Forecasting**: Once you have a satisfactory model, you can use it to forecast future values of the time series.

Algorithm of SARIMA

```
// Load your time series data into Stata
use "your_datafile.dta", clear

// Estimate a SARIMA model
sarima y, arima (2,1,2) sarima (1,0,1,12)

// Check diagnostic statistics
estat ic
estat respect
```

In this example, y represents your time series variable. The arima () option specifies the non-seasonal ARIMA structure (in this case, ARIMA (2,1,2)), while the sarima () option specifies the seasonal part of the model (in this case, SARIMA (1,0,1,12) with a seasonal period of 12). Remember to replace your_datafile.dta with the actual filename and path of your data file, and adjust the SARIMA model specification according to the characteristics of your data. Additionally, always perform diagnostic checks to ensure the validity of your model.

Lecture: Complex survey technique for agricultural production and market information
Dr. Sheikh Mohammad Sayem, BAU

🕮 Survey Design

Survey design is the process of preparing a complete plan of operations to be followed in conducting a survey and disseminating its intended results.

🕮 Census

If data are collected on all the elements of a population, the process is known as census.

Example: Population Census
Agricultural Census
Economic Census

Bangladesh Bureau of Statistics (BBS) is the government organizations who conduct that entire census.

🕮 Sample Survey

Sample survey is a method by which detailed information on the population characteristics are collected on the basis of sample element.

Example: Demographic Health Survey
Household Income Expenditure Survey
Labour Force Survey

🕮 Principal Steps in a Sample Survey

1) Defining the objectives
2) Defining the population to be sampled
3) Sampling frame and sampling units
4) Collecting of data
5) Data collection method

Example: Interview method, Mailed questionnaire methods

6) The schedule of questionnaire
   - ➢ Create the questionnaire
   - ➢ Pre-test the questionnaire
7) Sampling design
   - ➢ Sampling methods for selecting representative sample with appropriate sample size
   - ✓ Probability sampling
     - Simple random sampling
     - Stratified random sampling
     - Systematic sampling
     - Cluster sampling
     - Multistage sampling

- ✓ Purposive sampling
  - Area sampling
  - Quota sampling
  - Snowball sampling
- ✓ Mixed sampling
  - ➢ Select variables on which data are to be collected.

8) Administration of the survey
   - ➢ Conduct training program before starting the survey for increasing the skill of the field workers about sampling units, recording the information, data collection methods

9) Data processing and analysis
   - ➢ Editing of the data
   - ➢ Tabulation and presentation of data
   - ➢ Statistical analysis
   - ➢ Reporting and conclusion

🕮 Errors of Sample Survey

Two types of error may involve in the collection, organization and analysis of data.

- ➢ Sampling error
- ➢ Non sampling error

🕮 Sample size determination

The sample size is typically determined at the domain level, where a separate estimate is derived for each specific domain or group being studied. This approach ensures that the sample size is appropriate for the unique characteristics of each domain, allowing for accurate and reliable estimates. In general, the minimum required sample size is calculated using the standard formula for estimating the proportion, which is given by

- ➢ $n = \frac{N \times n_0}{N + n_0} \times Deff$
- ➢ $n_0 = \frac{z_{\frac{\alpha}{2}}^2 \times p \times (1-p)}{E^2} \times Deff$

Where; Deff represents the design effect, which accounts for the impact of stratification and clustering within the sampling process. p is the a priori proportion of the characteristic of interest in the population, providing an estimate of how prevalent the characteristic is. $z_{\frac{\alpha}{2}}^2$ is the value of standard normal variate corresponding to the confidence level α , N refers to the total population size and E is the margin of error rate, indicating the acceptable level of uncertainty around the estimate.

🕮 Design Effect

The design effect is the ratio of two theoretical variances for an estimator.

$$Design\ Effect = \frac{\text{Actual variance of the sample estimate obtained from a particular design}}{\text{Variance of a SRS estimate of the same sample size}}$$

$$\textit{Sample size} \text{ of complex design } (n) = \textit{Sample} \text{ size obtained by SRS} \times \textit{Design Effect}$$

🕮 Sampling weights and probability of selections

The sampling probability was calculated independently for each sampling stage and for each Primary Sampling Unit (PSU) within a sub-stratum. Within the sampling strategy described here, p_{1h_i} refers the probability of firs stage sampling of the ith PSU in the stratum h. n_h be the number of PSUs selected in the stratum h. M_{h_i} is the number of households of the ith PSU according to the stratum h and $\sum M_{h_i}$ is the total number of households in stratum h. The probability of selection of the ith PSU in stratum h was

$$p_{1h_i} = n_h \times \frac{M_{h_i}}{\sum M_{h_i}}$$

Again; S_{h_i}be the number of households selected within PSU i in the stratum h. In this sampling design $S_{h_i} = 20$, where H_{h_i} is the total number of household within PSU i in the stratum h .

$$p_{2h_i} = \frac{S_{h_i}}{H_{h_i}}$$

The ultimate sampling weight w_i is the inverse of the overall probability which is the product of the two selection probabilities mentioned above.

$$w_i = \frac{1}{p_{1h_i} \times p_{2h_i}}$$

Use of Multiple and Dynamic regression model for long term forecasting

Dr. Md. Akhtarul Alam
Professor
Department of Agricultural and Applied Statistics
Bangladesh Agricultural University, Mymensingh

Lecture: Autoregressive distributed lag model (ARDL)

Lagged Variables

- A possible source of any problem with the functional form is the lack of a lagged structure in the model.
- One way of overcoming autocorrelation is to add a lagged dependent variable to the model.
- However, although lagged variables can produce a better functional form, we need theoretical reasons for including them.

Inclusion of Lagged variables

- Inertia of the dependent variable, whereby a change in an explanatory variable does not immediately effect the dependent variable.
- The overreaction to 'news', particularly common in asset markets and often referred to as 'overshooting', where the asset 'overshoots' its long-run equilibrium position, before moving back towards equilibrium
- To allow the model to produce dynamic forecasts.

Types of Lag

- Autoregressive refers to lags in the dependent variable
- Distributed lag refers to lags of the explanatory variables
- Moving average refers to lags in the error term

- Recent development in econometrics have however, revealed that often times, most time series are not stationary as was conventionally thought. Therefore, different time series may not display the same features. Hence, it is possible to see some time series that display the feature of diverging away from their mean over time while others may converge to their mean over time. Time series that diverge away from their mean over time are said to be non-stationary.
- The classical estimation of variables with this relationship most times gives misleading inferences or spurious regression.
- To overcome this problem of non-stationarity and prior restrictions on the lag structure of a model, econometric analysis of time series data has increasingly moved towards cointegration. The reason is that, cointegration is a powerful way of detecting the

presence of steady state equilibrium between variables. Cointegration has become an over-riding requirement for any economic model using non-stationary time series data.

- If the variables do not cointegrate, then the problems of spurious regression and the results therein become almost meaningless. On the other hand, if the variables do cointegrate then we have cointegration.
- In applied econometrics, the Granger (1981) and, Engle and Granger (1987), Autoregressive Distributed Lag (ARDL) cointegration technique, Johansen and Juselius (1990) cointegration techniques have become the solution to determining the long run relationship between series that are non-stationary, as well as the Error Correction Model (ECM).
- The ECM result gives the short-run dynamics and long run relationship of the underlying variables.

ARDL Models

- An Autoregressive Distributed lag model or ARDL model refers to a model with lags of both the dependent and explanatory variables. An ARDL (1,1) model would have 1 lag on both variables

$$y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 x_{t-1} + \alpha_3 y_{t-1} + u_t$$

- ARDL cointegration technique does not require pretests for unit roots unlike other techniques. Consequently, ARDL cointegration technique is preferable when dealing with variables that are integrated of different order, I(0), I(1) or combination of the both and, robust when there is a single long run relationship between the underlying variables in a small sample size.
- Although ARDL cointegration technique does not require pre-testing for unit roots, to avoid ARDL model crash in the presence of integrated stochastic trend of I(2), we are of the view the unit root test should be carried out to know the number of unit roots in the series under consideration.
- Since each of the underlying variables stands as a single equation, endogeneity is less of a problem in the ARDL technique because it is free ofresidual correlation (i.e. all variables are assumed exogenous). Also, it enables us analyze the reference model.
- The major advantage of ARDL lies in its identification of the cointegrating vectors where there are multiple cointegrating vectors.

- The Error Correction Model (ECM) can be derived from ARDL model through a simple linear transformation, which integrates short run adjustments with long run equilibrium without losing long run information.
- The associated ECM model takes a sufficient number of lags to capture the data generating process in general to specific modeling frameworks.

$$\Delta y_t = \alpha_0 + \alpha_1 t + \alpha_2 (y_{t-1} - \theta x_t) + \sum_{i=1}^{p-1} \psi_{yi} \Delta y_{t-i} + \sum_{i=0}^{q-1} \psi'_{xi} \Delta x_{t-i} + u_t$$

The steps of the ARDL Cointegration Approach

- **Step 1: Determination of the Existence of the Long Run Relationship of the Variables**

  At the first stage the existence of the long-run relation between the variables under investigation is tested by computing the Bound F-statistic (bound test for cointegration) in order to establish a long run relationship among the variables. This bound F-statistic is carried out on each of the variables as they stand as endogenous variable while others are assumed as exogenous variables.

  In practice, testing the relationship between the forcing variable(s) in the ARDL model leads to hypothesis testing of the long-run relationship among the underlying variables.

  The null of non-existence of the long-run relationship is defined by;

  Ho: $\delta_1 = \delta_2 = 0$ (null, i.e. the long run relationship does not exist)

  H1: $\delta_1 \neq \delta_2 \neq 0$ (Alternative, i.e. the long run relationship exists)

- **Step 2: Choosing the Appropriate Lag Length for the ARDL Model**

  If a long run relationship exists between the underlying variables, then ARDL approach to cointegration can be applied. The issue of finding the appropriate lag length for each of the underlying variables in the ARDL model is very important because we want to have standard normal error terms that do not suffer from non-normality, autocorrelation, heteroskedasticity etc.

- **Step 3: ARDL into Error Correction Model**
- As we discuss before, when non-stationary variables are regressed in a model we may get results that are spurious.
- One way of resolving this is to difference the data in order to achieve stationarity of the variables. In this case, the estimates of the parameters from the regression model may be correct and the spurious equation problem resolved.
- However, the regression equation only gives us the short-run relationship between the variables. It does not give any information about the long run behaviour of the parameters in the model. This create a problem since researchers are mainly interested in long-run relationships between the variables under consideration.
- To solve this problem, the concept of cointegration and the ECM becomes imperative. With the specification of ECM, we now have both long-run and short-run information incorporated.

$$\Delta y_t = -\phi(1, \hat{p})EC_{t-1} + \sum_{i=1}^{k} \beta_{i0}\Delta x_{it} + \delta\Delta w_t - \sum_{j=1}^{p-1} \phi_j \Delta_{t-j} - \sum_{i=1}^{k} \sum_{j=1}^{\hat{q}-1} \beta_{ij}\Delta x_{i,1-j} + \mu_t$$

EC_t is the error correction term defined by;

$$EC_t = \varepsilon_t = y_t - \sum_{i=1}^{k} \hat{\theta}_i x_{it} - \psi' w_t$$

- The term EC_t as the speed of adjustment parameter or feedback effect is derived as the error term from the cointegration models. The ECt shows how much of the disequilibrium is being corrected, that is, the extent to which any disequilibrium in the previous period is being adjusted in *yt*.
- A positive coefficient indicates a divergence, while a negative coefficient indicates convergence. If the estimate of ECt = 1, then 100% of the adjustment takes place within the period, or the adjustment is instantaneous and full, if the estimate of ECt = 0.5, then 50% of the adjustment takes place each period/year. ECt = 0, shows that there is no adjustment, and to claim that there is a long-run relationship does not make sense any more.

- If the trace or Maximal eigenvalue or the F-statistics establishes that there exists a single long-run relation among the variables (i.e underlying variables), ARDL approach can be applied rather than applying Johansen and Juselius approach. The ARDL technique

provides a unified framework for testing and estimating of cointegration relations in the context of a single equation

- When there are multiple long-run relationships, ARDL approach cannot be applied. Hence, an alternative approach like Johansen and Juselius (1990) becomes more appropriate.
- If one wants to understand the dynamic relationship between two variables, there is a number of possible cases:
- Both are I(0), i.e. stationary. Then an OLS on the variable levels will be unbiased and efficient.
- The variables are integrated of the same order (eg. I(1)) but not cointegrated.
- Appropriate differentiation (i.e. first difference for first order integration) allows for OLS estimation.
- The variables are integrated of the same order and co-integrated. Then a level OLS provides the long-run relationship, whereas an Error Correction Model (ECM) (which can be estimated using OLS) represents the short-run dynamics.
- Data might be of different orders and/or co-integrated ("things are not as clear cut"). ARDL analyzes both short-run dynamics and long-run relationships.

Auto regressive distributed lag model (ARDL) and its advantages

- Autoregressive Distributed Lag Models (ARDL) model plays a vital role when comes to a need to analyze an economic scenario. In an economy, change in any economic variables may bring change in another economic variable beyond time. This change in a variable is not reflected immediately, but it distributes over future periods. Not only macroeconomic variables, but other variables such as loss or profit earned by a firm in a year can also affect the brand image of an organization over the period.
- On the other hand, the ARDL model addresses the issue of collinearity by allowing the lag of the dependent variable in the model with other independent variables and their lags.
- Assumptions for ARDL Model

  The absence of autocorrelation is the very first requirement of ARDL. The model requires that the error terms should have no autocorrelation with each other.

  There should not occur any heteroscedasticity in the data. In simple terms, the variance and mean should remain constant throughout the model.

The data should follow a normal distribution.

Data should have stationary either on I(0) or I(1) or on both. In addition to this, if any of the variables in the data has stationary at l(2), ARDL Model cannot run.

References

- Emeka Nkoro and Aham Kelvin Uko (2016) Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation, Journal of Statistical and Econometric Methods, vol.5, no.4, 2016, 63-91.
- Samuel Asumadu Sarkodie and Phebe Asantewaa Owusu (2020) How to apply the novel dynamic ARDL simulations (dynardl) and Kernel-based regularized least squares (krls) MethodsX 7, 101160.

Lecture: VAR Model using STATA
Dr. Md. Akhtarul Alam, BAU

Vector Autoregression (VAR)

1. Box–Jenkins and VAR approaches to economic forecasting are alternativesto traditional single- and simultaneous-equation models.

2. The VAR approach to forecasting considers several time series at atime. The distinguishing features of VAR are as follows:

- It is a truly simultaneous system in that all variables are regarded asendogenous.
- In VAR modeling the value of a variable is expressed as a linear functionof the past, or lagged, values of that variable and all other variables includedin the model.
- If each equation contains the same number of lagged variables in the system, it can be estimated by OLS without resorting to any systemsmethod, such as two-stage least squares (2SLS) or seemingly unrelated regressions(SURE).
- This simplicity of VAR modeling may be its drawback. In view of the limited number of observations that are generally available in most economicanalyses, introduction of several lags of each variable can consume alot of degrees of freedom.
- If there are several lags in each equation, it is not always easy to interpreteach coefficient, especially if the signs of the coefficients alternate. For this reason, one examines the impulse response function (IRF) in VARmodeling to find out how the dependent variable responds to a shock administeredto one or more equations in the system.

- There is considerable debate and controversy about the superiority ofthe various forecasting methods. Single-equation, simultaneous-equation, Box–Jenkins, and VAR methods of forecasting have their admirers as well asdetractors. All one can say is that there is no single method that will suit allsituations. If that were the case, there would be no need for discussing thevarious alternatives. One thing is sure: The Box–Jenkins and VAR methodologieshave now become an integral part of econometrics.

For example, three variables- GDP, investment and consumption from BBS. Whether these three variables are cointegrated we have to run Johansoncointegration test. It is to be noted that most of the time series data are non-stationary in the label form. But if we take the first difference or second difference they become stationary. First of all, let us assume that they are non-stationary in the level form. To test cointegration we have to follow this algorithm.

Open Stata data from Stata program or obtained from Excel.

Click statistics ⟶ Multivariate time series ⟶ Cointegrating rank of a VECM

```
. vecrank gdp consumption investment, trend(constant) max

                    Johansen tests for cointegration
Trend: constant                                      Number of obs =      10
Sample:  2008 - 2017                                       Lags =       2
```

maximum rank	parms	LL	eigenvalue	trace statistic	5% critical value
0	12	-219.97126	.	41.0744	29.68
1	17	-207.37922	0.91941	15.8903	15.41
2	20	-201.2501	0.70648	3.6321*	3.76
3	21	-199.43405	0.30456		

maximum rank	parms	LL	eigenvalue	max statistic	5% critical value
0	12	-219.97126	.	25.1841	20.97
1	17	-207.37922	0.91941	12.2583	14.07
2	20	-201.2501	0.70648	3.6321	3.76
3	21	-199.43405	0.30456		

Here, the null hypothesis is that there is no Co-integrated equation among the three variables. Here, rank= o, 1, 2, 3 means no cointegrated equation, 1 cointegrated equation, 3 cointegrated equation. It appears from both the trace statistics and max statistics; they are greater than 5 percent critical value. So, the null hypothesis is rejected of no cointegration. That means they are integrated. And they move together in the long run. If they is no cointegrated equation we can run unrestricted VAR model. Since they are cointegrated we shall have to run Vector Error Correction Model (VECM). So, just to know the unrestricted VAR technique we assume that the three variables- GDP, consumption and income are not cointegrated.

Click statistics ⟶ Multivariate time series ⟶ Vector auto regression (VAR)

```
Vector autoregression

Sample:  2008 - 2017                            Number of obs      =         10
Log likelihood =  -199.4341                     AIC                =   44.08681
FPE            =   7.68e+15                     HQIC               =   43.38975
Det(Sigma_ml)  =   4.22e+13                     SBIC               =   44.72224

Equation           Parms      RMSE     R-sq      chi2      P>chi2
------------------------------------------------------------------
gdp                    7   373.729   0.9977   4326.837   0.0000
consumption            7   2247.23   0.8493   56.37059   0.0000
investment             7   83.9678   0.9989   9082.565   0.0000
------------------------------------------------------------------
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gdp						
gdp						
L1.	-1.007433	.3220233	-3.13	0.002	-1.638587	-.376279
L2.	-1.153967	.3244265	-3.56	0.000	-1.789831	-.5181025
consumption						
L1.	.1293952	.0337201	3.84	0.000	.0633051	.1954853
L2.	.0982085	.0334457	2.94	0.003	.0326562	.1637609
investment						
L1.	8.535973	2.07238	4.12	0.000	4.474184	12.59776
L2.	1.24857	1.42186	0.88	0.380	-1.538225	4.035366
_cons	2255.329	596.0245	3.78	0.000	1087.143	3423.516
consumption						
gdp						
L1.	-.5150686	1.936327	-0.27	0.790	-4.310199	3.280062
L2.	-3.522278	1.950777	-1.81	0.071	-7.345731	.3011756
consumption						
L1.	.2828373	.2027588	1.39	0.163	-.1145627	.6802372
L2.	.5560854	.2011091	2.77	0.006	.1619188	.950252
investment						
L1.	-38.59592	12.46122	-3.10	0.002	-63.01947	-14.17237
L2.	55.40518	8.54965	6.48	0.000	38.64817	72.16219
_cons	11411.18	3583.897	3.18	0.001	4386.875	18435.49
investment						
gdp						
L1.	.0010332	.0723508	0.01	0.989	-.1407717	.1428381
L2.	-.1059217	.0728907	-1.45	0.146	-.2487848	.0369415
consumption						
L1.	.0150162	.0075761	1.98	0.047	.0001674	.029865
L2.	.0075127	.0075144	1.00	0.317	-.0072153	.0222407
investment						
L1.	1.500825	.4656131	3.22	0.001	.5882397	2.41341
L2.	-.0811791	.3194574	-0.25	0.799	-.707304	.5449458
_cons	-74.02248	133.9122	-0.55	0.580	-336.4855	188.4405

Here, GDP, consumption and investment are dependent variables and their lag variables are independent variables. It appears from the table that most of the coefficients are significant because their p values are less than 5 percent. As per example Z value of GDP L1= -3.13 and p value is 0.002 which is less than 5 percent. So, the coefficient of GDP L_1 is significant to explain GDP. Z value of investment L1is .01 and p value is 98.9 percent which is greater than

5 percent. That indicates that investment L1 has not any significant effect to explain investment.

In case of VAR model, all the coefficients are short run causalities. To run causality, we shall have to run Granger causality test.

Click Statistics ⟶ Multivariate Time Series ⟶ VAR diagnostics and tests ⟶ Granger Causality tests

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
gdp	consumption	23.705	2	0.000
gdp	investment	34.462	2	0.000
gdp	ALL	37.56	4	0.000
consumption	gdp	4.3206	2	0.115
consumption	investment	42.965	2	0.000
consumption	ALL	51.908	4	0.000
investment	gdp	3.3547	2	0.187
investment	consumption	4.9913	2	0.082
investment	ALL	8.3426	4	0.080

It appears from the table that most of the coefficients are significant since the p values are very small. As per example there is short run causality running from consumption to GDP. But In case of GDP there is no short term causality running from GDP to investment because the p value is 18.7 percent. If we consider both GDP and Consumption, we see that p value is 0.000 that indicates there is short term causality running from GDP and consumption to investment.

Application of panel regression for forecasting agricultural product and price

Dr. Md. Fuad Hassan
Associate Professor
Department of Agricultural and Applied Statistics
Bangladesh Agricultural University, Mymensingh

Lecture: Basics of panel regression model

🕮 Panel Data

A panel data set (also longitudinal data) has both a cross-sectional and a time series dimension, where all cross section units are observed during the whole time period.

Balanced Panel: If each cross-sectional unit has the same number of time series observations, then is called balanced panel.

Unbalanced Panel: If each cross-sectional unit has not the same number of time series observations, then is called unbalanced panel.

Balanced Panel:					**Unbalanced Panel:**				
Person	Year	Income	Age	Sex	Person	Year	Income	Age	Sex
1	2001	1300	27	M	1	2001	1300	27	M
1	2002	1600	28	M	1	2002	1600	28	M
1	2003	2000	29	M	2	2001	2000	37	F
2	2001	2000	38	F	2	2002	2000	38	F
2	2002	2300	39	F	2	2003	2300	39	F
2	2003	2400	40	F	3	2001	2400	34	M

🕮 Panel Data Regression

Panel regression models are based on panel data. Panel data faces several estimation and inference problems. Since panel data involve both cross-section and time dimensions, problems that plague cross-sectional data (say, heteroscedasticity) and time series data (say, autocorrelation) need to be addressed.

Panel data analysis has three more-or-less independent approaches:

i) Pooled OLS Regression model;
ii) Fixed effects model;
iii) Random effects model.

Example (1): Data for General Electric (1), General Motor (2), U.S. Steel (3), and Westinghouse (4) on three variables real gross investment (Y), real value of the firm (X_2) and real capital stock (X_3) are available for the period 1935-1954 (appendix 1). Thus, there are four cross-sectional units and 20 time periods. In all, therefore, we have 80 observations. A priori, Y is expected to be positively related to X_2 and X_3. The objectives of this study are to develop panel regression model, estimate the parameters and test it.

Lecture: Hands-on exercise on different Panel regression models

🕮 Pooled OLS Regression model

If intercept and slope coefficients are constant across time and space and the error term captures differences over time and individuals, the pooled OLS regression can be used.

Algorithm 1: (Pooled regression model by Stata)

Step 1: *Open Stata* $\rightarrow$ *File* $\rightarrow$ Im*port Data* $\rightarrow$ *Excel Spreadsheet* $\rightarrow$ *click or click Enter*

Step 2: In Import Excel dialog box

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step 3: $Statistics \rightarrow Linear\ models\ and\ related \rightarrow Linear\ Regression$

In "regress-linear regression" dialog box-

- Select Dependent Variable
- Select Independent Variable

Click OK

The Stata results are given below-

```
. import excel "C:\Users\Hp\Desktop\ICF STATA\Data\2nd day\panel data example.xlsx", sheet("panel") firstrow

. regress Y x1 x2
```

Source	SS	df	MS	Number of obs	=	80
				F(2, 77)	=	119.63
Model	4849457.34	2	2424728.67	Prob > F	=	0.0000
Residual	1560689.71	77	20268.6975	R-squared	=	0.7565
				Adj R-squared	=	0.7502
Total	6410147.05	79	81141.1019	Root MSE	=	142.37

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.1100955	.0137297	8.02	0.000	.0827563	.1374348
x2	.3033932	.0492957	6.15	0.000	.2052328	.4015535
_cons	-63.30414	29.6142	-2.14	0.036	-122.2735	-4.334735

🕮 Fixed effect regression model

If slope coefficients constant across individuals or over time but the intercept varies across individuals (each individual's intercept does not vary over time), fixed effect model can be used.

Algorithm 2: (Fixed effect regression model by Stata)

Step 1: *Open Stata → File → Import Data → Excel Spreadsheet → click or click Enter*

Step 2: In Import Excel dialog box

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

  Click OK

Step 3:

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Setup\ and\ Utilities \rightarrow Declare\ dataset\ to\ be\ panel\ data \rightarrow click$$

In "xtset-declare dataset to be panel data" dialog box-

- Select Panel ID variable
- Select Time variable

  Click OK

Step 4:

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Linear\ models \rightarrow Linear\ regression\ (FE, RE, PA, BE) \rightarrow Click$$

In xtreg…dialog box-

- Select dependent variable
- Select independent variable
- Tick Fixed effect

Click OK

The Stata results are given below-

```
. import excel "C:\Users\Hp\Desktop\ICF STATA\Data\2nd day\panel data example.xlsx", sheet("panel") firstrow

. xtset company year
       panel variable:  company (strongly balanced)
        time variable:  year, 1935 to 1954
                delta:  1 unit

. xtreg Y x1 x2, fe

Fixed-effects (within) regression               Number of obs     =         80
Group variable: company                         Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0.8068                                         min =         20
     between = 0.7304                                         avg =       20.0
     overall = 0.7554                                         max =         20

                                                F(2,74)           =     154.53
corr(u_i, Xb)  = -0.1001                        Prob > F          =     0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1079481   .0175089     6.17   0.000     .0730608    .1428354
          x2 |   .3461617   .0266645    12.98   0.000     .2930315    .3992918
       _cons |  -73.84947   37.52291    -1.97   0.053    -148.6155    .9165626
-------------+----------------------------------------------------------------
     sigma_u |  139.05116
     sigma_e |  75.288896
         rho |  .77329632   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3, 74) = 67.11                     Prob > F = 0.0000
```

Interpretation:

Since the p-value of the F test is less than 0.05, at least one of the independent variables effect is significant. So, our model is ok.

Again, the p-value of individual t test for both independent variables is less than 0.05. So, dependent variable is significantly influenced by both independent variables.

Rho is known as the intra class correlation which is calculate using the formula-$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2+(sigma\_e)^2}$

Here, $sigma\_u = Standard\ deviation\ of\ the\ residuals\ within\ group\ u_i$

$sigma\_e = Standard\ deviation\ of\ the\ residuals\ (overall\ error\ terms)\ e_i$

Rho=0.7733 refers that 77.33% of the variance is due to difference across panels.

🕮 Random Effect Regression model (REM)

In REM, we assume that the intercept value of an individual unit is a random drawing from a much larger population with a constant mean.

Algorithm 2: (Random effect regression model by Stata)

Step 1: *Open Stata* → *File* → Im*port Data* → *Excel Spreadsheet* → *click or click Enter*

Step 2: In Import Excel dialog box

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step 3:

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Setup\ and\ Utilities \rightarrow Declare\ dataset\ to\ be\ panel\ data \rightarrow click$$

In "xtset-declare dataset to be panel data" dialog box-

- Select Panel ID variable
- Select Time variable

Click OK

Step 4:

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Linear\ models \rightarrow Linear\ regression\ (FE, RE, PA, BE) \rightarrow Click$$

In xtreg…dialog box-

- Select dependent variable
- Select independent variable
- Tick GLS random effect

Click OK

The Stata results are given below-

```
. import excel "C:\Users\Hp\Desktop\ICF STATA\Data\2nd day\panel data example.xlsx", sheet("panel") firstrow

. xtset company year
      panel variable:  company (strongly balanced)
       time variable:  year, 1935 to 1954
               delta:  1 unit

. xtreg Y x1 x2, re

Random-effects GLS regression                   Number of obs     =        80
Group variable: company                         Number of groups  =         4

R-sq:                                           Obs per group:
     within  = 0.8068                                         min =        20
     between = 0.7303                                         avg =      20.0
     overall = 0.7554                                         max =        20

                                                Wald chi2(2)      =    317.79
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =    0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1076555   .0168169     6.40   0.000     .0746949     .140616
          x2 |   .3457104   .0265451    13.02   0.000     .2936829    .3977378
       _cons |  -73.03531   83.94957    -0.87   0.384    -237.5734    91.50283
-------------+----------------------------------------------------------------
     sigma_u |  152.15823
     sigma_e |  75.288896
         rho |  .80332023   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

🕮 Hausman Specification Test

The Hausman test can be used to decide between fixed or random effects where the null hypothesis is that the random effects model is preferred vs. the alternative the fixed effects. Here, run a fixed effects model and save the estimates, then run a random model and save the estimates, then perform the test.

Algorithm 3: (Hausman test by Stata)

Step 1 (Import data):

$$Open\ Stata \rightarrow File \rightarrow \text{Im}\ port\ \ Data \rightarrow Excel\ Spreadshee\,t \rightarrow click\ or\ click\ Enter$$

In Import Excel dialog box-

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step 2 (Declare dataset to be panel data):

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Setup\ and\ Utilities \rightarrow Declare\ dataset\ to\ be\ panel\ data \rightarrow click$$

In "xtset-declare dataset to be panel data" dialog box-

- Select Panel ID variable
- Select Time variable

Click OK

Step 3 (run fixed effect model and store the results):

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Linear\ models \rightarrow Linear\ regression\ (FE, RE, PA, BE) \rightarrow Click$$

In "xtreg…" dialog box-

- Select dependent variable
- Select independent variable
- Tick Fixed effect

Click OK

For record results:

$$Statistics \rightarrow Postestimation \rightarrow Mange\ estimation\ results \rightarrow Store\ current\ estimation \rightarrow Launch$$

Write the filename (say, fixed) in "estimate store…" dialog box.

Click OK

Step 4 (run random effect model and store results):

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Linear\ models \rightarrow Linear\ regression\ (FE, RE, PA, BE) \rightarrow Click$$

In "xtreg…" dialog box-

- Select dependent variable
- Select independent variable
- Tick GLS random effect

Click OK

For record results:

$$Statistics \rightarrow Postestimation \rightarrow Mange\ estimation\ results \rightarrow Store\ current\ estimation \rightarrow Launch$$

Write the filename (say, random) in "estimate store…" dialog box.

Click OK

Step 5 (Hausman specification test):

$$Statistics \rightarrow Postestimation \rightarrow Specification, diagnosis\ ... \rightarrow Hausman\ specification\ test \rightarrow Launch$$

- Select consistent estimator (say, fixed)

Click OK

The Stata results are given below-

```
. import excel "C:\Users\Hp\Desktop\ICF STATA\Data\2nd day\panel data example.xlsx", sheet("panel") firstrow

. xtset company year
       panel variable:  company (strongly balanced)
        time variable:  year, 1935 to 1954
                delta:  1 unit

. xtreg Y x1 x2, fe

Fixed-effects (within) regression               Number of obs     =         80
Group variable: company                         Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0.8068                                         min =         20
     between = 0.7304                                         avg =       20.0
     overall = 0.7554                                         max =         20

                                                F(2,74)           =     154.53
corr(u_i, Xb)  = -0.1001                        Prob > F          =     0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1079481   .0175089     6.17   0.000     .0730608    .1428354
          x2 |   .3461617   .0266645    12.98   0.000     .2930315    .3992918
       _cons |  -73.84947   37.52291    -1.97   0.053    -148.6155    .9165626
-------------+----------------------------------------------------------------
     sigma_u |  139.05116
     sigma_e |  75.288896
         rho |  .77329632   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3, 74) = 67.11                     Prob > F = 0.0000

. estimates store fixed

. xtreg Y x1 x2, re

Random-effects GLS regression                   Number of obs     =         80
Group variable: company                         Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0.8068                                         min =         20
     between = 0.7303                                         avg =       20.0
     overall = 0.7554                                         max =         20

                                                Wald chi2(2)      =     317.79
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1076555   .0168169     6.40   0.000     .0746949     .140616
          x2 |   .3457104   .0265451    13.02   0.000     .2936829    .3977378
       _cons |  -73.03531   83.94957    -0.87   0.384    -237.5734    91.50283
-------------+----------------------------------------------------------------
     sigma_u |  152.15823
     sigma_e |  75.288896
         rho |  .80332023   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random

. hausman fixed .

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |    fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
          x1 |   .1079481     .1076555        .0002926        .0048738
          x2 |   .3461617     .3457104        .0004513        .0025204
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =         0.07
                Prob>chi2 =       0.9678
```

Comment: Since the P-value of Hausman test is greater than 0.05, the null hypothesis cannot be rejected. Hence, random effect model is more appropriate for the panel data.

🕮 Breusch and Pagan LM test

The LM test helps us to decide a model among a random effects regression and a simple OLS regression. The null hypothesis in the LM test is that variances across entities are zero. This is no significant difference across units (i.e. no panel effect).

Algorithm 4: (Breusch and Pagan LM test by Stata)

Step 1 (Import data):

Open Stata → File → Im*port Data → Excel Spreadsheet → click or click Enter*

In Import Excel dialog box-

- Browse Excel File
- Select Worksheet
- Tick Import first row as variable name

Click OK

Step 2 (Declare dataset to be panel data):

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Setup\ and\ Utilities \rightarrow Declare\ dataset\ to\ be\ panel\ data \rightarrow click$$

In "xtset-declare dataset to be panel data" dialog box-

- Select Panel ID variable
- Select Time variable

Click OK

Step 3 (run OLS regression and store the results):

$$Statistics \rightarrow Linear\ models\ and\ related \rightarrow Linear\ regression \rightarrow Click$$

In "regress-linear regression" dialog box-

- Select dependent variable
- Select independent variable

Click OK

For record results:

$$Statistics \rightarrow Postestimation \rightarrow Mange\ estimation\ results \rightarrow Store\ current\ estimation \rightarrow Launch$$

Write the filename (say, OLS) in "estimate store…" dialog box.

Click OK

Step 4 (run random effect model and store results):

$$Statistics \rightarrow Longitudinal\ or\ Panel\ data \rightarrow Linear\ models$$
$$\rightarrow Linear\ regression\ (FE, RE, PA, BE) \rightarrow Click$$

In xtreg…dialog box-

- Select dependent variable
- Select independent variable
- Tick GLS random effect

Click OK

For record results:

$Statistics \rightarrow Postestimation \rightarrow Mange\ estimation\ results \rightarrow$
$Store\ current\ estimation \rightarrow Launch$

Write the filename (say, random) in "estimate store…" dialog box.

Click OK

Step 5 (Breusch and Pagan LM test):

$$Statistics \rightarrow Postestimation \rightarrow Specification, diagnosis \ldots$$
$$\rightarrow Breusch\ and\ Pagan\ LM\ test \ldots \rightarrow Launch$$

Click OK

The Stata results are given below-

```
. import excel "C:\Users\Hp\Desktop\BBS Training\Second day_BBS\Data Second Day\panel data example.xlsx", sheet("panel") firstrow

. xtset company year, yearly
       panel variable:  company (strongly balanced)
        time variable:  year, 1935 to 1954
                delta:  1 year

.
. regress Y x1 x2
```

Source	SS	df	MS	Number of obs	=	80
				F(2, 77)	=	119.63
Model	4849457.34	2	2424728.67	Prob > F	=	0.0000
Residual	1560689.71	77	20268.6975	R-squared	=	0.7565
				Adj R-squared	=	0.7502
Total	6410147.05	79	81141.1019	Root MSE	=	142.37

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.1100955	.0137297	8.02	0.000	.0827563	.1374348
x2	.3033932	.0492957	6.15	0.000	.2052328	.4015535
_cons	-63.30414	29.6142	-2.14	0.036	-122.2735	-4.334735

```
. estimates store OLS

. xtreg Y x1 x2, re

Random-effects GLS regression                   Number of obs     =         80
Group variable: company                         Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0.8068                                         min =         20
     between = 0.7303                                         avg =       20.0
     overall = 0.7554                                         max =         20

                                                Wald chi2(2)      =     317.79
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000
```

Y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	.1076555	.0168169	6.40	0.000	.0746949	.140616
x2	.3457104	.0265451	13.02	0.000	.2936829	.3977378
_cons	-73.03531	83.94957	-0.87	0.384	-237.5734	91.50283
sigma_u	152.15823					
sigma_e	75.288896					
rho	.80332023	(fraction of variance due to u_i)				

```
. estimates store random

. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        Y[company,t] = Xb + u[company] + e[company,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+------------------------------
                       Y |     81141.1       284.8528
                       e |    5668.418        75.2889
                       u |    23152.13       152.1582

        Test:   Var(u) = 0
                             chibar2(01) =   379.08
                          Prob > chibar2 =   0.0000
```

Comment: Since p-value of the LM test is less than 0.05, the null hypothesis (H_0: *The random effect panel regression is not appropriate*) can be rejected. So, the data consist random and panel effect. Hence, Random effect model is appropriate.

Appendix 1

Y	x1	x2	company	year	Y	x1	x2	company	year
33.1	1170.6	97.8	1	1935	230.4	1957.3	312.7	3	1939
45	2015.8	104.4	1	1936	361.6	2202.9	254.2	3	1940
77.2	2803.3	118	1	1937	472.8	2380.5	261.4	3	1941
44.6	2039.7	156.2	1	1938	445.6	2168.6	298.7	3	1942
48.1	2256.2	172.6	1	1939	361.6	1985.1	301.8	3	1943
74.4	2132.2	186.6	1	1940	288.2	1813.9	279.1	3	1944
113	1834.1	220.9	1	1941	258.7	1850.2	213.8	3	1945
91.9	1588	287.8	1	1942	420.3	2067.7	232.6	3	1946
61.3	1749.4	319.9	1	1943	420.5	1796.7	264.8	3	1947
56.8	1687.2	321.3	1	1944	494.5	1625.8	306.9	3	1948
93.6	2007.7	319.6	1	1945	405.1	1667	351.1	3	1949
159.9	2208.3	346	1	1946	418.8	1677.4	357.8	3	1950
147.2	1656.7	456.4	1	1947	588.2	2289.5	341.1	3	1951
146.3	1604.4	543.4	1	1948	645.2	2159.4	444.2	3	1952
98.3	1431.8	618.3	1	1949	641	2031.3	623.6	3	1953
93.5	1610.5	647.4	1	1950	459.3	2115.5	669.7	3	1954
135.2	1819.4	671.3	1	1951	12.93	191.5	1.8	4	1935
157.3	2079.7	726.1	1	1952	25.9	516	0.8	4	1936
179.5	2371.6	800.3	1	1953	35.05	729	7.4	4	1937
189.6	2759.9	888.9	1	1954	22.89	560.4	18.1	4	1938
317.6	3078.5	2.8	2	1935	18.84	519.9	23.5	4	1939
391.8	4661.7	52.6	2	1936	28.57	628.5	26.5	4	1940
410.6	5387.1	156.9	2	1937	48.51	537.1	36.2	4	1941
257.7	2792.2	209.2	2	1938	43.34	561.2	60.8	4	1942
330.8	4313.2	203.4	2	1939	37.02	617.2	84.4	4	1943
461.2	4643.9	207.2	2	1940	37.81	626.7	91.2	4	1944
512	4551.2	255.2	2	1941	39.27	737.2	92.4	4	1945
448	3244.1	303.7	2	1942	53.46	760.5	86	4	1946
499.6	4053.7	264.1	2	1943	55.56	581.4	111.1	4	1947
547.5	4379.3	201.6	2	1944	49.56	662.3	130.6	4	1948
561.2	4840.9	265	2	1945	32.04	583.8	141.8	4	1949
688.1	4900	402.2	2	1946	32.24	635.2	136.7	4	1950
568.9	3526.5	761.5	2	1947	54.38	732.8	129.7	4	1951
529.2	3245.7	922.4	2	1948	71.78	864.1	145.5	4	1952
555.1	3700.2	1020.1	2	1949	90.08	1193.5	174.8	4	1953
642.9	3755.6	1099	2	1950	68.6	1188.9	213.5	4	1954
755.9	4833	1207.7	2	1951					
891.2	4924.9	1430.5	2	1952					
1304.4	6241.7	1777.3	2	1953					
1486.7	5593.6	2226.3	2	1954					
209.9	1362.4	53.8	3	1935					
355.3	1807.1	50.5	3	1936					
469.9	2673.3	118.1	3	1937					
262.3	1801.9	260.2	3	1938					

Reference

- Damodar N. Gujarati (2003), *Basic Econometrics*, 4th edition, McGraw Hill.
- A. Colin Cameron and Pravin K. Trivedi (2009), *Microeconometrics Using Stata*, a Stata Press Publication, Texas.
- Spyros Makridakis, Steven C. Wheelwright and Rob J. Hyndman (1998*), Forecasting Methods and Applications*, 3rd edition, John Wiley & Sons. Inc.

Future research scope in socio-economic perspectives: Advancements and challenges in forecasting

Dr. Md. Shofiqul Islam
Principal Scientific Officer (PSO)
Agricultural Economics and Rural Sociology (AERS) Division
Bangladesh Agricultural Research Council, Farmgate, Dhaka
Email address: shafiqbau07@gmail.com
Mobile no.: 01704-778929

Agriculture remains the backbone of many economies, especially in developing nations where it provides employment and sustains livelihoods for millions. However, for agriculture to continue contributing effectively to poverty reduction and food security in the face of changing global dynamics, it must evolve. As technology advances, markets become more interconnected and policy frameworks shift, it is crucial to assess how these changes impact rural communities, food production, and economic development. This manual delves deeply into key areas of research in agriculture, focusing on market intelligence, policy research, innovative commercial agriculture, fallow land utilization and investment analysis of agricultural projects. By exploring these areas, this topic aims to guide stakeholders in designing inclusive, sustainable and economically viable agricultural strategies.

Socio-Economic Perceptions and Impact of Agricultural Technologies

The adoption of new agricultural technologies is a key driver of increased productivity and efficiency in the agricultural sector. However, the uptake of these technologies is heavily influenced by socio-economic perceptions and local contexts. This research focuses on understanding these perceptions and the impact of these technologies on rural communities.

Key Areas of Research:

- **Adoption Barriers**: Understanding the barriers that prevent farmers, particularly smallholders, from adopting new technologies. These include access to credit, training, perceived risk, and cultural resistance to change.
- **Technological Acceptance and Perceptions**: Investigating how farmers perceive the potential of agricultural technologies, including benefits such as increased yield, reduced labor, and profitability, and concerns like high costs, lack of support, and potential environmental impact.
- **Gender and Social Equity**: Exploring how different groups within rural communities especially women, youth, and marginalized groups experience the impacts of

technology adoption. Ensuring that technologies are inclusive and accessible for all socio-economic groups is critical.

Advancements:

- **Gender-Sensitive Technologies**: Ensuring that innovations are designed with gender considerations in mind, such as technologies that empower women through access to labor-saving devices or improved farming methods.
- **Participatory Research**: Involving farmers in the design and testing of technologies to ensure they meet the needs and preferences of end-users, ensuring better adoption rates.

Challenges:

- **Affordability**: Many new agricultural technologies are too expensive for small-scale farmers to afford, especially in developing countries.
- **Cultural Resistance**: Traditional farming practices may create resistance to adopting new technologies, especially if they are perceived as conflicting with local knowledge or cultural practices.

Efficient Production and Marketing Systems Development

Developing efficient agricultural production and marketing systems is essential to the long-term sustainability of agriculture. An efficient system reduces waste, ensures fair pricing, and improves farmers' access to profitable markets.

Key Areas of Research:

- **Market Linkages and Access**: Investigating mechanisms to link farmers directly with markets, cutting down intermediaries who often absorb a significant portion of the farmer’s potential income. This includes exploring cooperatives, contract farming, and digital platforms.
- **Supply Chain Management**: Researching how to optimize the agricultural supply chain from production to market. This includes reducing transportation costs, improving storage facilities, and minimizing post-harvest losses.
- **Price Discovery and Market Transparency**: Ensuring farmers have access to accurate and timely market price information through market intelligence platforms which help them make informed decisions about when and where to sell their products.

Advancements:

- **Cold Chain Infrastructure**: The development of cold chain solutions especially in developing regions can help preserve perishable products reducing waste and extending market reach for farmers.
- **Digital Platforms and E-Commerce**: Digital solutions like online marketplaces, mobile apps, and e-commerce platforms that enable farmers to reach broader markets directly, eliminating middlemen, and enhancing profitability.

Challenges:

- **Infrastructure Gaps**: Many rural regions still lack the basic infrastructure necessary to support efficient marketing systems, including roads, storage facilities and reliable electricity.
- **Market Volatility**: Agricultural prices are often highly volatile due to seasonal changes, global demand shifts and climate change impacts which can affect farmers' income stability.

Market Intelligence

Market intelligence is the process of collecting, analyzing, and interpreting data about the market environment, consumer demand, competitive landscapes and pricing trends to make better business decisions. This research area helps farmers and agribusinesses align their strategies with current and future market trends.

Key Areas of Research:

- **Consumer Behavior Analysis**: Understanding consumer demand and preferences, including emerging trends such as organic, fair trade and sustainable products. This allows farmers to identify high-demand markets and align their production strategies accordingly.
- **Pricing Mechanisms and Forecasting**: Researching ways to improve price transparency through market intelligence tools and forecasting systems that help farmers plan better and reduce risk.
- **Global Market Trends**: Exploring international trade policies, shifts in global demand, and market entry strategies for farmers and agribusinesses, particularly in export-oriented agriculture.

Advancements:

- **Big Data and AI Integration**: The use of big data analytics and artificial intelligence (AI) to analyze trends and predict market movements, consumer preferences and pricing patterns, enabling farmers and agribusinesses to anticipate changes and make informed decisions.
- **Mobile Apps and SMS-Based Market Information**: Mobile applications that provide real-time updates on market prices, weather forecasts and other relevant data, empowering farmers with information at their fingertips.

Challenges:

- **Data Gaps**: Accessing accurate and timely market data can be difficult, especially in rural areas where internet and technology penetration is low.
- **Market Fragmentation**: In many regions, markets are fragmented and there is often a lack of coordination between smallholder farmers and other market participants, which makes data collection and sharing challenging.

Policy Research

Agricultural policies shape the environment in which farmers operate. Effective policy can promote sustainable practices, foster innovation and ensure equitable distribution of agricultural benefits. This research area focuses on understanding the constraints and opportunities in policy frameworks.

Key Areas of Research:

- **Constraints to Technology Adoption**: Identifying how government policies can either enable or restrict the adoption of innovative agricultural technologies particularly in terms of subsidies, land tenure and infrastructure support.
- **Input Access and Affordability**: Examining how policies can improve access to agricultural inputs such as seeds, fertilizers and machinery especially for smallholder farmers in rural areas.
- **Policy for Fair Pricing and Market Access**: Investigating policies that ensure fair pricing for agricultural products and improve farmers' access to both domestic and international markets.

- **Research-Extension Linkages**: Studying the effectiveness of policies that link agricultural research institutions with extension services to ensure that farmers benefit from the latest technologies and innovations.

Advancements:

- **Agricultural Policy Frameworks**: Designing national and regional agricultural policies that support sustainable practices, market access and fair competition while addressing the specific needs of vulnerable populations.
- **Public-Private Partnerships (PPP)**: Fostering collaboration between governments, private enterprises and farmers to create policies that drive inclusive agricultural growth and innovation.

Challenges:

- **Policy Gaps and Inefficiency**: Many developing countries struggle with inefficient policy implementation, which leads to poor outcomes for farmers especially in terms of access to markets, fair pricing and input availability.
- **Global Trade Policies**: International trade agreements can either facilitate or hinder access to global markets depending on how they are structured.

Innovative Commercial Agriculture

Innovative commercial agriculture is about moving beyond subsistence farming to creating profitable agribusinesses that add value through processing, marketing and technological innovation. This area of research focuses on exploring new business models and fostering entrepreneurship in agriculture.

Key Areas of Research:

- **Value-Added Processing**: Exploring opportunities for farmers and agribusinesses to process raw agricultural products into higher-value items. This could include processing fruits into juices or grains into flour which increases profit margins and job creation.
- **Agri-Entrepreneurship and Business Models**: Identifying sustainable and innovative business models that can be scaled and replicated across different regions. This includes contract farming, cooperatives and franchising.

- **Agribusiness Incubators and Support**: Developing agricultural incubators and accelerators that support entrepreneurs with the resources, mentorship and access to financing needed to succeed in commercial agriculture.

Advancements:

- **Agri-Tech Innovation**: Encouraging the development of agri-tech solutions that enhance production efficiency, reduce waste and increase sustainability such as precision farming and smart irrigation systems.
- **Contract Farming and Rural Industries**: Scaling up contract farming and establishing rural industrial clusters that process local agricultural products providing both market access and added value.

Challenges:

- **Capital Constraints**: Securing funding for innovative agribusinesses is often difficult especially for small-scale entrepreneurs due to high risk and long repayment periods.
- **Market Competition**: Smaller agribusinesses often face stiff competition from larger companies which dominate the value chains and markets.

Fallow Land Utilization Research

Fallow land, often left idle after years of farming, can be a valuable resource if managed correctly. This research focuses on finding productive uses for fallow lands to improve soil health and increase productivity.

Key Areas of Research:

- **Soil Restoration Techniques**: Researching techniques to restore the fertility of fallow lands, including agroforestry, crop rotation and organic farming practices.
- **Alternative Uses for Fallow Land**: Identifying non-cultivation alternatives for fallow land, such as grazing or growing bioenergy crops which could help farmers generate income while maintaining land health.
- **Sustainable Land Management**: Studying the long-term ecological and economic impacts of different land management practices to optimize fallow land use.

Advancements:

- **Agroecological Practices**: Encouraging the use of sustainable agricultural practices such as agroecology and agroforestry that integrate both food production and environmental conservation.
- **Precision Agriculture**: Using technology to monitor soil conditions and optimize the use of fallow land, ensuring it is restored to its full productivity potential.

Challenges:

- **Economic Viability**: Convincing farmers to invest in fallow land restoration can be challenging due to the upfront costs and the time required to see results.
- **Land Tenure Issues**: Many farmers are hesitant to invest in land restoration if they do not have secure land tenure or face uncertain land rights.

Investment Analysis of Agricultural Projects

Investment analysis is critical for understanding the financial viability of agricultural projects. Proper investment evaluation ensures that resources are allocated efficiently and that projects are sustainable in the long term.

Key Areas of Research:

- **Cost-Benefit Analysis**: Conducting detailed cost-benefit analysis of agricultural projects, taking into account factors like capital investment, operational costs, market conditions and risk.
- **Risk Management and Financial Products**: Developing risk management strategies for agricultural projects including insurance and hedging instruments to mitigate weather-related and market-related risks.
- **Impact Investment**: Studying how to attract impact investors to the agricultural sector with an emphasis on both financial returns and social/environmental benefits.

Advancements:

- **Innovative Financing Models**: The development of financial products such as weather-indexed insurance or blended finance mechanisms that provide capital for agricultural projects while addressing risks.

- **Agri-Financing Platforms**: Building platforms that connect agricultural projects with potential investors and lenders helping farmers access necessary funding.

Challenges:

- **Capital Access**: Farmers, particularly smallholders face challenges in accessing financing due to the perceived riskiness of agriculture.
- **Long Investment Horizon**: Agriculture is a long-term investment requiring patience and careful planning to manage risks and rewards effectively.

To merge the elements of market intelligence, policy research, innovative commercial agriculture, fallow land utilization and investment analysis, forecasting plays a pivotal role in guiding decisions at all levels of the agricultural sector. A well-structured forecasting system enables farmers, policymakers, agribusinesses, and investors to make informed decisions that maximize productivity, optimize resource allocation, and reduce risks associated with market volatility and environmental changes.

Key Takeaways:

- Accurate market forecasting helps in demand planning, pricing strategies and mitigating risks associated with market volatility.
- Policy forecasting ensures that agricultural systems are prepared for future changes in regulations, climate adaptation strategies and international trade policies.
- Technological advancements particularly in agri-tech allow for more precise forecasting of production outcomes, resource use and environmental impacts.
- Fallow land utilization forecasting helps optimize land use practices and predict long-term environmental and economic benefits.
- Investment analysis forecasting aids in predicting the financial viability of agricultural projects and improving access to finance reducing the risk for investors and farmers.

By improving market intelligence, formulating effective policies, fostering innovative commercial agriculture, utilizing fallow land and conducting comprehensive investment analysis, we can support sustainable agricultural growth that benefits all stakeholders. Moving forward, these research areas should be prioritized and refined to create resilient, profitable and equitable agricultural systems.