Introduction to Stata

Part 1: Introduction

Prakash Singh (Institute of Economic Growth)

Introduction to Stata

Econometric Softwares

- Eviews, Rats/Cats, Microfit, Limdep, SPSS, Gauss, Matlab, R, Gretl, Octave X-12-ARIMA, TRAMO/SEATS and so on…..

Proprietary software and open source/free software

Stata is

Integrated statistical analysis packaged.

- Strengths:-
- Ability to handle and manipulate large data sets (e.g. millions of observations! Depending upon the type of Stata you have).
- Large number of users which makes development of newer programme fast compare to other software.
- Stata has a nice Graphic User Interface and simply to use, (started with Stata 8)

- Weakness:-
- Proprietary in nature.
- Sensitive to Uppercase alphabets.
- Stata being a purely commercial software makes difficult to for user to upgrade to the latest version every time.
- Stata can only open a single dataset at any one time. Stata holds the entire dataset in (random-access or virtual) memory… so the computational ability is restricted to the computer capacity also.

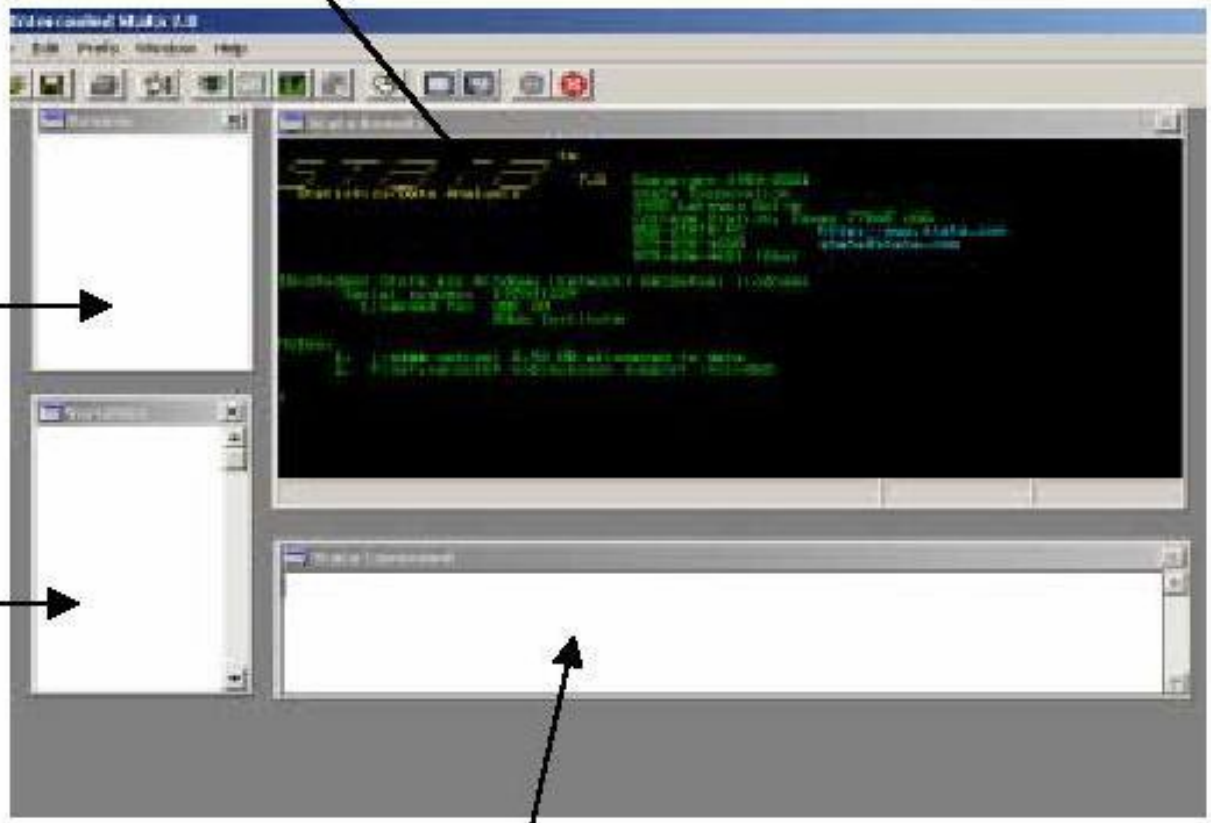
There are four major builds of each version of Stata

- ***Stata/MP:-*** for multiprocessor computers (including dual-core and multicore processors)
- ***Stata/SE:-*** for large databases
- ***Stata/IC:-*** is the standard version
- ***Small Stata:-*** is a smaller, student version

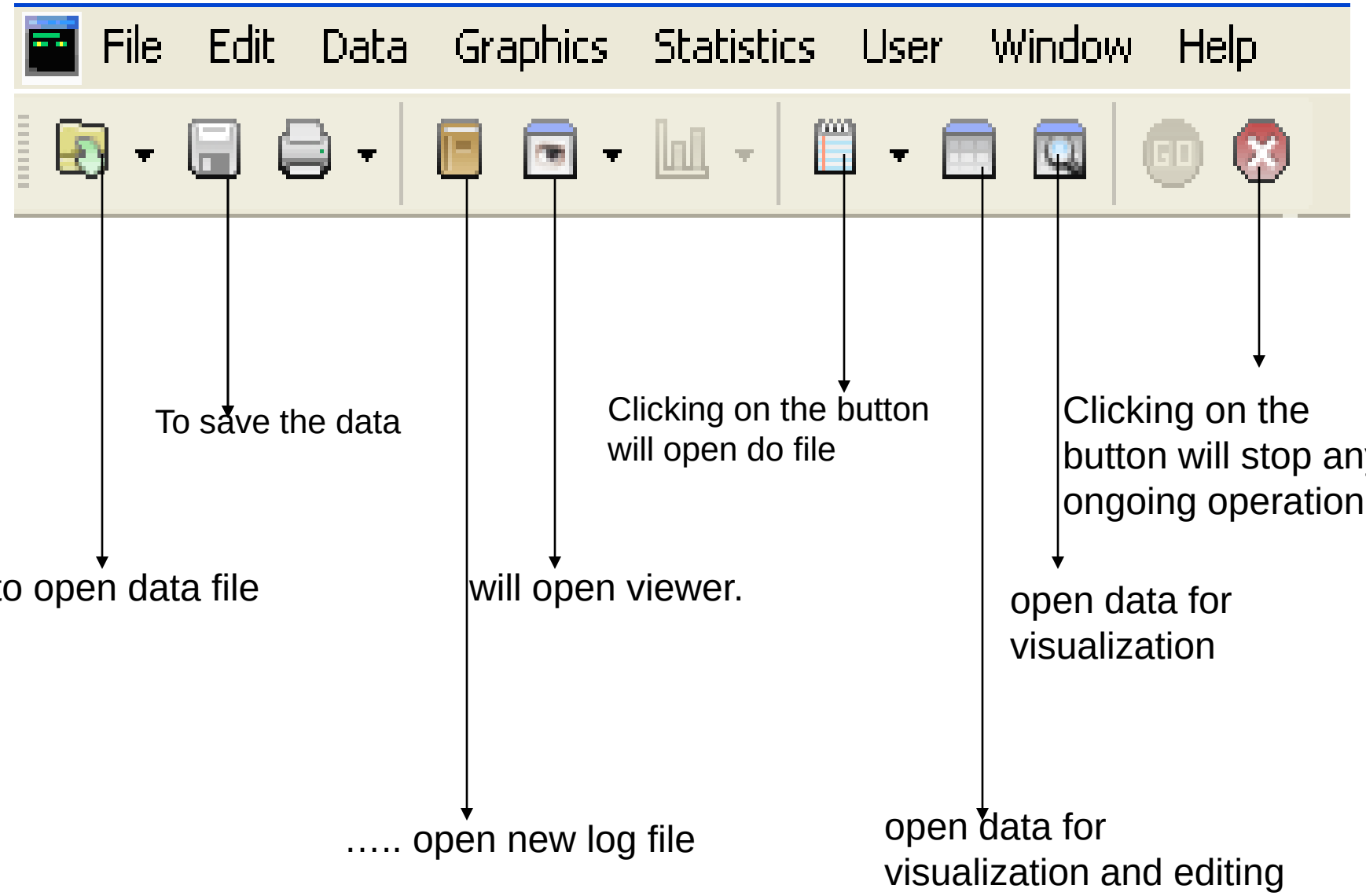
2. Results window – results are displayed here.

3. Review Window – past commands appear here
   a) Click on command in the review window, and it will appear in the command window.

4. Variables window – variable list appears here.
   a) Any variable labels will also appear here.
   b) Click on variable and it will appear in the command window.



1. Command window – commands are typed here, hit return to execute


File
Edit
Data
Graphics
Statistics
User
Window
Help
to open data file
To save the data
….. open new log file
will open viewer.
Clicking on the button will open do file
open data for visualization and editing
open data for visualization
Clicking on the button will stop any ongoing operation

Using permanent Stata data files

Bold words are the command to be typed in command box

- ***clear*:-** clearing Stata's memory, i.e. erases data from the memory of the software. Important for getting new data set into use.
- ***set memory*:-** allowing enough space for the data. Gives flexibility for changing the memory allocated to the software for handling data according to the size of data file.
- ***use*:-** copying the file into memory. Basically call the data file form source location.
- ***save*, *replace*:-** saving changes in the operational data set.

Describing the data

- **describe:** Describe data in memory or in file
- **summarize:** the variable in terms of noumber of observation, mean, std. dev, min, and max of variables
- **codebook:** more univariate statistics, i.e. describe data contents
- **tabulate:** provides frequencies and cross-tabulations of the variable(s).

Groups and subsets of data

- **if:** execute command for a subset of observations
- **sort:** order observations by the values of a variable
- **by:** execute command for groups of observations (requires sort)
- **in:** execute command for a range of observations relational, logical, and arithmetic operators missing values

Changing the data

- **replace:** change the values of a variable
- **recode:** change the values of a variable
- **rename:** change a variable name
- **label:** labeling variables, values, and data files
- **drop:** drop one or more variables
- **drop if:** drop observations conditional on one or more variables
- **edit:** editing the data file directly

Ways of running Stata

- There are the ways to operate Stata.
  - Drop Down Menu: Using drop down window we can specify the commands
  - Interactive mode: Commands can be typed directly into the Command window and executed by pressing Enter.
  - Batch mode: Commands can be written in a separate file (called a do-file) and executed together in one step.

Getting data into Stata

- From excel/spreadsheet or database program

  if the data is in excel format then one can either copy paste the data from excel or can save in csv format and then import using following command.

insheet using “path of the file\filename.csv”, clear

insheet using “path of the file\filename.txt”, clear

Other options: delimiter(), case

Use of infile command

Infile command is used to get data into stata if your data satisfy following precondition.

- The file should NOT have variables names on the first line.
- Character variables that have spaces in them, such as full names, must be enclosed in quotes.
- Numbers can have commas and minus signs, but not dollar or percent signs.
- **infile** command assumes that the variables have spaces between them and that there are no blank spaces where it expects data (missing numeric data need to be represented by something, either a number or a single period, '.').

infile name of the variables using “path of the file\file name.txt", clear

infile str13 make mpg weight price using “F:\STATA_IHD\auto4.txt”,clear

Use of infix command

- infix command is used to import data in stata memory if the data set is in fixed width format.

infix var name postition var name postition var name postition …. using **“path of the file\file name.txt”, clear**

Note: here position means column location length of the variable i.e. 1-5, 6-8 and so on.

infix str make 1-13 mpg 15-16 weight 18-21 price 23-26 using “F:\stata_class\auto5.raw”,clear

Creating directory

- *pwd:-* locates the present directory
- *cd:-* changes the location of the directory to new and drive:\folder “path of the forlder\” **the biggest advantage of directory file is that one can use and save… file without writing the full path of the file.**

Of course we can create a new directory “folder” within the system (computer)

Creating directory

- *mkdir:-* creats new directory “path of the forlder\name of the directory you want to mak\”
- *dir :-* list the files/content in the directory.

Data type

Numeric and string

Numeric Numbers are stored as byte, int, long, float, or double

Byte: integers of two or three digit (-127 to 100 generally)

Int: -32740 to 32740

Long: -2,147,483,647 to 2,147,483,620 (+/- 2.14 billion)

Floot: real number with seven digit of precision

Double: real number with 15 digit of precision

Str: for string variable

Save and Delete the stata data

save **“path of the file\file name.dta”**

save file name.dta

if dictionary is created or changed to the folder where you intend to save your data file

saveold “**path of the file\file name.dta**”

To save the data for reopening it in a lower or older version.

Save and Delete the stata data

If you are saving the data already saved in the desired folder the command save goes with replace in the end i.e.

save **“path of the file\file name.dta”, replace**

save file name.dta, replace

saveold “**path of the file\file name.dta**”, replace

and so on….

Save and Delete the stata data

To delete the data file from folder we use command

erase **“path of the file\file name.dta”**

Now once we have your data set into stata we can use

browse and ***edit*** command for browsing and editing the data

Log and cmdlog file

- log file: acquaints you to make and keep record of stata session. You can make log file both in smcl (.smcl) and text (.txt) format.
- cmdlog file (command log): acquaints you to make and keep a record of all the executed commands during Stata session.

Log and cmdlog file

To open and create a log file the command is

log using “path\log file name.smcl”

log close will close the log file i.e. recording of the session.

log off and ***log on*** will temporary off and the on the recording in the middle of the session.

log using can go with two sufix also, namely replace and append.

Log and cmdlog file

log using “path\log file name.smcl”, replace

Will record session but will overwrite the existing log file i.e. previous session record will not available.

log using “path\log file name.smcl”, append

Will record the session but will not overwrite the existing log file rather it will continue the existing log file record.

Log and cmdlog file

To open and create a ***cmdlog*** file the command is

cmdlog using “path\cmdlog file name.smcl”

cmdlog close will close the cmdlog file i.e. command recording of the session.

cmdlog off and ***cmdlog on*** will temporary off and on the command recording in the middle of the session.

cmdlog using can also go with two sufix also replace and append.

Log and cmdlog file

cmdlog using “path\cmdlog file name.smcl”, replace

Will record session but will overwrite the existing cmdlog file i.e. previous session command record will not available.

cmdlog using “path\cmdlog file name.smcl”, append

Will record the session commands but will not overwrite the existing cmdlog file rather it will continue the existing cmdlog file record.

Do File

- We have earlier seen the how log and cmd log file can be helpful in keeping the record of session. In most of the research work we do need to reproduce our result. This is one of the most challenging part as not only it requires taking note of what all you have done but also a lot of time. Do file in stata provides you the ability to store and reproduce the sequence of command executed for further use.

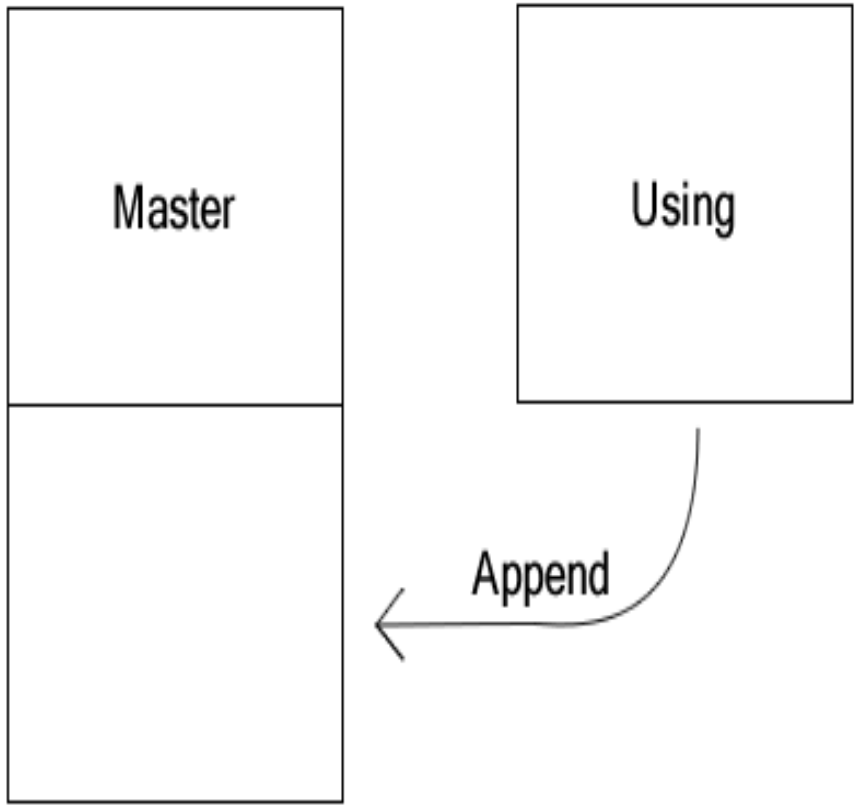
Appending data file

append:- appends a Stata-format dataset stored/saves on disk (existing data) to the end of the dataset in memory. If filename is specified without an extension, .dta is assumed.

Syntex to append two stata data files is

append using “path\file name”

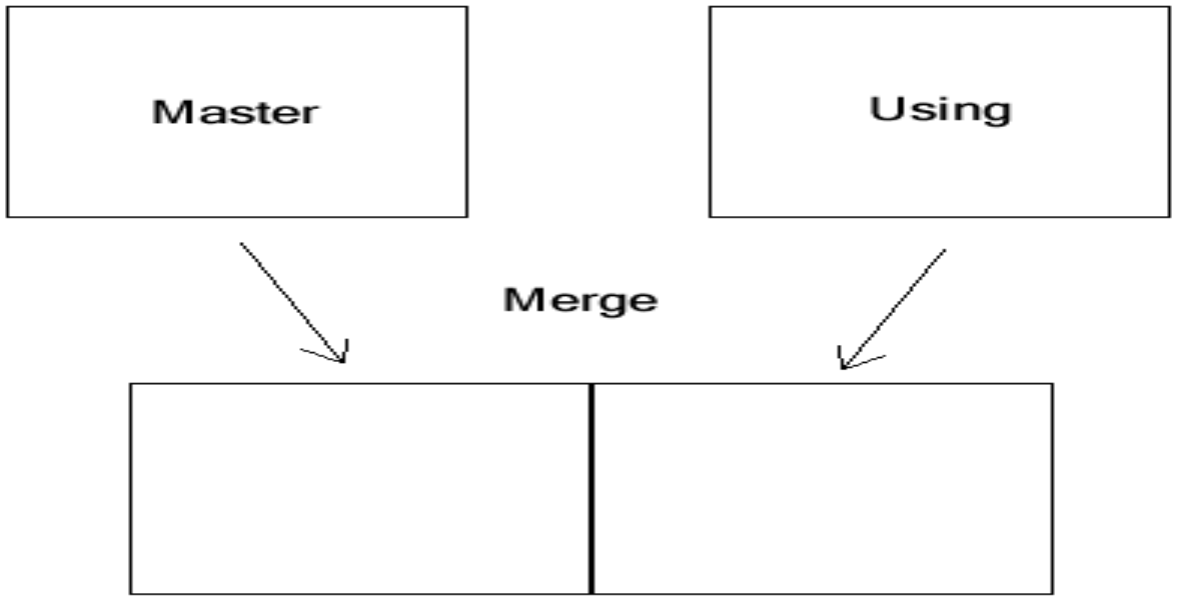
How append works



Merging data file

merge:- joins observations from the dataset currently in memory (called the master dataset) with those from Stata-format datasets stored as filename (called the using datasets). If filename is specified without an extension, .dta is assumed.

How merge works



Syntex
merge var using “path of data/file name”
tab _merge

Data management

- set memory/ set mem
- compress
- count
- browse
- edit
- list
- describe
- assert
- codebook
- summarize
- drop
- keep

Data management

set memory:- to change the allocated memory to stata, the default is 10md

set memory/set mem 200m

compress:- to compress the data set. Some times the computational limitation of your system and or stata hinders the data handling. In order to reduces the size of data set we compress the data set. set to handle the lagre data set.

compress

compress var

Data management

count is the command for counting the number of observation.

count

count command can also go with by and if i.e. counts the number of observation satisfying the condition.

count if rep78>4

by foreign: count if rep78>4

Note: use ex3 data

Data management

browse var with if condition, for example

browse a6a if a3==3

browse a6a if a3>1

Similarly, edit can also go with **if** option

edit var if condition, for example

edit a6a if a3==3

edit a6a if a3>1

Note: use ex2 data

Data management

list size a3 a3a

You can also get the total/mean/n of the variables. See ex-

list size a3 a3a in 1/5, sum(size a3 a3a) labvar(size)

You can also use if option with list command

list a3 a3a if size==1

Note: use ex2 data

Data management

- Assert command is used in validation of data.

assert k3bc==50

assert size<5

des *a will describe the variables which ends with a

des a* will describe the variables which starts with a

des a2x- b2a will describe the variables from a2x to b2a

Note: use ex2 data

Data management

summarize gives you the Summary statistics of the variable/variables in dataset with or without condition.

Data Manipulation

sort: arranges the observations into ascending order based on the values of the variables in *varlist*.

sort varname

gsort: arranges observations to be in ascending or descending order of the specified variables and so differs from sort in that sort produces ascending-order arrangements only.

gsort + varname

gsort - varname

Data Manipulation

generate:- is used to create variable

Use caselli_handbook.dta

gen tot_worker =yafao + ynona

**Generate command take all the algebraic expression to the right hand of "="*

gen ysch15_2= ysch15^2

**gen command also go with if exp*

gen ysch15_21= ysch15^2 if africa==1

**gen command also go with by exp*

by country: gen ysch15_2= ysch15^2

Note: before using gen command with "by" exp, do sort the variable to be used with by exp.

Data Manipulation

tabstat:- Display table of summary statistics

tabstat ysch15

tabstat ysch15, by(country)

tabstat ysch15, by(country) statistics(N mean sd min max)

tabulate oneway -- One-way tables of frequencies

tab country

tabulate twoway -- Two-way tables of frequencies

tab country oecd

- tsset year
- gen lgdp_1=l.lgdp
- gen lgdp_2=l2.lgdp
- gen lgdp_d=d.lgdp
- drop x2 x3 x4 X8
- drop x2-x8 x11-x15 x17-x23