UNIVERSITY OF NIGERIA, NSUKKA

FACULTY OF THE SOCIAL SCIENCES

PGC 601 LECTURE

TOPIC: INTRODUCTION TO STATA ECONOMETRIC SOFTWARE

Jonathan E. Ogbuabor, PhD and Anthony Orji, PhD

Department of Economics

University of Nigeria, Nsukka

jonathan.ogbuabor@unn.edu.ng; and

anthony.orji@unn.edu.ng

Introduction

- You will soon have to write your thesis/dissertation
- For that purpose, you might decide to so some empirical work using data and therefore an econometric/statistical software
- Some of these packages are general in nature, i.e. can do (almost) any kind of statistical analysis, e.g. STATA 16, SPSS 25.0, etc.
- Others are specialized packages:
  - Eviews 11, Gauss v19, Microfit 5.5 (for time series analysis)
  - R (for non parametric estimation)
  - GeoDA (for spatial data analysis)
  - LIMDEP (for categorical dependent variables), etc.
- There are thousands of specialized packages out there

Objectives of the Lecture

At the end of this lecture, participants should be able to:

- **Launch STATA**
- **Upload data into STATA**
- **Understand the command syntax in STATA**
- **Perform basic data management operations**
- **Perform descriptive data analysis using STATA**
- **Perform regression analysis using STATA**
- **Compare regression models and select an appropriate model**
- **Prepare and run do-files**
- **Generate and maintain log files**
- **Appreciate the purpose of *STATA Technical Bulletin***

The Data

We use the Educational Attainment and Earnings Functions (eaef) Dataset:

- The National Longitudinal Survey of Youth 1979 (NLSY79) is one of the most important data bases available to social scientists working with US data
- The eaef dataset is an extract from the NLSY79
- The eaef dataset is quite relevant for social policy as it enables analysis on the determinants of educational attainment and earnings, and how differences in these may be attributed: to sex, marital status, ethnicity, genetic endowment, etc; to interactions in the effects of these factors; and to changes through time

Examples of Micro Data for Nigeria

- **World Bank Enterprise Survey (World Bank)**
- **Demographic Household Survey (World Bank)**
- **Nigeria General Household Survey (World Bank/NBS)**
- **Interested students can practice the methods provided in this lecture using these data**

Input the Data into STATA from Excel

- **Open the Excel file named eaef (notice the variable names on the first row and the variable descriptions in the second sheet)**
- **Highlight the entire data and copy it (*ctrl C*)**
- **Launch STATA and notice its *outlook***
- **Follow the work flow: *data* ➡ *data editor* ➡ *data editor (edit)* ➡ *ctrl V (to paste)* ➡ *treat first row as variable names***
- **Follow the path *Data* ➡ *Variables Manager* to modify the properties of the variables e.g. label or name them as appropriate**

Basic Data Management Operations

- It is good practice to always view the data before beginning analysis
- Useful commands here include:

  list in 1/5 (to list the first five observation)

  describe (to describe the variables)

  summarize (for descriptive statistics of the variables, try *detail* option)

  tabulate age (to list each distinct value of the data)
- You can summarize a given variable by gender using the *table* command with *contents* option e.g.

  table female, contents (N earnings mean earnings sd earnings p50 earnings)
- The tabstat command provides more flexibility than the summarize command, e.g.

  tabstat earnings s hours, stat (count mean min p50 max sd skew kurt) col(stat)

  Generate the natural log of "earnings" i.e. "learnings" and repeat the tabstat command including "learnings"

  generate learnings=ln(earnings)

Basic Data Management Operations Cont'd

- **Instead of using the Variable Manager, you can also use the *rename* and *label variable* commands. E.g.:**

  rename ethblack black

  label variable age "age of individual"

Basic Data Management Operations Cont'd

- Use *generate* command to create new variables, e.g. *generate agesq = age^2*
- Use replace command to replace one or more values of a variable, e.g. *replace age=85 if age==37*
- Use *generate* command to create interactive variables, e.g. *generate baearn=educba*earnings*
- Use *save* command to save the data (include replace option to overwrite existing dataset, e.g. *save eaef1.dta, replace*

Basic Data Management Operations Cont'd

- **Sample Selection: Sometimes, we may want to restrict the sample or drop some observations from a given analysis**
- **E.g., we may want to restrict analysis to a particular age group or gender**
- **To do this, use the *if* qualifier after the command, e.g.**

su earnings hours if age>=35 & age<=40

su earnings hours if age<=39 | age>=41

su earnings s if male==1

Descriptive Statistics

- By now, we are already familiar with the *summarize* command
- The *correlate* command is used to compute the correlation matrix of the variables, e.g.

  correlate age earnings hour

- To obtain the statistical significance of these correlations, try the *pwcorr* command with the *sig* option

  pwcorr age earnings hours, sig

STATA Command Syntax

- **The *"command varlist, options"* structure is a fairly general structure that appears over and over in STATA**

For instance:

- **We use the *corr* command with the variable list "earnings s" and add the option *"cov"* to obtain the covariances rather than the correlation.**
- **We use the *summarize* command with the variable list "earnings s" and add the option *"detail"* to obtain detailed summary statistics**
- **Overall, a necessary step before any regression analysis is the use of summary statistics to gain some understanding of the data**

A Little Graphics Can Be Fun

- **Sometimes we would like to represent our data graphically.**
- **STATA has a very rich plotting capabilities.**
- **Let us begin with simple scatter plots. The general form is:**

  graph twoway scatter yvariable xvariable

  E.g.: ***graph twoway scatter earnings s, title("Scatter Plot of Earnings on Years of Schooling")***
- **If we want to look at differences between men and women, we need to use the *twoway* option.**

  graph twoway (scatter earnings s if male==1) (scatter earnings s if female==1)

Introducing Regression Analysis

- **Linear regression analysis is often the starting point of an empirical investigation**
- **It is often used to:**

  Summarize the data

  Make conditional predictions

  Test & evaluate the role of specific regressors
- **Here, we focus on basic regression analysis on cross-section data of a continuous dependent variable**
- **Our set-up is for a single equation and exogenous regressors**
- **We consider alternative models and some diagnostics**

The Regress Command

- The *regress* command performs OLS regression and yields: 1. coefficient estimates;
  2. standard errors; 3. *t* statistics; 4. *p*-values;
  5. confidence intervals; etc
- The syntax of the command is:
  regress depvar indepvars [if], options
- E.g., consider the simple regression model:

$$EARNINGS = \beta_1 + \beta_2 S + \varepsilon_i$$

regress earnings s

We split the sample:

regress earnings s if male==1

regress earnings s if female==1

What do the results show?

The Regress Command Cont'd

- **We move from simple to multiple regression**
- **Consider the inclusion of ability variable in the model:**

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 ASVABC + \varepsilon_i$$

regress earnings s asvabc

- **How has the results changed?**
- **Now, include the education of the father (SF) and education of the mother (SM) in the model:**

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 ASVABC + \beta_4 SF + \beta_3 SM + \varepsilon_i$$

- **How has the results changed? Include "exp"**

Postestimation Diagnostics

- Use the *estat hettest* command to check for heteroskedasticity
- Use the *estat ovtest* command for Ramsey regression specification-error test for omitted variables
- Use *help regress postestimation* to see all postestimation commands for regress
- To solve the heteroskedasticity problem, use the *robust* or *vce(robust)* option

Hypothesis Tests

- The *test* command performs hypothesis tests using the Wald test procedure
- We can test for equality of coefficients

$$H_0: \beta_4 = \beta_5$$

test sf=sm

- We can test for joint statistical significance of several variables

$$H_0: \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

test s asvabc sf

Suggested Hypothesis Testing Questions

- **The following questions are for those students who would want to *dirty* their hands**
- ***Do earnings depend on work experience as well as education?***
- ***Does the sex of an individual affect earnings?***
- ***Is the effect of education on earnings different for males and females?***
- ***Are there ethnic variations in the effect of the sex of a respondent on educational attainment?***

Reproducibility of your work is important

- This implies that you want to use do files and log files
- Do files are files that contain sets of STATA commands
- Log files are files where you save the output produced by STATA
- You should save the sequence of commands you have used for your empirical analysis and the output emanating from them
- Write your do files using comments so that you will be able to follow your do files even a year or two later
- Try to name and label your variables sensibly
- Try to label do file and output file (dta or log files) in a way that the names are helpful to remember their contents

Organize the contents of your do files

- **You should organize your work in different do files**
- **A good practice is to create a set of separate do files corresponding to different parts of your empirical work:**

1. **Data manipulation,**
2. **Descriptive statistics,**
3. **Empirical Analysis**

- **You need then a master.do file which summarizes the sequence in which you have to run the do files and give some comments on what each one does**

Do file for the data manipulation

- **Data manipulation includes:**

1. **loading the data set(s),**
2. **deriving new variables,**
3. **dropping unnecessary variables,**
4. **recoding (e.g missing & inapplicable cases as .),**
5. **labelling or renaming your variables,**
6. **selecting the estimation subsample(s),**

- **Try to name do files sensibly for example the data manipulation file could be called just data.do**

Do file to produce descriptive statistics

- A do file to produce some basic descriptive statistics is always useful
- To get to know mean, median, SD, min, max values and other features of each variable
- To detect if there are errors or anomalies
- Inspect each variable using the commands sum, tab and corr, and using graphs. This can help to identify potential mistakes you have made or anomalies in the data.

Do file to produce the main empirical analysis

Main Empirical Analysis may include

- **Estimation of models**
- **Diagnostic test of the models**
- **Testing procedure to compare different models**
- **Estimation replicated using different subsamples or different definition of variables**
- **Other types of sensitivity analysis**

Folders (directories) organization

- **It is good practice to organize your empirical work in a folder and to create three separate subfolders:**

1. **For the datasets**
2. **For the do files**
3. **For the log files**

Working with .log and .do Files

- Two types of files that are extremely efficient in Stata applications:
  - One stores Stata commands and results for later review (.log files),
  - and the other stores commands for repeated executions later.
  - The two types of files can work interactively, which is very helpful in debugging commands and in getting a good “feel” for the data.

.log Files

- One often wants to save the results of Stata commands and perhaps to print them out. Do this by creating a .log file. Such a file is created by issuing a "log using" command and closed by a "log close" command; all commands issued in between, as well as corresponding output (except graphs) are saved in the .log file. Use hh_98.dta.
- Assume that you want to save only the education summary of heads by household gender. Here are the commands:

```
. log using educm.log
. by sexhead, sort:sum educhead
. log close
```

.log Files

- What happens here is that Stata creates a text file named *educm.log* in the current folder and saves the summary output in that file.
- If you want the .log file to be saved in a folder other than the current folder, you can specify the full path of the folder in the .log creation command. You can also use the File option in the Menu, followed by Log and Begin.
- If a .log file already exists, you can either replace it with “log using educm.log” or replace or append new output to it with “log using educm.log, append.”

.log Files

- If you really want to keep the existing .log fi le unchanged, then you can rename either this file or the file in the .log creation command.
- If you want to suppress a portion of a .log file, you can issue a “log off” command before that portion, followed by a “log on” command for the portion that you want to save.
- You have to close a .log file before opening a new one; otherwise, you will get an error message.

A Sample Do File

```
*this is a sample do file
*our data is already loaded
/* so, we are good to go
first, we open the log file to store our results*/
log using output.log, replace
* we then set more off
set more off
*lets create some variables
generate learnings=ln(earnings)
generate smale=s*male
generate agesq=age^2
*lets compute some descriptive statistics
pwcorr earnings s hours age, sig
tabstat earnings s hours age, stat(count mean min p50 max sd skew kurt) col(stat)
*perform some regression analysis
regress earnings s hours age agesq
estat hettest
estat ovtest
regress earnings s hours age agesq, robust
*end the do file
log close
exit, clear
```

Saving a Data Set

- If you make changes in an open Stata data file and want to save those changes, you can do so by using the Stata “save” command.
- For example, the following command saves the hh_98.dta file:

```
. save hh_98, replace
file hh_98.dta saved
```

Saving a Data Set

- You can optionally omit the file name here (just “save, replace” is good enough). If you do not use the replace option, Stata does not save the data but issues the following error message:

```
. save hh_98
file hh_98.dta already exists
r(602)
```

- The replace option unambiguously tells Stata to overwrite the preexisting original version with the new version. If you do *not* want to lose the original version, you have to specify a different file name in the “save” command.

Exiting Stata

- An easy way to exit Stata is to issue the command “exit.” However, if you have an unsaved data set open, Stata will issue the following error message:

```
. exit
no; data in memory would be lost
r(4)
```

Exiting Stata

- To remedy this problem, you can save the data file and then issue the “exit” command. If you really want to exit Stata without saving the data file, you can first clear the memory (using the “clear” or “drop _all” command as shown before) and issue the “exit” command. You can also simplify the process by combining two commands:

. exit, clear

Stata Help

- Stata comes with an excellent multivolume set of manuals. However, the on-computer help facility in Stata is extensive and very useful; if you have access to the Web, an even larger set of macros and other useful information are available.
- From within Stata, if you know which command or keyword you want the help information about, you can issue the command “help” followed by the command name or keyword. This command works only if you type the full command name or keyword with no abbreviations.

Stata Help

- For example, the following command will not work:

. help mem

help for mem not found

try **help contents** or **search mem**

- However, this command will:

. help memory

[output omitted]

Stata Help

- If you cannot recall the full command name or keyword, or if you are not sure about which command you want, you can use the command “lookup” or “search” followed by the command name or keyword. So the following will work:

. search mem

[output omitted]

- This command will list all commands associated with this keyword and display a brief description of each of those commands.
- Then you can pick the command that you think is relevant and use help to obtain the specific reference.
- The Stata Web site (http://www.stata.com) has excellent help facilities, such as an online tutorial and frequently asked questions (FAQ).

Notes on Stata Commands

- Here are some general comments about Stata commands:
  - Stata commands are typed in lowercase.
  - All names, including commands or variable names, can be abbreviated as long as no ambiguity exists. For example, “describe,” “des,” and simply “d” do the same job because no confusion exists.
  - In addition to typing, some keystrokes can be used to represent a few Stata commands or sequences. The most important of them are the Page-Up and Page-Down keys. To display the previous command in the Stata Command window, you can press the Page-Up key. You can keep doing so until the first command of the session appears. Similarly, the Page-Down key displays the command that follows the currently displayed command in the Stata Command window.
  - Clicking once on a command in the Review window will put it into the Stata Command window; double-clicking it will tell Stata to execute the command.This can be useful when commands need to be repeated or edited slightly in the Stata Command window.

STATA Technical Bulletin (STB)

- **STB (ISSN 1097-8879) is a printed journal containing articles related STATA and software additions to STATA**
- **Accompanying each issue of STB is a software that can be installed into STATA free of charge from the internet**
- **STATA is backward compatible**
- **STB promotes communication among STATA users of all disciplines and levels of sophistication**
- **It contains articles written by STATA users, StataCorp employees and others**
- **It is a vehicle by which new features are first added to STATA and distributed**

Where to find resources to learn Stata

An excellent resource for getting started with Stata is provided by

- **Princeton and UCLA**
  http://data.princeton.edu/stata/
  http://www.ats.ucla.edu/stat/stata/
- **Stata introduction is downloadable from**
  http://www.stata.com/links/stataintro.pdf
- **Frequently Asked Questions site:**
  http://www.stata.com/support/faqs
- **Statalist user group**
  http://www.stata.com/support/statalist
- **Further stata learning sources**
  http://www.stata.com/links/resources.html
- **Use help and the help menu button and findit**
- **Try 'googling' your question**

Some References:

Some useful references for STATA programming are:

- **Cameron, A.C. & Trivedi, P.K. (2010) *Microeconometrics using Stata*, Stata Press.**
- **Cameron, A.C. & Trivedi, P.K. (2005) *Microeconometrics: Methods and Applications*, Cambridge University Press**
- **Greene, W.H. (2003) *Econometric Analysis*, Prentice Hall**
- **Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press**

Thank You