221.644.01 Econometric Methods for Impact Evaluation of Health Programs

Prof. Antonio J. Trujillo, Ph.D.

Lab 1

Caroline Hanson

March 30, 2017

**Objectives for Today**

1. Review essential Stata commands that will be necessary for completing the assignments
2. Familiarize ourselves with a dataset that will be analyzed as part of the Assignment 1
3. Start practicing data analysis relevant to program impact evaluation
4. Answer any questions you may have

Software

Just to reiterate, you are free to use any statistical analysis software, but in this course we will use Stata. If you don’t have a copy, please make arrangements to get one or make sure you can do your work on the lab computers (not recommended). All of the most popular software packages (Stata, R, SAS, etc.) have pros and cons.

Stata’s strengths in particular:

* Intuitive syntax;
* Designed with “rectangular” datasets as its main target;
* Useful econometric techniques built in (very economist-friendly);
* Has full graphing and matrix processing capabilities
* Relatively robust and has become extendable with the use of user-written ado-files.

Some of its weaknesses are:

* Loading the data in the RAM memory (therefore limited capability to work with “big data”);
* May not have the latest estimators or statistical techniques

**Workflow**

As a future program impact evaluator you want to make sure your analysis work is efficient, replicable, and easy to communicate. I highly recommend you build a workflow that is conducive to these goals. One book I found useful is *The Workflow of Data Analysis Using Stata,* by Scott Long (<http://www.indiana.edu/~jslsoc/web_workflow/wf_home.htm>). It has a good framework and some very useful ideas and tricks you can use for your own workflow.

As we move along in the course, I will show you some of my preferred approaches and tricks I use when writing Stata code. I AM BY NO MEANS AN EXPERT. I always learn new things and I love to “compare notes” with other analysts. I encourage you to do the same. I will try to be explicit in my demos, but if you don’t understand something PLEASE ASK! Also, if you have a better way of doing something, feel free to let me know – that way everyone can learn.

**Do Files**

I assume most of you are familiar with do files, Stata’s file format for writing code. We will use them almost EXCLUSIVELY to communicate with Stata. Do files are in fact nothing more than text files. Why are they good? For at least a few reasons:

1. Code portability
2. Analysis Replication
3. Efficiency
4. Documentation

**Commenting Code**

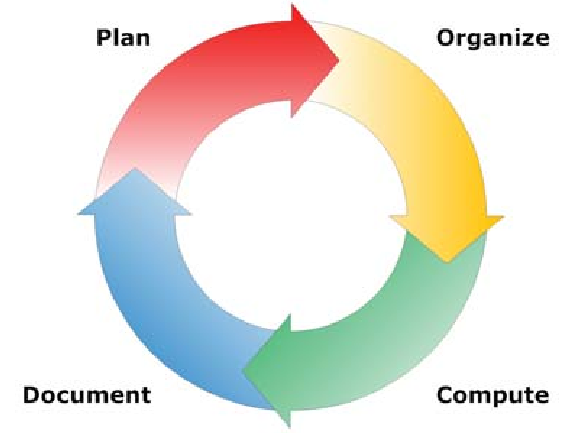
In Stata there are multiple ways to “comment out” code so that it is not run when the do-file is run:

* The “\*” symbol at the beginning of a line comments out a whole line
* The “//” symbol anywhere in the line tells Stata to ignore the rest of the line
* Using /\* *text on multiple lines \*/* you can comment multiple lines of text

**Some Useful Principles**

1. Keep your files organized – it is very easy to forget what you did a few days (or even hours) ago, let alone a few weeks or months
2. Document your analyses as you go – there is a trade-off between having detailed documentation and the cost to writing it, so try to figure out what is JUST ENOUGH for you
3. Write ROBUST do files – meaning files that minimize dependency on changes in software versions, computers, analysts, time of day, etc.
4. Minimize direct handling of data – instead, try to do any processing with code
5. Minimize direct copy/paste from the console – instead, try to export results automatically
6. Try to invest in learning some keyboard shortcuts and habits that speed up your analysis – it will be worth it in the long term

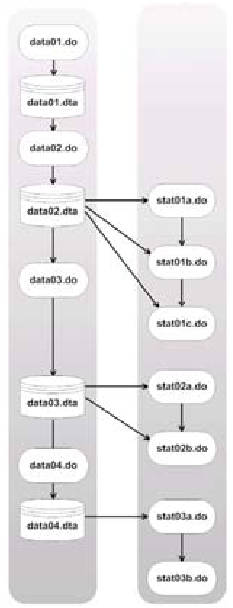
**The Ideal**



In reality, things are much more complicated and not purely iterative, with many moving parts and changing elements as the analysis progresses.

**Dual workflow**

I suggest separating your workflow into two parts: data management and data analysis. In the first one, you alter the structure of the data. In the second one, you obtain results but the data largely remains unchanged. This is not always easy, but I find it results in much better outcomes.



DATA ANALYSIS

DATA MANAGEMENT

**Essential Stata Commands**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Category** | **Exploring data** | **Processing/cleaning**  **data** | **Analyzing data** | **Exporting results** |
| **Most common commands for this class**  [There are many more out there!] | use  summarize  codebook  ameans  tabstat  tabulate  inspect  hist  scatter | save  gen / replace / egen  label [var]/[values]  sort / order  keep / drop  merge / append  collapse | reg  predict  logit/probit  margins  xtreg  xtlogit/xtprobit | est store  outreg2  esttab  tabout  putexcel |

**User-Written Ado Files**

Stata benefits from a repository of user-written commands. Think of them as additional packages that expand the functionality beyond the Stata core.

You can search for them by typing:

findit *name*

and if the command exists in the standard repository (which is hosted by the Social Science Center or SSC at Boston University), you can type the following command to install it automatically (you need an internet connection):

ssc install *name*

**Subsetting and Command Options**

In Stata subsetting is done mainly via the *if* and *in* statements. *If* is followed by a condition that selects a group based on the values of variable and *in* is followed by a range of observations. The *by* statement, where allowed, stratifies the command by groups determined by a variable or a variable list.

***Example:***

reg income education if age >=18

restricts the estimation sample to those who are older than 18

reg income education in 1/18

restricts the estimation sample to *observations* 1 through 18. This may change depending on how the data is sorted, so be careful!

Most commands also have options that, when specified, can substantially alter the behavior of a command. These options come after a comma in the command syntax:

***Example:***

tab race

tab race, missing

These two commands display different results. The second one displays the observations with missing values for race, whereas the first one ignores them.

**Today’s Exercise**

* We will use the file *wisdometalworkeddata.dta*
* The corresponding paper is :

Wisdom, J., Downs, J., Loewenstein, G. 2010 Promoting Healthy Choices: Information versus Convenience. American Economic Journal, Applied Economics. Vol 2(2), 164-178.

* I will start from scratch and demo some of the basic commands for data management and data analysis on this dataset
* Please follow along on your laptops and try to refresh your Stata skills
* Should not take a long time
* If you know this stuff already and want to leave, I will not be offended
* **If you have questions, PLEASE ASK!**