Programming with Stata - Stats II Introduction Sessions

December 2015

In the first session, students will receive an introduction to reproducible research and literate programming and learn some tools for applying what they have learnt to their work with Stata. Students will be taken through an exercise designed to strengthen their Stata skills. During the exercise, students will import, clean and analyse a messy dataset, learning how programming using Stata can help them accomplish difficult tasks and simplify easy ones. Students will receive a homework assignment which will require them to put their knowledge into practice. Students will have time to discuss what they have learnt while going over the homework assignment in the next session, where some further techniques and applications of programming with Stata will be discussed.

Some particular features of Stata to be included are listed below.

1 Reproducible Research and Literate Programming

Students should develop the tools necessary to produce research output that meaningfully links data, code and narrative interpretation.

1.1 About Reproducible Research and Literate Programming

- Explanation of the principles of Reproducible Research and Literate Programming
  - "The standard of reproducibility calls for the data and the computer code used to analyse the data be made available to others" [1]
  - Literate programming ties together data, code and the actual research output, enhancing reproducibility
- How these can help students avoid errors and unnecessary repetitive tasks

1.1.1 Options for Creating Documents that include Stata Output

- Why is copying, taking screenshots, and pasting into word suboptimal?

- Option 1: Use word's -link- or -includetext- fields to include log output that can be updated automatically
  - Text fields can include output from log files that automatically updates when you run your do file. However, you would have to run a script to clean these log files [Need to check if this works on school computers. Also, it may be easier to make this a stata function & this needs to be extended to control formatting. Perhaps it's possible to write a stata function that allows for including text and document writing instructions within a do file].
  - Word is easy to use but can be frustrating when what you want to do is very specific. Formatting can be an issue.
- Other options: using LaTeX or Statweave and LaTeX will be briefly mentioned as further options, but not required [we may still want to check software availability on school computers]

2 Programming with Stata

Students should develop a clearer understanding of how Stata works, expand their vocabulary of commands, and gain experience in solving problems and simplifying routines by programming Stata.

2.1 General Pointers

- The stata environment: command line, do file, data, logs
- Command syntax and return values
- Where to look for help, and how to understand it
- How to write a good do file

2.2 Directory Structure

- File paths and the working directory
- How thinking about directory structure can make things easier

2.3 Reading Data

- Reading data types not formatted for stata: xlsx, csv, etc. Overcoming issues with delimiters and unhelpfully formatted data files
- Reading data from web sources

2.4 Data transformation and processing

- Data types
- Scalars and matrices
- Applying computations conditionally
  - How to use -if-, how to combine if conditions
  - Using -cond()-
- Repeating computations
  - -For- loops
  - Using -by- to repeat computations across groups
  - Using macros to store results or instructions so that they can be reused
- Cleaning data
  - Handling missing data
  - -keep- and -drop-
- Transforming data
  - -generate-, -replace-, -rename-
  - -egen-, -egen- by
  - Counting from _n to _N
  - Sorting data using -sort- and -gsort-
  - Using the -by- command

3 Presenting Results with Stata

Students will learn how to manipulate the results of regression analyses to present well-formatted tables and graphs.

3.1 Post-estimation

- Accessing and using post-estimation results in e()
- Using -predict- to generate new variables

3.2 The -estout- program

- Storing regression output and presenting using -esttab-
- Computing additional statistics to add to results tables
- Saving tables to various file formats for use in output

3.3 Producing Graphs

- Different types of graphs using stata
- Combining graphs
- Customising the appearance of graphs in stata
- Saving and presenting graphs in your output

Some Further Resources

- `http://data.princeton.edu/stata/`
- `https://github.com/HertieDataScience/SyllabusAndLectures`

References

[1] Roger D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.