

ICPSR Summer Program

Managing Data for Reproducible Results

July 24 – July 28, 2017

Scott Long, Indiana University - jslong@indiana.edu

Nic Bussberg, Indiana University - nbussber@indiana.edu

This workshop presents methods for planning, organizing, documenting, and executing sophisticated quantitative analyses that are reproducible. The social sciences are paying increasing attention to reproducible results with the expectation that other researchers have access to your analysis files so that they can analyze your data and obtain exactly the same results. Creating reproducible results depends on a coordinated workflow that begins with the conceptualization of the research and ends with the preservation of files. This workshop provides you the tools to do this efficiently. Topics include creating datasets with metadata documentation; writing robust script files; methods for organizing and preserving files; efficient approaches to documentation; and tools to maintain the provenance of statistical results. While we use Stata in class, the methods apply to any statistical package.

Suggestions for the workshop

1. You should have some familiarity with Stata. If you have not used Stata, please look at the introductory videos on Stata's YouTube channel (Google: youtube stata channel).
2. Bring a laptop to class. If you do not have Stata, we will install a temporary Stata license for use during the class. If you want to use other packages during the class, you need that package installed on your laptop.
3. Bring an external drive or USB stick for the files you use in class. This limits the chances of losing files on your laptop as you experiment with new methods.
4. Many attendees like to apply the methods to their own data during the workshop. Feel free to bring your own data and script files (e.g., do-files in Stata). *If you plan to work on your own files during class, back these up before coming to ICSPR!*
5. In the past, the classroom varied between too hot and too cold. Dress accordingly!

Texts

Scott Long. 2017. *Lecture and Lab Notes for Strategies for Reproducible Research*. The lecture and lab notes are all you will need during the workshop. Electronic copies will be available.

For after the workshop, I recommend these books. You might want to look at them during class, but that is not necessary.

Long, J.S. 2008. *The Workflow of Data Analysis Using Stata*. Stata Press. College Station, TX.

Wong, Dona M. 2010 (2nd edition). *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures*. Highly recommended for tables and graphs.

Baum, C. F. 2016. *An introduction to Stata programming*. Stata Press. College Station.

Tentative schedule

Each day will include a combination of lectures, labs, discussion, and independent work. Lectures begin at 9:00am and formal class sessions will end by 4:00pm. The instructors will be available from 8:30am till 5:30pm.

Monday

- Part 1: Introduction
- Part 2: Tools
- Part 3: Digital asset management
- Part 4: Protecting files
- Part 5: Getting started with Stata

Tuesday

- Part 6: Planning, organizing and documenting
- Part 7: Workflow for computing
- Part 8: Using do-files
- Part 9: Macros and returns

Wednesday

- Part 10: Dataset
- Part 11: Importing data
- Part 12: Variables
- Part 13: Loops
- Part 14: Extended WF for names and labels (as time permits)
- Part 15: Debugging

Thursday

- Part 16: Cleaning
- Part 17: Adding variables
- Part 18: Analysis

Friday

- Part 19: Presentations
- Part 20: Replication
- Part 21: Review of Workflow
- Discussion