Strategies for Reproducible Results

Scott Long, Indiana University
jslong@indiana.edu | www.indiana.edu/~jslsoc/

ICPSR Summer Program in Quantitative
August 1 – August 4, 2016

This workshop will show you how to plan, organize, document, and execute sophisticated quantitative analyses that are reproducible by other researchers. The social sciences are paying increasing attention to reproducible results—the expectation that other researchers have access to your data and analysis files so that they can reproduce your work to obtain the same results. Increasingly, journals require authors to deposit the datasets and script files that produce the findings from the paper. While these changes benefit the scientific enterprise, they place new demands on authors. 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 will provide you with all of the essential elements to do just that. Topics covered in the class include creating datasets with metadata documentation, writing robust script files that allow results to be reproduced, methods for organizing and preserving files, efficient ways to document research, and how to maintain the provenance of statistical results. While the strategies introduced apply to any software, the course illustrates these ideas using Stata and labs give attendees a chance to try these methods. Attendees are asked to bring a laptop. If you have Stata that's great, if not a temporary installation will be provided. If you prefer, ICPSR will also provide a computer for your use during the class.

Getting ready for the workshop

1. **Back up any files you plan to use during the class!**
2. You should have some familiarity with Stata. If you have not used Stata or would like a refresher, look at the introductory videos on Stata’s YouTube channel.
3. Ideally, bring a laptop along with copies of your files. If you do not have Stata, local staff can install a temporary Stata license for use during the class.
4. Bring an external drive or USB stick to hold the files you use during the class. This prevents doing something to the files on your laptop that you will later regret!
5. It is useful to have a real data to experiment with. If possible bring copies of files from prior research, including any script files (e.g., do-files in Stata). You won’t be able to finish serious work with your data, but you can use it to try things.
6. Try to reproduce the results from earlier research you have done or from assignments in a statistics class. Or, try to replicate results from a published paper. Did it work? What would have made it easier?

Texts

Scott Long. 2019. Lecture and Lab Notes for Strategies for Reproducible Research. These handouts are copies of the overheads used in lectures and materials for lab. The lecture and lab notes are all you will need during the workshop. Electronic copies will be available.
After the workshop, the following books are recommended. You might want to look at them during class, but that isn’t necessary.


**Schedule**

The schedule could change depending on the background of students, local facilities, and how tired you get! Mornings will be primarily lectures, with computer work and lectures in the afternoon.

**Monday**

*Part 1: Introduction*
*Part 2: Tools*
*Part 3: Digital asset management*
*Part 4: Protecting files*
*Part 5: Getting started with Stata*
*Part 6: Planning, organizing and documenting*

**Tuesday**

*Part 7: Workflow for computing*
*Part 8: Using do-files*
*Part 9: Macros and returns*
*Part 10: Datasets*
*Part 11: Importing data*

**Wednesday**

*Part 12: Variables*
*Part 13: Loops*
*Part 14: Extended WF for names and labels*
*Part 15: Debugging*
*Part 16: Cleaning*

**Thursday**

*Part 17: Adding variables*
*Part 18: Analysis*
*Part 19: Presentations*
*Part 20: Replication*
*Part 21: Review of Workflow*