Data analysis is technically demanding, time consuming, and hard work. A dataset must be prepared, statistical analysis performed, and results incorporated into papers. Invariably, reviewers want revisions that require additional passes through the data. Increasingly, journals expect authors to distribute the datasets and the script files that produced the paper’s results. If others find errors in these files, the article might be retracted.

This workshop presents a workflow that is guided by the demands of producing accurate and reproducible results as efficiently as possible. Using this approach, work is completed faster, findings are more trustworthy, and the results are reproducible. Non-trivially, the process is easier and less stressful. Topics considered in the class include:

- Planning, organizing, and documenting your work.
- Methods to manage, document, and preserve digital files.
- Computing strategies that support reproducibility by documenting the provenance of results.
- Tools for writing robust programs that use automation to increase accuracy and efficiency.
- Methods for preparing datasets that use consistent names and labels, have metadata for documentation, and verify that variables are correct.
- Strategies for sophisticated data analyses that are reproducible and efficient.
- Methods to accurately and quickly incorporate results into a paper.
- Strategies and techniques that simplify the revisions of papers.

The principles taught in the workshop can be used with any statistical package. Examples of applying these methods are illustrated using Stata. In lab you can explore how to the same ideas can be implemented in other software (please have the software you want to used installed on your laptop). Lab also provides time to discuss how ideas from lecture can be applied to your work. A handbook is provided to help you adapt materials from the workshop to your own research and to guide the development of your workflow after the workshop ends.

**Web links about the class**

- [www.icpsr.umich.edu/icpsrweb/sumprog/courses/0235](http://www.icpsr.umich.edu/icpsrweb/sumprog/courses/0235)
- [www.indiana.edu/~jslsoc/teaching_icpsrworkflow.htm](http://www.indiana.edu/~jslsoc/teaching_icpsrworkflow.htm)
Suggestions for the workshop

1. If you are not a Stata users, look at the introductory videos on Stata's YouTube channel (Google: YouTube Stata channel).

2. Bring a laptop. 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, 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 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. 2018. 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.

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

**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 are available from 8:30am till 5:30pm. Class ends at 3PM on Friday.

**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: Robust and legible scripts, syntax, and do-files
- Part 9: Macros and returns

**Wednesday**
- Part 10: Datasets
- Part 11: Importing data
- Part 12: Variables and labels
- Part 13: Loops
- Part 14: Extended WF for names and labels (as time permits)
- Part 15: Debugging programs that won’t run or produce incorrect results

**Thursday**
- Part 16: Cleaning datasets and verifying the accuracy of variables
- Part 17: Adding new variables
- Part 18: Data analysis

**Friday**
- Part 19: Presentations and revisions
- Part 20: Replication
- Part 21: Review of Workflow
- Discussion

wficpsr18 syllabus 2018-04-23.docx