Lab Guide: ICPSR Workshop on *Managing Data for Reproducible Results*

Scott Long and Nic Bussberg | July 24 – July 28, 2017 wficpsr17 labguide 2017-07-14.docx

These activities reinforce ideas from lecture. Since there is not enough time in class to complete all of the exercises, use them as a starting point for work that you will do after the workshop ends. While it might be useful to experiment with your own data, *make sure you have multiple backups of your files and do not expect to get "real work" done!*

These exercises are designed to get you started. I want you to explore other ideas, evaluate your own workflow, and try those things from lecture that you think will be most useful for you.

Files for lab

1. Files for the class are in \WFclass with subdirectories:

   o \WFclass\Resources has PDFs of class materials (e.g., lectures, handouts).

   o \WFclass\Work has files for Stata exercises. Hereafter, simply \Work. This will be your working directory in Stata.

   o \WFclass\Work-nonStata has files for non-Stata exercises. Hereafter, simply \Work-nonStata

   o \WFproject is a second project folder used in exercises.

2. Accessing these files:

   o You can obtain the files from http://www.indiana.edu/~jslsoc/ftp/ where they are zipped.

Install these Stata packages

   o Run `search spost13_ado` and install.

   o Run `ssc install estout, replace`

Part 2: Tools

You might not have time during the workshop to install any software except for Stata. After the class, I encourage you to try the following.

1. Install and use a dual-pane file manager: Total Commander for Windows or Path Finder for Mac.

2. Install and use a macro program such as AutoHotKey (free) for Windows or Automator (free) or Keyboard Maestro for Mac. Start by creating a macro to insert the date.

3. If you do a lot of programming, learning a full featured text editor is a worthwhile investment. Two excellent freeware programs are Notepad++ for Windows and TextWrangler for Mac.

Part 3: Digital asset management

I do **not** recommend that you reorganize files during the workshop! Instead, get familiar with the directory structures that I propose and *plan* the changes you want to make after the class.

1. Using your file manager, explore \WFclass.

2. In \Work-nonStata\p03dam modify the templates for the general directory structure and the project directory structure to meet your needs. You can use a spreadsheet to plan your directory structures.

After the workshop

1. Take an inventory of your files from all computers and storage devices that you use.

2. Think about re-organizing your hard drive. Does the structure in \p03dam make sense? How do you want to modify it? Experiment with reorganizing the structure.

3. Plan how to populate the new directory structure with your files.

4. Before making any changes, study the section of WFDAUS that describes moving into a new directory structure. Back files up before you begin!

Part 4: Protecting files

Think about your current workflow for protecting files, but do not start new procedures during the workshop. Talk to your local IT support staff about what resources are available in your workplace.

1. Diagram your current procedures for protecting files.

2. Create a revised plan. Diagram how your plan will work.

After the workshop

1. Take an inventory of your files and how they should be protected. A spreadsheet is useful.

2. After the inventory, review your plan.

3. Implement the plan and stick with it!

Part 5: Getting started with Stata

1. Open Stata and practice changing the working directory from \WFclass\Work to \WFproject\Work.

2. For lab exercises your working directory should be \WFclass\Work

   o Suppose your USB drive is drive U. Then, `cd u:\wfclass\work`

   o Suppose your files are in c:\Desktop\WF. Then, `cd c:\desktop\WF\WFclass\work`

3. From the command line enter `do cda14-gettingstarted.do`. If the program does not run, you are probably not in the correct working directory. Enter `pwd` to find out where you are and `dir` to see files in the working directory.

4. If you are new to Stata, start with *cda2014 StataGettingStarted 2014-06-04color.pdf* located in \Resources. Alternatively, watch YouTube videos on the Stata channel (Google: stata youtube).

5. Familiarize yourself with the `help` command. For example, `help regress`. Click on the blue `regress` to open the PDF manual.

Part 6: Planning, organizing and documenting

1. For most people, digital asset management (DAM) is the biggest challenge for organizing work. I suggest you focus on DAM in class. After class, consider other aspects of how you organize your work.

2. During the workshop think about changing your workflow to prioritize planning. That is, plan how to plan!

3. \Work-nonStata\p06pod has a Word diary template for Windows and Mac. Copy one of these to be your diary for the class. Use this diary to experiment with heading levels and keystrokes to create headings. Experiment with the navigation pane. You might need to tweak the document depending on the word processor you use.

4. Pick a current project. Start a diary for the project and begin planning.

5. Create a Workflow Diary for keeping notes on your workflow. Practice using the diary throughout the week to document your workflow. Where will you save this document?

6. Think about how the reinforcing levels of documentation discussed in lecture apply to your work. What changes do you need to make to create more effective documentation?

Part 7: Workflow for computing

Think about the computing workflow

1. I want you to think carefully about how the computing workflow operates. I don't want you to do any analyses, but to imagine what you would do. Make up a hypothetical project where you create a dataset and run analyses. What do-files will you need? How does the dual workflow apply? How are the files named? How will the posting principle be applied? Later questions ask you to evaluate this workflow.

2. Does the dual workflow make sense for your research? If not, why?

3. Does posting make sense? How will you make sure it happens? If you don't plan to post, why?

4. How does run-order naming work? Decide on a naming template.

5. Think about how your past work would be modified using the computing workflow.

6. What parts of the computing workflow don't make sense? What problems do you anticipate?

Try the posting workflow

1. \Work-nonStata\p07computewf has a directory structure where \Active\CWF is a project directory for experimenting with the posting workflow.

2. Practice *moving* the files in \CWF\Work to the \PrePosted, \Posted, and \Vault.

3. Focus on understanding the process by which files are copied and moved as you complete work.

4. Later we will see examples of a dual workflow and run order naming.

Part 8: Using do-files

1. Customize `wflec-template-dofile.do` for your own use.

2. `wflec-do-step1-fragile.do` and `wflec-do-step2-fragile.do` are <u>not</u> robust. `wflec-do-step1V2-robust.do` and `wflec-do-step2V2-robust.do` are robust. Start with the fragile files and make them robust and legible. Compare them to the robust versions I provide.

3. How do these ideas apply to your work? If you have time, make copies of some old do-files and make the copies robust and legible.

Part 9: Macros and returns

1. Complete the exercises in *Guide to Automating Your Work in Stata* in \Resources for display and macros. Start the section on returns if you have time.

2. Things to try:

    a. Use **display** to display the local **tag** in your do-file.

    b. Define a set of nested models using locals and estimate them. Use **display** to list the sets of variables.

3. `wflec-auto*.do` are examples from lecture and more. Explore the ones that would be useful to you. The `wfauto-*.do` files are from the automation guide.

Part 10: Datasets

1. Use `rr-issp-family02.dta` in Stata. Run `notes` to see how metadata was used to document the dataset.

2. `wflec-data-provenance*.do` creates a series of datasets using metadata to document provenance. Review and run them. Display notes in the last dataset and trace the provenance.

3. Revise `wflec-template-datasets.do` to be your template for saving datasets.

4. Merge two datasets using various methods. The files in `wflec-data-merge*.do` provide examples*.

Part 11: Variables

1. `wflec-var-commands.do` illustrates commands for adding metadata to variables.

2. Experiment with the commands and applying them to other variables.

3. Create a template `template-var-commands.do`. Where will you save this?

4. Begin planning the names and labels for your own research.

Part 12: Loops and returns

1. Complete the exercises in *Guide to Automating Your Work in Stata* for loops and returns. You are likely to need to do more work on these after the class ends.

2. There are sample do-files `wflec-auto*.do` and `wflec-auto-*.do` in /Work.

3. Things to try:

    a. Use a loop to list a sequence of names and another to list a sequence of numbers.

    b. Use a loop to summarize variables. Use `display` to echo the `summarize` command.

    c. Using `k5` in `wf-lfp.dta` create binary variables indicating 0 vs. more children; 0 or 1 vs. more; etc. Add labels and tabulate `k5` against the new variables.

4. Think of ways to use loops for data management and analysis in your research.

Part 14: Debugging

1. You will get lots or practice debugging you own work later! Here you can replicate the example used in lecture.

2. `wflec-debug-graph1.do` has an error in the `graph` command. Try to debug the program following the methods from lecture.

3. As time permits, look at the other `wflec-debug-*.do` files.

Part 15: Cleaning

1. `wflec-clean-graph-loop.do` creates graphs for cleaning variables. Explore these commands. Adapt the do-file to other data. Modify `wflec-template-clean-graph-loop.do` for you own use. What should you name the file and where should you save it?

2. `wflec-clean-examples.do` has example of commands for cleaning data. Run this file and create new do-files to explore other datasets.

    • If you are using a Mac, `use wflec-add-examples-mac.do`

3. Modify `wflec-template-clean-graph-loop.do`, `wflec-template-clean-list-if-missing-loop.do`, and `wflec-template-clean-list-random-cases.do` for you own use. What will you name them and where will you save them?

Part 16: Adding variables

1. `wflec-add-examples.do` has examples of adding new variables. Run the do-file and study how it works. Make changes to experiment with the commands.

    • If you are using a Mac, run `wflec-add-examples-mac.do`

    • If you get an error in section #8 your version of Stata might not have a large enough `set matsize` #. You can comment out the section of code that produces an error.

2. `wflec-add-russia*.do` is a series of do-files that add variables in steps. Study the sequence of files and how `notes` is used. Make changes to experiment with the commands. Run the master do-file.

Part 18: Analysis

1. `wflec-analysis-collect*.do` has examples of collecting results from analyses. Explore the ones that look useful. Extend the ideas to your own analyses.

2. `wflec-analysis-local-specification.do` uses locals to specify models. Explore the code and extend it to your own analyses.

3. `wflec-dual*.do` illustrate the dual workflow for data and analysis. Study how the do-files work. Add additional steps.

Part 19: Presentations

1. The document *template-paper-2016-02-09.docx* in \Work-nonStata\p19presenting is a template for a paper that uses hidden text. It was created in Windows and might not format correctly in Mac. Try adding hidden text to document provenance.

2. `wflec-present-graph-formats.do` creates graphs in various formats. Add these graphs to a document and examine how they print (see \Work-nonStata\p19presenting*templates-stata-graph-formats-2017-06-20.docx*). If we don't have convenient access to a printer, you will have to try this at home. Before you create graphs for research, decide which format works best so you don't need to redo your graphs later.

   - If you are using a Mac, run `wflec-present-graph-formats-mac.do`

3. `wflec-present-provenance-caption.do` illustrates adding captions to graphs.

   - If you are using a Mac, run `wflec-present-provenance-caption-mac.do`

4. `wflec-present-tables-esttab.do` is a simple example of using `esttab` to create a table.