a WORKFLOW primer

An introduction to WORKFLOW

In order to be robust & replicable, all research & related analysis requires a process of documented decision-making. "Workflow" is the general process of documenting the myriad of decisions that undergird the research process.

As we believe the process of implementing & abiding by a strong workflow is essential to good data analysis, we ask that ICPSRCDA15 students abide by a modicum of workflow rules in their coursework. This document provides an overview of some of the most important workflow ideas. More detailed information on workflow fundamentals in Stata can be found in *The Workflow of Data Analysis Using Stata* (Long, 2009).

A recommended directory structure

If you can't find your files, you can't fix your files, or replicate your files, or publish your files. A clean, clear directory structure is a key part of your workflow and, subsequently, your success in this class. Prior to starting your first assignment, we recommend implementing a full directory structure for your anticipated CDA work. A model is provided below.

****N.B.: As the computers as Newberry do not provide secure or private storage space, we HIGHLY recommend saving all files on a personal jump drive or portable hard drive.**

…/Thesis
…/Coursework SP17
…/Coursework ICPSR17
/MLE
/Bayesian
/CDA
/Assignments // once complete, move files from 'work' to here
/01Math
/02BRM-Pt1
/03BRM-Pt2
/04T&F
/05ORM
/06MNLM
/07Count

/Data
/Original // for data as its been downloaded
/Cleaned // for data you save after making changes

/LabResources
/Readings
/StataGuide
/Working // keep active assignments, do-files & data here

The fundamental ideas of *Posting and Run Order*

***Posting* a file:** Posting refers to moving a file from your working directory into a specific folder (e.g., the 02BRM-P1 file). Files [do-files, log-files, assignments] should **ONLY** be posted when **COMPLETE** (i.e., when your assignment is ready to be handed in). Once a file is posted, it should **NEVER** be changed. Prior to posting, files should be considered working files & should be kept in your 'working' directory.

***Run order*:** For assignments or project that require multiple do-files, files should be designed to run in a specified order—i.e., the do-file used to clean the data for a paper should be run before the analysis do-file that presupposes the clean data. Naming can be a useful way to indicate run order. While we do not require you to use multiple do-files for your assignments in this class, the idea of run-order should be considered essential for larger research projects.

- **For example:**
  `Icpsrcda-a2-1-clean.do` /Cleaning file for assignment 2
  `Icpsrcda-a2-2-analysis.do` /Analysis file for assignment 2

a few Workflow rules

Organization:

RULE #1: NEVER change a file after it has been posted. If you discover a mistake after a file has been posted, move the file back to your working subdirectory, **rename** (e.g., "`icpsrcda-orm-01cleanV2.do`") & post this new file when it's complete.

- **HINT:** Keep all 'in progress' assignment files in your 'Working' sub-directory until complete; then move to the appropriate 'Assignment' file.

RULE #2: NEVER name anything **'Final'**. Because in all likelihood, it's not ☺

Computing:

RULE #3: Don't hardcode directories/directory paths within your do-files. Rather, set your 'working' subdirectory as your central directory before you open or run any files in Stata.

- **HINT:** Be sure to include **both** the data & relevant do-files in this directory before running.

RULE #4: As best practice, do-files should include: (a) version control; (b) info section, including the do-file name; assignment #; your initials; & date; (c) appropriate header.

- **HINT:** Use the do-file templates as examples (both available in the class folder on Z:\ drive):
  Step 1—Prepare data: `prep_template.do`
  Step 2—Analyze data: `analysis_template.do`

RULE #5: ALWAYS save data that has been modified under a new name [e.g., source: `wls.dta` > modified: `icpsrcda-wls-orm.dta`].

RULE #6: NEVER change (rescale, recode, etc.) a variable without giving it a new name.

RULE #7: All new variables should have:

(a) ***variable label***

(b) categorical variables must have ***value labels*** {e.g., 1=1Single; 2=2Married; 3=3Divorced/separated/widowed, etc.}
HINT: For value labels, include the value of the category in the value label [e.g., for four-category Likert scale: 1SA; 2A; 3D; 4SD]
HINT: Be sure to keep variable labels & value labels to ≤ 8 characters to avoid truncation.

(c) a ***note*** indicating how variable was constructed, by whom & on what date.

RULE #8: Give **BINARY** variables ***positive, easily interpretable*** names. For example, if your binary variable is coded 1=1Female and 0=0Male, name the variable "Female" not "Gender." This prevents having to remember which category of gender was coded as 1 for the analyses and also minimizes interpretation mistakes.

RULE #9: Use **comments** throughout your do-file to indicate different tasks—e.g., in this course, indicate the question that is being answered.

RULE #10: Ensure do-files are **robust** (i.e., give the same results every time).

- **HINT:** Make do-files self-containing—e.g., don't open a data file and make changes in one do-file, and then perform analyses on the modified data in another do-file without first saving the modified data under a new name and reopening it in the analysis file.

Documentation:

RULE #11: To avoid output that is difficult to read, don't let do-files or log-file output 'wrap.'

- **HINT:** Use `set linesize 80` command in header of do-file & keep columns to <80 of do-file to ensure logs don't wrap.

RULE #12: Use a fixed font (e.g., `Courier` or `Courier New`) when copying output from log files to assignments.

RULE #13: Although not a requirement, consider documenting your work in a research log, including:

(a) name of project & date
(b) name & paths of do-file/log-file
(c) names & paths of original & created data sets
(d) tasks contained in do-files
(e) short documentation of any new variables created.

RULE #14: Have a back-up system in place, preferably one with immediate, midrange & long-term backup solutions. Talk to Shawna or one of the TAs for ideas as to what this might look like.

Due to the quick pace of this course, late assignments are not accepted. No exceptions.