Data Management for Analysis
EDLF 5500, Fall 2010
Time: Thursday 1-3:45
Location: Ruffner Hall 175

Professor: Daphna Bassok
Email: dbassok@virginia.edu
Phone: 982-5415
Office Hours: T 2:00-4:00 or
by Appointment (Ruffner 260)

Professor: Jim Wyckoff
Email: wyckoff@virginia.edu
Phone: 924-0842
Office Hours: M 3:00-5:00 or
by Appointment (Ruffner 258)

Overview: Managing data is an important skill for successful research. This class has three primary goals: to develop skills to clean, organize and manage data for data analysis, to develop skills in the descriptive analysis of data, and to develop good programming habits. These skills will be developed in a very applied setting through the use of a variety of databases and Stata software. By the completion of the class students will feel comfortable with: program organization and documentation, file management, data cleaning and checking, working with data from different sources, reshaping data (wide to long), recoding data, transforming variables, preparing macros to reduce data errors, creating clear and polished results tables, handling missing values, merging data, and creating graphs, plots and charts.

Motivation: Research, like many professions, relies heavily on the reputation and credibility of the researcher, think of surgeons or lawyers. Empirical researchers with strong reputations are those who are productive and whose work is reliable. Strong data management skills provide the foundation for improving both. While everyone makes mistakes, having systems that minimize those mistakes is critical. Gwande provides strong evidence that hospitals can substantially improve ER outcomes by employing simple checklists.[1] The premise of this class is that adherence to a set of data preparation and management protocols can substantially improve the quality and reliability of research. Developing good data management habits early on will also make you a more productive researcher. Stata is a powerful general programming language that greatly facilitates these efforts and goes well beyond to include a variety of data analysis funcitons. While other software may have important specialized purposes, Stata provides a data management and analysis backbone that we strongly recommend.

Course Pre-requisites: There are no formal pre-requisites for this course. Some experience with (and interest in) large-scale datasets is a plus.

While Stata is a powerful statistical software package, the focus of this course will not be on Stata's analytic capabilities and no formal knowledge of statistics is expected or needed.

[1] A. Gawande, "The Checklist" The New Yorker, December 10, 2007, accessed at: http://www.newyorker.com/reporting/2007/12/10/071210fa_fact_gawande

That said, students who plan to use Stata to conduct complex statistical analyses are encouraged to enroll.

Course Requirements: Learning how to work with data takes a lot of hands-on practice. Towards this end, this course will be structured as a very interactive workshop. There are three primary requirements:

Weekly homework assignments (50 percent): There will be 11 required homework assignments throughout the semester. Each assignment will involve significant data work in Stata. The assignments will build on the material presented in class, but will require students to apply the basic concepts presented in new ways. Note that for most of us, working with data always seems to take longer than we anticipate. Please start early to avoid problems. Late assignments will not be accepted. When calculating final grades- we will drop your lowest homework score. That said- it is not acceptable or suggested to skip an assignment. The material in the homework builds from week to week, and skipping one week, will make the following weeks substantially more difficult.

Homework assignments are typically comprised of three components—a) a word document that clearly and succinctly answers the homework questions, b) the Stata do.file for the assignment. This do.file needs to be clear, with each homework question clearly delineated, and c) the Stata log file of the do.file in (b).

Electronic versions of all three components of the homework assignments should be put in the Collab dropbox by the **beginning of class on the due date**. Please label your files as: your last name_HWx.zzz e.g., Smith_HW1.log. Note: If you are unable to answer any of the homework questions, describe how you tried to answer the question, and where you were stuck.

- ***Class participation (25 percent)****:* This is a course where "class participation" really counts. We have designed the course in a way that we hope will facilitate many opportunities to learn from your peers.

    - *Homework Presentations:* We will begin each class with a discussion of the homework assignment. Each week we will randomly select one student to lead this discussion. The discussion leader will walk the class through their approach to the assignment, highlight any "discoveries" they've made about new commands or strategies, discuss any challenges, attempt to address questions from the rest of the class, etc.

    - *In Class Explorations:* Most weeks, class will involve hands on practice using data. Either individually, in small groups, or as a class, we will apply data management techniques to large, education policy datasets. Students are expected to be actively engaged in these in-class assignments and discussions.

- *Posting online:* Students are encouraged to post questions and comments on our class COLLAB site.

- ***Final project "The State of Education in the States" (25 percent):*** There is one final project for this class **due December 2nd**. The project is meant to serve as a way to apply everything that is learned in this course to real, messy, large-scale education data. The weekly homework assignments in the second half of the course will all be related to the final project. As part of the final project you will be expected to (1) resubmit all "state education" assignments revised as needed based on weekly homework feedback; (2) present your work during the last class session (December 2); (3) write a brief final paper summarizing your findings. More information on the final project will be provided in class.

A note on how work will be evaluated: In this class getting the "right" answer will not always be enough for getting full credit on assignments. Your grade on assignments will reflect the extent to which your work incorporates the techniques and approaches presented in the class. There is a heavy premium on clarity. A program that "gets the job done" but is very difficult for us to follow, does not make use of comments, and is generally sloppy will be marked down. Similarly, if in class we specifically discuss an elegant, more reliable (or FASTER) way to code something otherwise onerous, we expect you to make use of this strategy. That said, there will probably be many times throughout the semester you discover an even more elegant, more reliable, better or just different approach to doing something in Stata. This is great. There are many ways to do the same thing in Stata, and we look forward to learning from you as you discover new methods.

What to do when you get stuck: While working with large-scale datasets you will, inevitably, get frustratingly stuck. Sometimes tasks that seem like they should be totally SIMPLE seem strangely impossible to code properly. One goal of this course is to learn how to resolve these challenges quickly so you can get on with your work. Here are seven suggestions for what to do in these situations.

1. Look over your notes, class PowerPoint's, and class examples to see if we did something fairly similar in class that might prove useful.
2. Search the Stata help files and documentation (more on this soon).
3. Ask one of your classmates if they have any suggestions. Your peers will most definitely be your best resources.
4. Post a question on the COLLAB website. [Everyone should look at the posts regularly to see if they might be able to help a classmate out]
5. See if some other Stata user has run into the same issue (http://www.stata.com/statalist/archive/)
6. Ask Google!
7. Email the instructors. Please note: this option is listed seventh of seven options. While we are happy to help, sending us an email with specific programming questions should be a last resort after making a good faith effort to resolve the matter using suggestions 1-6. Emails to us on such issues should include the

statement "I've already tried getting unstuck using approaches 1-6." For fastest responses, please cc both instructors on your emails.

Accessing Stata: The latest version of Stata is available by "virual access" to all UVa students through the Hive (http://itc.virginia.edu/hive/). We will spend some time familiarizing ourselves with the Hive during the first day of class.

Collab Site: Our course has a "COLLAB" site which you should check frequently (https://collab.itc.virginia.edu/portal). All lecture slides, homework assignments, and datasets will be posted on this website. Also, students are expected to use the COLLAB site to post their questions, answers to questions, and other enlightening discoveries.

Readings: There is one required textbook for this course, and you may purchase it directly from the instructors for 50 dollars. It is:

Mitchell, M. "Data Management Using Stata: A Practical Handbook." Stata Press, 2010.

Other useful references:

A new and very handy book about research "workflow" and data management:
Long, J.S. "The Workflow of Data Analysis Using Stata."Stata Press, 2009.

An excellent website out of UCLA with many helpful examples:
http://www.ats.ucla.edu/stat/stata/

A searchable archive of questions about Stata: http://www.stata.com/statalist/archive/

The International Archive of Education Data: An excellent source for education data:
http://www.icpsr.umich.edu/IAED/

A note on academic honesty: We assume and expect that all students will approach the work they do both in this class and outside of it, with academic honesty. It is the student's responsibility to become familiar with and adhere to the guidelines outlined in the University of Virginia Honor Code. See also:
http://www.virginia.edu/honor/proc/fraud.html

Given the collaborative nature of the learning process employed in this class, academic honesty dictates that you make substantial efforts ensure you are actually discovering/developing good data management strategies to solve assignments rather than copying the work of others. While we encourage you to speak to and collaborate with your classmates when working on assignments, each student must submit their own programs, files, etc. If you have worked closely with another student(s) on your assignment, please note them as a collaborator on your homework write-up.

Tentative Class Schedule (Subject to change)

Date	Topic	Reading (to be done before this class session)	Assignment (due at the beginning of class
Aug. 26	Introduction to Data Management		
Sept. 2	Working with Data in Stata	Mitchell: Ch. 1, 2	HW 1
Sept. 8	Data Cleaning & Labeling	Mitchell: Ch. 3, 4	HW 2
Sept. 16	Creating Variables	Mitchell: Ch. 5	HW 3
Sept. 23	Writing Effective do.files	Mitchell: Ch.9, 9.1-9.5	HW 4
Sept. 30	Project Planning, Organization & Efficiency		HW 5
Oct. 7	Intermediate Data Mgmt: Combining Datasets	Mitchell: Ch. 6	HW6
Oct 14	Intermediate Data Mgmt: Restructuring Datasets	Mitchell: Ch. 8	HW 7
Oct. 21	Repeating Commands	Mitchell: Ch. 7	HW 8
Oct. 28	Using Macros and Other Short Cuts	Mitchell: Ch. 9 9.6-9.13	HW 9
Nov. 4	No Class (APPAM Conference) Individual meetings about final project.		
Nov. 11	Presenting Data: Graphics in Stata	Kohler & Kreuter, "Creating and changing graphs" (Handout)	HW 10
Nov. 18	Presenting Data: Making Tables		HW 11
Nov. 25	No Class (Thanksgiving Holiday)		
Dec. 2	Class Presentations	Final Project Due	

Tentative Class Learning Goals: We have developed learning goals that indicate the skills you should be able to perform by the end of that that week. These closely correlate with the reading and homework assignments, and should guide your out of class efforts and our in class discussions. If you find that by the end of class you are uncomfortable with your mastery of the learning goals for that week, please talk with one of us.

Week 1: Introduction to Data Management

a) Understand the structure of the class, the nature of class assignments, expectations for your performance and how you will be evaluated.
b) Be able to describe the basic principles of data management
c) Be able to access Stata (in and out of class)
d) Understand the organization of Stata Windows
e) Be able to open a database using Stata
f) Be able to use basic Stata syntax and commands

Week 2: Working with Data in Stata

a) Be able to describe different types of types of databases and understand the implications for working in Stata
b) Be able to move across different data directories to store files
c) Be able to enter and save data in Stata
d) Be able to import existing data organized in a variety of varying formats into STATA
e) Be able to set memory appropriately to most efficiently analyze data in Stata
f) Be able to construct and run a simple do.file

Week 3: Data Cleaning and Labeling

a) Be able to employ common strategies to verify that data only take on allowable values and are internally consistency
b) Be able to modify and flag data when problems arise
c) Be able to construct a master data base when there are multiple sources for the same variables
d) Be able to create variable labels, value labels and database labels

Week 4: Creating Variables

a) Be able to modify existing variable values and generate new variables
b) Be able to limit the observations to a subset of original data
c) Understand the difference between string and numeric formats and how to change from one to another
d) Be able to identify, code and use missing variables and observations
e) Be able to read and modify time and data variables

Week 5: Writing Effective do.files

a) Be able to clearly identify the goals of the program
b) Be able to build a template do.file that includes basics for most programs

c) Understand and practice program documentation
d) Be able to debug do.files when errors occur
e) Be able to construct a master do.file that calls other do.files

Week 6: Project Planning, Organization and Efficiency

a) Be able to describe the components of the research project, develop a directory structure that fits the needs of the project and choose a program naming convention
b) Conceptualize and organize the structure of multiple datasets for projects with multiple goals and sources of data.
c) Structure the work flow into a series of programs with specific purposes
d) Be able to clearly identify the goals of the program and build the program to meet these goals
e) Include program documentation to describe each phase of the program, including program modifications that respond to intermediate results

Week 7: Intermediate Data Management: Combining Datasets

a) Be able to correctly perform various merges: 1 to 1, 1 to many, many to 1 and many to many
b) Be able to combine datasets without merging

Week 8: Intermediate Data Management: Restructuring Datasets

a) Be able to convert panel datasets from long to wide and the reverse
b) Be able to aggregate data
c) Be able to return to the data in its original format

Week 9: Repeating Commands

a) Be able to repeat commands across different groups
b) Be able to assign and employ variable subscripts to repeat commands and create subscript specific variables

Week 10: Using Macros and other Short Cuts

a) Be able to construct and use macros for variable and option lists
b) Be able to create loops that perform a series of commands
c) Be able to recover values from Stata commands

Week 11: Presenting Data: Graphics in Stata

a) Be able to create publication ready graphs illustrating various types of analysis

Week 12: Presenting Data: Making Tables

a) Be able to create publication ready one-way and n-way tables