EDLF 5310: Data Management for Social Science Research 3 Credit Hours
Leadership, Foundation, and Policy Fall, 2019

08/27/2019 - 12/06/2019
Mondays 6 pm – 8:30 pm
Mechanical Engineering 215
Final project due: Monday, December 9th

Instructor
Luke Miller, MPP, Ph.D., Research Assistant Professor
Office: Ruffner, 260
Office Hours: Tuesdays, 10:00–11:30 am
Email: lmiller@virginia.edu

Teaching Assistant
Brian Heseung Kim, MPP, current Ph.D. Student in Education Policy
Office Hours: Thursdays, 12:00–1:45pm
Office Hours Location: Ruffner, 279 (center hallway benches)
Email: bhk5fs@virginia.edu

Description

This course introduces strategies for effectively working with large-scale quantitative data for social science research. Topics covered include: data cleaning, recoding, and checking; merging data from multiple sources; reshaping data; documenting processes; writing programs and macros to reduce errors; and presenting descriptive data through tables and graphs. Students will utilize Stata, a statistical software package.

Learning Objectives

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 and replicable. Strong data management skills provide the foundation for these traits. While everyone makes mistakes, having systems that minimize those mistakes is critical. 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 statistical software package that greatly facilitates these efforts and goes well beyond to include a variety of data analysis functions. While other software may have important specialized purposes, Stata provides a data management and analysis backbone that I strongly recommend.

By the end of the course students should:

- Be able to practice the basic principles for robust, responsible, and collaborative data management.
- Be able to transform raw data into usable, analysis-ready datasets.
- Be able to write efficient and well-documented programs to facilitate reproducible analysis.
- Be able to describe data using tables, graphs, and other statistical techniques.
- Be able to find help and solutions when faced with a data management challenge.

Instructional Methods

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. Every week, class will involve practice using data. Either individually, in small groups, or as a class, we will apply data management techniques to large datasets. Six homework assignments and a final project are the "heart" of this class. Doing these assignments, with lots of support in class, is where you will really develop the skills necessary to independently manage large, quantitative data. This class is regularly characterized as "high effort/high reward."

Our course has a "Canvas" site which you should check frequently. All lecture slides, homework assignments, and datasets will be posted on this website. Also, students are expected to use the Canvas site to post their questions, answers to questions, and other enlightening discoveries.

Course Software and Texts

Textbooks There are no required textbooks for this course. Two helpful optional books are:

- *Mitchell, M. (2010) "Data Management Using Stata: A Practical Handbook." Stata Press,*
- *Long, J.S. (2009) "The Workflow of Data Analysis Using Stata." Stata Press, 2009.*

Stata 16: All students in the course are required to purchase a copy of Stata 16 (Intercooled-IC). It is available at a discounted student rate at this link.

A six-month license is $48 and a "perpetual license" (one that continues indefinitely) is $225. Either is fine for the class, so purchase based on what you believe your future needs will be. Please order Stata through the link above. There is a download delivery option; if you order by 4 pm EST you can receive an activation key within the same day. Note that if you end up wanting to upgrade your license to a "bigger" version of Stata than the intercooled version (bigger versions can handle larger datasets with more variables and observations), you are always able to upgrade to the bigger package. We will be using the software extensively, both during class and for all required homework assignments and projects.

Course Outline

Class # and Date	Topic	Assignments DUE by 5pm
1. September 2	Introduction to Data Management	
2. September 9	Files! (Data files, do files, log files)	Intro discussion post on Canvas
3. September 16	Data cleaning & labeling	HW #1
4. September 23	Creating Variables	HW #2
5. September 30	Repeating Commands	HW #3
October 7	***No class. Reading Day***	
6. October 14	Combining Datasets	**Final Project Proposal**
7. October 21	Restructuring Datasets	HW #4
8. October 28	Restructuring Wrap-Up & Example	HW #5
9. November 4	Using Macros and Other Short Cuts	HW #6
10. November 11	***Applied Data Management Task***	
11. November 18	Graphics in Stata	
12. November 25	Simple Descriptive Tables & Analysis	**Final Project Progress Report**
13. December 2	Data Management Highlights/Celebration	**Final Project Presentation**
DECEMBER 9th	***No Class ~ Examinations.***	**Final Project DUE by 5 pm**

Grading and Assessments

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 four primary requirements:

- ***Homework assignments (30 percent)***: There will be 6 required homework assignments throughout the semester, each worth 5 percent of your final grade. 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.**

    - Homework assignments should be submitted as "log" files. We will learn how to create these files in Week 2. Your log file should show (1) the code you used to answer the homework problems; (2) code comments that clearly explain what you're doing at each major step; (3) the output as it appeared in your results windows; (4) your answers (written clearly in code comments using sentences, tables, or otherwise). Your log file needs to be clearly formatted, with each homework question delineated. We will show examples of how you can do this in class.

    - Electronic versions of the homework assignments should be uploaded in the assignments section of Canvas by **5pm on the due date**. **Please name your files as: your last name_HWx.log e.g., Kim_HW1.log.** You will find the assignment in that week's canvas module.

- Note: If you are unable to answer a particular homework question, instead describe how you tried to answer the question, and where you were stuck. This will help us provide more structured feedback.

- ***Class participation (5 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.

  - *Posting online:* Throughout the semester, students are required to participate in the discussion boards at least two times, and are encouraged to do so much more often. Please post questions and comments and respond to others' questions on our class Canvas site. Please also post cool "aha" moments (e.g. "I can't believe I got this awesome loop to run! Check it out" or "I spent all of last week trying to figure out a way to do XYZ with my dataset and finally figured out a way to make it happen!" The discussions can be a tremendously helpful resource for getting "unstuck", but only works if all students visit and contribute regularly. Asking and answering questions in a shared space provides a public service; this style of online community is a *critical* element of coding for all levels of research and projects (statalist, stackoverflow, twitter, etc.). Levels of participation will be factored into your final grade.

  - At times, we will ask people to present on solutions to their homework assignments, class assignments, and work-in-progress from their final projects.

  - *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 datasets. Students are expected to be actively engaged in these in-class assignments and discussions.

- ***In-Class Applied Data Management Exercise (20 percent)*** On **November 11** we will spend the class period doing an applied data management task. Using real-world data students will be asked to complete an in-class assignment that provides them with a hands-on opportunity to practice all the skills and techniques covered in the course.

- ***Final project (45 percent):*** There is one final project for this class **due December 9th**. 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, and to do so independently. The final project will also provide you with an opportunity to work with data that's particularly interesting to you or relevant to your own research (if it meets the requirements of the project). As part of the final project you may be asked to present your work during the last class sessions. 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 and care. A program that "gets the job done" but is very difficult to follow, does not make use of comments, and is generally sloppy or haphazard will be marked down. Statistical analyses and coding are increasingly done in collaborative environments, and knowing how to write readable, well-documented code is an essential skill in research and industry.

Similarly, if in class we specifically discuss an elegant, more reliable (or FASTER) way to code something otherwise onerous, I 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 I look forward to learning from you as you discover new methods.

Resources

What to do when you get stuck/wish to learn more

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 PowerPoints, and class examples to see if we did something fairly similar in class that might prove useful. We intend to at least introduce you to all concepts you need to know for homework assignments.
2. Search the Stata help files and documentation (more on this soon).
3. Ask Google! Use thoughtful keywords that describe what you're trying to do, and what error you're getting (plus "stata" to filter out results for other statistical languages. We recommend prioritizing results from Statalist and StackOverflow)
4. Ask one of your classmates if they have any suggestions. Your peers will most definitely be your best resources.
5. Post a question on the Canvas website. Try to be as specific as possible with what you're trying and what errors you're getting so others can actually understand what you're struggling with [Everyone should look at the posts regularly to see if they might be able to help a classmate out]
6. See if you can attend either my or Brian's office hours (listed at the top of this document). We look forward to seeing students and learning how we can best support you! To make the best use of time, come prepared to explain what you've already tried and what your best understanding of the issue is. Bring a laptop if possible.
7. Email the instructors. Please note: this option is listed **seventh** of **seven** options. While Brian and I are eager 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."** Both I and Brian do our best *not* to access emails on the weekend (and we encourage our students to find similar ways of supporting their own mental well-being throughout school and work), so this is yet another reason why **starting early** is crucial.

There are also a number of terrific resources both within and outside UVA that may be useful.

UVA Resource

- University of Virginia StatLab, provides workshops and statistical consulting (SPSS, SAS, R, etc.): http://statlab.library.virginia.edu/

External Resources

- An excellent website out of UCLA with many helpful examples: https://stats.idre.ucla.edu/stata/
- Links to many Stata supports including video tutorials: https://www.stata.com/support/
- A listserv for Stata questions: https://www.statalist.org/forums/
- Very helpful cheat sheets: http://geocenter.github.io/StataTraining/

University Email Policy

Students are expected to activate and then check their official U.Va. email addresses on a frequent and consistent basis to remain informed of University communications, as certain communications may be time sensitive. Students who fail to check their email on a regular basis are responsible for any resulting consequences.

University of Virginia Honor System

I 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.

Given the collaborative nature of the learning process employed in this class, academic honesty dictates that you make substantial efforts to ensure you are actually discovering/developing good data management strategies to solve assignments rather than copying the work of others. While I strongly 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.

Special Needs

It is the policy of the University of Virginia to accommodate students with disabilities in accordance with federal and state laws. Any student with a disability who needs accommodation (e.g., in arrangements for seating, extended time for examinations, or note-taking, etc.), should contact the Student Disability Access Center (SDAC) and provide them with appropriate medical or psychological documentation of their condition. Once accommodations are approved, it is the student's responsibility to follow up with the instructor about logistics and implementation of accommodations.

If students have difficulty accessing any part of the course materials or activities for this class, they should contact the instructor immediately. Accommodations for test taking should be arranged at least 14 business days in advance of the date of the test(s). Students with disabilities are encouraged to contact the SDAC*:* 434-243-5180/Voice, 434-465-6579/Video Phone, 434-243-5188/Fax. For more information, visit their website.

Classroom Civility Statement

Students are asked to refrain from conducting private conversations (both in-person and electronically) in class, and are requested to use appropriate language and behavior that are not demeaning or disruptive to either the instructor or the other members of the class.

On Well-Being

If you are feeling overwhelmed, stressed, isolated, or otherwise unwell, there are many individuals at UVa who are ready and wanting to support you. The Student Health Center offers Counseling and Psychological Services (CAPS) for its students, and they are an incredible resource to be aware of as a student here. They offer free consultations and a number of free sessions as appropriate. Mental and emotional well-being is a wide spectrum, and any reason is a good reason to seek support if you feel you need it (grad school is *hard*). Call 434-243-5150 (or 434-972-7004 for after hours and weekend crisis assistance) or visit their website to get started and schedule an appointment. If you prefer to speak anonymously and confidentially over the phone, call Madison House's HELP Line at any hour of any day: 434-295-8255.

If you or someone you know is struggling with gender, sexual, or domestic violence, there are many specialized community and University of Virginia resources available. The Office of the Dean of Students, Sexual Assault Resource Agency (SARA), Shelter for Help in Emergency (SHE), and UVA Women's Center are all fantastic and eager to help.

Rubric for Homework Assignments Grades

- There are six homework assignments that will together account for 30% of your grade.
- We will post answers to homework assignments on Canvas after they are submitted.
- Note that for most of us, working with data always seems to take longer than we anticipate, and difficulties can pop up in unexpected ways. Please start early to avoid problems. Late assignments will not be accepted.
- Given the collaborative nature of the learning process employed in this class, academic honesty dictates that you make substantial efforts to ensure you are actually discovering/developing good data management strategies to solve assignments rather than copying the work of others. While I 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.
- Homework will be graded on a 7-point scale, where:
  - o *A score of 7* means that the homework is perfect (or nearly perfect). You have mastered the material, and perhaps even taught us some new tricks! Bravo!!!
  - o *A score of 6* means that you have done well on the homework. You could review a few *minor points* (which we will point out to you in the comments or address in class).
  - o *A score of 4-5 means* that you have done an acceptable job on the homework. However, there are a number of mistakes throughout the homework, or you have missed some key concepts.
  - o *A score of 2-3* means that you there are *1 or 2 major concepts* that you should review. You should see me at office hours so that we can address any questions that you might have.
  - o *A score of 1* means that your homework was largely incomplete or that there are several major concepts or skills you have missed. You should come meet with us during office hours to remedy this and to develop a plan to help you get back on track.