**SOC 561: Programming for the Social Sciences**

Spring 2016

S SCI 415

M W 3:30-4:45

Professor Jennifer Earl

Office: Social Sciences 440

Phone: (520) 621-3296

Office Hours: See <http://tinyurl.com/profearl-hours> for hours for any given week—this gets updated all the time, so please check it the week and day you plan to come in

**COURSE DESCRIPTION**

Although standard graduate statistics courses prepare students to design and run statistical analyses, courses generally do not spend a great deal of time discussing data management and workflows, which are critical to making research replicable, efficient, and accurate. This is unfortunate because the best designed statistical analysis is easily undone by poor data management, whether through misconstructed variables, unreplicable workflows, and/or poorly commented or documented workflows and programs.

This course addresses this oversight largely using Stata. Although most students are aware that Stata is a very powerful statistical package, it is also an excellent environment for managing and preparing data before analyses. This course is designed to develop the data management and programming skills of researchers who are new to Stata as well as individuals who already have moderate experience with Stata. *The course is not designed to develop statistical testing skills in Stata*, which are already covered in basic and advanced statistics courses. If you are already an advanced Stata user, you may complete the course but you will need to work out alternative assignments and grading criteria that involve you teaching the class how to perform a similar function in R. Students interested in applying these skills in other programs (e.g., R) will find that although the specific commands and formatting differs between packages, the logics are quite similar and thus skills are surprising portable between programs—it is like learning regional dialectics of the same language.

There are several common goals for all students in the class:

(1) To develop problem-solving skills that allow you to seek and find relevant information on solving Stata programming problems on your own and to allow you to continue developing your Stata skills after this course is over.

(2) To develop sufficient programming skills to allow you to program the same conceptual goal in multiple ways and to evaluate the most elegant path towards your goal.

(3) To develop strong skills in commenting, documenting, and code testing so that you can debug a program for both errors that Stata will notice and substantive errors that Stata won’t flag but that affect your analyses.

(4) Integrate you into a larger community of Stata users, which involves both helping others and asking for help from others.

Since the course addresses several types of students, additional course goals depend on where you enter the course:

*Relative novices*: If you have not worked with Stata before but have worked with other statistical software, or, if you have only worked with Stata but doing the kinds of tasks one completes in a statistics class, for the purposes of this class you are a relative novice. The primary learning goals for you are:

(1) Being able to write, document, and implement a workflow data management program that include basic to intermediate data management tasks, including recoding variables, merging and appending datasets, reshaping data, and creating composite scores through a variety of techniques.

(2) Being able to write do-file segments that use loops and programs to automate repetitive tasks within Stata.

*Intermediate Users:* If you have worked with Stata before and are familiar and comfortable with loading data, merging, appending, and recoding data, then you are an intermediate user for the purposes of this course. The primary learning goals for you are:

(1) To make your data management fully replicable so that only original data entry happens outside of Stata, but all other data management is internal to Stata, and your data management is appropriately documented and commented.

(2) Being able to write *elegant* do-file segments that use loops and programs, both inside Stata and in your operating system.

(3) To be able to automate repetitive tasks, including where that requires interfacing with other computer programs that complement your work in Stata

To reach these goals, the course is divided into several segments, which are tailored to the different and developing skill sets of these different types of students. Alternative assignments, such as substantial programming on an ongoing research project are possible, but must be discussed and agreed on with Professor Earl within the first two calendar weeks of class.

The course segments are as follows: (1) a general introduction, focused largely on goals of workflow management and programming; (2) basic data manipulation; (3) intermediate data manipulation; (4) loops and programs; and (5) interfacing with other common programs; and (6) student-driven projects and questions, which will involve participation credited toward course participation scores.

**REQUIRED READING**

Most readings are from freely available online materials, with links to the material in the syllabus. However, students do need to have access to the entirety of the following book for the entire term:

Long, J. Scott. 2009. The Workflow of Data Analysis Using Stata. College Station, TX: Stata Press. Available at: <http://www.amazon.com/Workflow-Data-Analysis-Using-Stata/dp/1597180475>

I also highly recommend *Stata Tips* (<http://www.stata.com/bookstore/stata-tips/>), which will be used frequently. However, I will make a single copy of this book available in the departmental library for temporary use by students who do not wish to purchase a copy.

**COURSE REQUIREMENTS AND EVALUATION**

Grades will be based on the following:

Assignment Grade per Assignment Total

Course Participation 10%

Homework assignments 5% each; 8 total 40%

Quizzes 7.5% each; 2 total 15%

Class presentation on add-ons 10% of overall grade 10%

Community Resource Building 25%

100%

Here are the details:

Regular attendance and participation in class discussions: Students are expected to have read and reviewed course material prior to class and to demonstrate that preparation through their in-class participation and interaction. These are not “gimme” points—just as in your later intellectual life, you are expected to participate in a community to be successful and that participation needs to reflect your pre-class preparation in terms of reading and playing in Stata.

Homework: There are several homework assignments (with respective dates indicated on the class schedule below).

Quizzes: The quizzes will be in Unit 2 and Unit 4.

Class presentation on add-ons: Each student will do presentations on assigned add-ons twice during the semester. Although the number of add-ons varies by day, the distribution is designed to account for differences in difficulty (so days with more add-ons tend to have easier add-ons bundled together). We will sign up for these in the first course. All presentations should address these topics: (1) What is the ado file and what does it do?; (2) How would someone find it? (3) What is the basic syntax? (4) What are major options? and (5) Run an example of the ado file. Grading will be based on detail and accessibility. It is recommended you make files available for your class prior to class, which will require you to send me your files before 7pm on the night before class.

Community Resource Building: For every class on Stata (so, any class before the last unit), at least one student must create an online resource to teach others about the material covered in that course day. This may include creating a video that is posted via a course account on YouTube, a post to a course-based blog, the creation of a detailed Wikipedia entry that may be linked to from the course blog, or other similar training tools. The goal is to make sure you really understand the material enough that you could explain it to someone else and at the same time build up the online knowledge base for current and future users. Since a prior class already made these resources and they are available online, please review the related resource for your assignment and be sure to in some way extend or improve the existing resource (i.e., don’t just give me a slightly revised version of what a student already posted).

I expect these to include examples and they will be graded based on their comprehensiveness and accessibility. We will agree on a course calendar where students choose course dates. The materials will be due a week after the class covering that material. You must also submit a PDF (for text based resources) or a copy of the video or audio file for other kinds of presentations so that if the online link ever goes down I can restore the material with this backup.

**COURSE POLICIES**

1) The course will use D2L to track and submit some of your assignments as well as host class information. Students should be automatically enrolled in the course and are responsible for all material on this site.

2) **Late assignments will receive no credit**.

3) Students requiring any disability-related accommodations should speak with me in a timely manner to secure those accommodations.

4) Students are expected to abide by the student code of conduct and all University regulations.

**CLASS SCHEDULE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week/**  **Day** | **Student Presentation** | **Lecture Topic, Readings** | **Assignments** |
| **Week 1** |  |  |  |
| W, 1/13 |  | Introduction to the Course  If you are a true novice, I recommend: Module 4, 5 & 6 from the Intro to Stata CDs and that you try to do these in the next week before class reconvenes on Wed 1/20  For mac users:  Stata tip 26: Maximizing compatibility between Macintosh and Windows  Stata tip 18: Making keys functional |  |
| **Week 2** |  |  |  |
| M, 1/18 |  | University Holiday for MLK Day—No Class |  |
| W, 1/20 | *Student Presentation Topic:* Working Across Different Versions of Stata  use13  saveold  version | **UNIT 1: GENERAL INTRODUCTION AND CONVENTIONS**  Replicability, Efficiency, and Elegance as Programming Goals  Naming Conventions for Files  Writing and Reading Documentation  All:  Help menu and items: http://youtu.be/UpXNMeTzmuI  PDF Documentation: http://youtu.be/KPHxC-HyrMk  Long, Ch 1: "Introduction"  Long, Ch 2: "Planning, organizing and documenting" | Homework 1 explained |
| **Week 3** |  |  |  |
| M, 1/25 | *Student Presentation Topic:* File Utilities  cleanlog  filefilter | An introduction to do files  Commenting  Debugging errors Stata finds  Finding errors Stata doesn’t notice  All:  Long, Ch 3: “Writing and Debugging Do Files”  Long, Ch 5.2-5.4 from “Names, Notes and Labels”  If you are still confused:  http://www.ssc.wisc.edu/sscc/pubs/sfr-dofiles.htm | Homework 1 due  Homework 2 explained |
| W, 1/27 | *Student Presentation Topic:* Ado Files  Demonstrate Installing and Updating Ado Files (<http://www.ssc.wisc.edu/sscc/pubs/4-16.htm>)  adoedit | Ado Files and the SSC Archives + Getting Help  All:  http://www.ssc.wisc.edu/sscc/pubs/sfr-dofiles.htm  Novices (Intermediates may skim):  General overview for Stata beginners and relative novices: http://www.wiwi.uni-muenster.de/ioeb/Downloads/Forschen/Pfaff/  Introduction\_to\_Stata\_with\_50+\_Basic\_Commands.pdf  Intermediate (Novices may skim):  Stata tip 30: “May the Source be With You (viewsource)”  Long, Appendix: “How Stata Works” | Homework 2 due  Review someone else’s code and output for Homework 2 |
| **Week 4** |  |  |  |
| M, 2/1 | *Class Topic:* Data storage types, variables names and labels  *Student Presentation Topic:* Labeling variables and changing storage type  labutil  recast | **UNIT 2: STATA BASICS**  **Data Storage types**  **Variable Naming**  **Labels and Notes**  All:  Variable Types and Precision and String heading: http://www.ssc.wisc.edu/sscc/pubs/sfr-data.htm  Long, Ch 5.5-5.10: "Names, notes, and labels"  Drop, Keep, Rename, Labels: http://www.ssc.wisc.edu/sscc/pubs/sfr-data.htm  Describe and Labels: http://www.ats.ucla.edu/stat/stata/modules/labels.htm  Importing Data: http://www.stata.com/support/faqs/data-management/converting-excel-files/ (from a common format like Excel)  Stata Tip 35: Detecting Whether Data Have Changed (Datasignature)  Other ways to get data into Stata: Pages 3-11 from http://fmwww.bc.edu/GStat/docs/StataInter.pdf  http://www.dawnteele.com/dt/technology\_files/The%20Stata%20Bible%202.pdf | Homework 3 assigned |
| W, 2/3 |  | NO CLASS  NOTE: If you are a true novice, and you have not done Module 4, 5 & 6 from the Intro to Stata CDs, you really need to do those and today would be a great day to draw a line in the sand and do them |  |
| **Week 5** |  |  |  |
| M, 2/8 | *Student Presentation Topic:* Output summary statistics  tabout  tabstat | **Importing and summarizing data**  All:  Long, Ch 6.1-6.2 & 6.4-6.5: “Cleaning your data”  Stata tip 66: “ds- a hidden gem” | Homework 3 due |
| W, 2/10 | *Student Presentation Topic:* creating factor variables “on the fly” | **Basic Data Manipulation**    All:  Long, Ch 6.3: “Cleaning your data”  Stata Tip 52: Generating composite categorical variables  Stata Tip 2: Building with floors and ceilings  help egen in Stata—mostly focus on the functions available  http://www.ats.ucla.edu/stat/stata/faq/cut.htm  If the Long was confusing to you, also read:  replace and recode:  http://www.ats.ucla.edu/stat/stata/modules/vars.htm  Replace and indicator variables: http://www.ssc.wisc.edu/sscc/pubs/sfr-data.htm (through Indicator variables)  Using tabulate to recode into dummies: “Answer 2 of 3: Use tabulate” http://www.stata.com/support/faqs/data-management/creating-dummy-variables/ &  http://kb.iu.edu/data/bajq.html  If egen was confusing to you, also read:  http://www.ssc.wisc.edu/sscc/pubs/sfr-data.htm  http://stataproject.blogspot.com/2007/12/step-4-thank-god-for-egen-command.html | Quiz 1 to be completed at the beginning of next week online through D2L |
| **Week 6** |  |  |  |
| M, 2/15 | *Student Presentation Topic:* Exploring Data  two\_way\_histogram\_gen  lgraph | **UNIT 3: INTERMEDIATE DATA MANIPULATION**  **By, System Variables, Return Codes, Egenmore**  All:  By from Stata PDF Documentation (see course webpage)  System variables: http://www.stata.com/help.cgi?\_variables  Using system variables: http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/misc/by\_detail  Return codes and values:  Long 4.2: “Information returned by Stata commands http://www.ats.ucla.edu/stat/stata/faq/returned\_results.htm  Help egenmore (from Stata)  Stata tip 14: Using value labels in expressions | Homework 4 explained |
| W, 2/17 | *Student Presentation Topic:* Exploring Data Cont.  mkcorr  fre | **Regular Expressions**  All:  Regular expressions available in Stata: http://www.stata.com/support/faqs/data-management/regular-expressions/  Regular expression commands in Stata: http://repec.org/wcsug2007/medeiros\_reg\_ex.pdf  Example of using a regular expression: http://www.ats.ucla.edu/stat/stata/faq/regex.htm  Stata Tip 60: Making fast and easy changes to files with filefilter  help filefilter (when in Stata) | Homework 5 explained |
| **Week 7** |  |  |  |
| M, 2/22 | *Student Presentation Topic:* more options for merging  mmerge | **Manipulating Entire Datasets: Append, Merge, Collapse**  All:  Stata Tip 5: Ensuring programs preserve data sort order  help append (in Stata)  Append: http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/example13 (including the Q&A below it)  Stata Tip 73: Append with care!  Long, Ch 6.6: Cleaning your data  help collapse (in Stata)  Collapse: http://www.ats.ucla.edu/stat/stata/modules/collapse.htm  Applying merge to data cleaning: Stata Tip 64: Cleaning up user-entered string variables  If Long was unclear to you:  Merge overview: http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/example8  1-to-1 merging: http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/example9 & http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/example10  One to many merging: http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/example11  Merge and append: http://dss.princeton.edu/training/Merge101.pdf | Homework 4 due |
| W, 2/24 | *Student Presentation Topic:* multiple imputation  mi  miset  mim  ice | **Reshaping Datasets and Transferring and Preserving Data**  All:  Reshape in general: http://www.ats.ucla.edu/stat/stata/notes/reshape.htm  Reshape to long: http://www.ats.ucla.edu/stat/stata/modules/reshapel.htm  Reshape to wide: http://www.ats.ucla.edu/stat/stata/modules/reshapew.htm  Stata Tip 45: Getting those data into shape (reshape)  StatTransfer manual, Pp 50-78 (Jenn will provide a master copy in the library)  Long, Ch 8: "Protecting your files" | Homework 5 due |
| **Week 8** |  |  |  |
| M, 2/29 | *Student Presentation Topic:* summary and regression output  esttab  estout  eststo | **UNIT 4: LOOPS AND PROGRAMS**  Readings:  Long, Ch 4 (pp. 83-105, except 4.3.3): "Automating your work" Macros: "Macros" on http://www.ssc.wisc.edu/sscc/pubs/stata\_prog1.htm -- up to, but not including, “Nested Loops” "B] macros" on http://pierrefrancois.wifeo.com/documents/Intro-Stata---LSE-III.pdf (pp 10-14) Tempfiles: http://www.stata.com/statalist/archive/2004-01/msg00542.html |  |
| W, 3/2 | *Student Presentation Topic:* summary and regression output cont.  estadd  estpost | Topics: forval versus foreach, forval in detail, i & local j, nested loops  Readings:  Long, Ch 4 (4.3.3 & 105-end of chapter): "Automating your work"  Forval: "For Loops & Nested Loops " on http://www.ssc.wisc.edu/sscc/pubs/stata\_prog1.htm  http://www.ats.ucla.edu/stat/stata/modules/acrossvars.htm |  |
| **Week 9** |  |  |  |
| M, 3/7 | *Student Presentation Topic:* regression output cont.  outreg  outreg2 | Topics: loops continued, preventing errors (assert), and debugging loops (trace, pause, and capture)  Readings:  http://www.stata-journal.com/sjpdf.html?articlenum=pr0005  Trace: http://www.stata.com/support/faqs/programming/debugging-program/  Stata Tip 32: Do Not Stop (do nostop)  Stata Tip 41: Monitoring loop iterations  help assert in Stata  help pause in Stata  help capture in Stata |  |
| W, 3/9 | *Student Presentation Topic:* logistic regression output  prgen  prtab | Post  Readings:  [P] postfile  http://datatoday.blogspot.com/2011/11/using-postfile-in-stata.html  http://www.michaelnormanmitchell.com/stow/posting-results-regression.html  Stata tip 54: Post your results (postfile and postclose)  http://www.stata.com/support/faqs/programming/post-command/ | Homework 6 explained |
| **Week 10** |  |  |  |
| M, 3/21 | *Student Presentation Topic:* logistic regression output cont.  margins  marginsplot | Topics: Programs versus do files: what’s the difference? Passing regular arguments  Readings:  http://www.ssc.wisc.edu/sscc/pubs/stata\_prog\_old.htm “Programs” section  http://www.cpc.unc.edu/research/tools/data\_analysis/statatutorial/labor\_saving/program (read entire page, including questions and answers at the bottom of the page)  http://www.stata.com/manuals13/u18.pdf , sections 18.4, 18.4.1 | Homework 6 due |
| W, 3/23 | *Student Presentation Topic:* logistic regression output cont.  spost  clarify | Gettoken and more on arguments  Bringing it all together (closing comments on Stata in isolation)  Readings:  http://www.stata.com/manuals13/u18.pdf , section 18.4.6 | Homework 7 explained |
| **Week 11** |  |  |  |
| M, 3/28 | *Student Presentation Topic:* specialized analysis tools (sequence analysis & geocoding)  sq  spmap & shp2dta  geocode3 & traveltime3 | **UNIT 5: INTERFACING WITH OTHER PROGRAMS**  Windows batch files (which have very similar, but often more powerful analogs in Unix, and hence Macs)  Reading:  http://commandwindows.com/batch.htm  http://www.computerhope.com/batch.htm  http://ss64.com/nt/ (skim in general, but read: dir, cd, copy, del) | Homework 7 due |
| W, 3/30 |  | NO CLASS  Reading (don’t wait until next week; space it out!):  *Python for Informatics*, Chapter 1-3  Find and install Python 2.7.3  <https://developers.google.com/edu/python/>  <https://developers.google.com/edu/python/set-up>  <https://developers.google.com/edu/python/introduction> | Quiz 2 will be available late this week |
| **Week 12** |  |  |  |
| M, 4/4 | python plug-in for Stata | Python, introduction  Reading:  *Python for Informatics*, Chapter 4  <http://docs.python.org/2/library/turtle.html#module-turtle> & read the sample triange.py and rectangle.py programs I send as you go to understand what they do | Homework 1B |
| W, 4/6 | *Student Presentation Topic:* robustness checks/confidence intervals  eclplot  rcentile | Python: strings and lists    Reading:  *Python for Informatics*, Chapter 6 & 8  <https://developers.google.com/edu/python/strings>  <https://developers.google.com/edu/python/lists>  <https://developers.google.com/edu/python/sorting>  <http://www.youtube.com/watch?v=tKTZoB2Vjuk>  <http://www.youtube.com/watch?v=EPYupizJYQI>  <https://www.youtube.com/watch?v=IcteAbMC1Ok> (start to 8 minutes on list comprehension) |  |
| **Week 13** |  |  |  |
| M, 4/11 | *Student Presentation Topic:* robustness checks/confidence intervals cont.  parmest  bootstrap | Python, file access & loops  Reading:  *Python for Informatics*, Chapter 5 & 7  Files segment in <https://developers.google.com/edu/python/dict-files>  <http://www.youtube.com/watch?v=haycL41dAhg> (file access starts at 10:50)  <https://developers.google.com/edu/python/utilities>  https://www.youtube.com/watch?v=uKZ8GBKmeDM |  |
| W, 4/13 |  | Reading:  <https://www.youtube.com/watch?v=Nn2KQmVF5Og> (on file exceptions from 2:30 to 9:15 in the video and on modularity from 9:30 to 16:00) |  |
| **Week 14** |  |  |  |
| M, 4/18 |  | Python, tuples and dictionaries  Reading:  *Python for Informatics*, Chapter 9 & 10  <https://developers.google.com/edu/python/dict-files>  <http://www.youtube.com/watch?v=haycL41dAhg> |  |
| W, 4/20 |  | Python, regular expressions  Reading:  *Python for Informatics*, Chapter 11  <https://developers.google.com/edu/python/regular-expressions>  <http://www.youtube.com/watch?v=kWyoYtvJpe4>  <https://www.youtube.com/watch?v=Nn2KQmVF5Og> (on urllib from 16:00-19:10) |  |
| **Week 15** |  |  |  |
| M, 4/25 |  | **UNIT 6: STUDENT WORK**  TBA | Homework 8 explained |
| W, 4/27 |  | TBA |  |
| **Week 16** |  |  |  |
| M, 5/2 |  | TBA | Homework 8 due |
| W, 5/4 |  | TBA |  |