This course is designed to familiarize you with some common statistical packages. We will focus most of our attention on R, and I will show you how to do almost everything we do in R in Stata as well. R is a free(!), open-source statistical package that is seeing increasingly wide use among the scientific community. It is a powerful tool that is flexible enough to meet almost any need you may have in your research. Stata is a commercially licensed program that some find easier to work with.

My main focus in this course is to cover the basic concepts that you will need so that you can teach yourself the more advanced methods. I will focus on good programming practices that (if followed) will save you a lot of time and heartbreak down the road. Due to the structure of the course (we will meet only weekly), we will need to move fairly quickly through the material. In my opinion, there is no better way to learn a programming language or statistical package than just to get down into the specifics of actual problems, so I will try to provide you with plenty of examples (in the form of homework assignments) as we go along.

Reading

There are no required texts for this course. I have compiled some online resources that you might find useful, but we will not be following a textbook.

Assessment

This course is offered on a pass/fail basis, and you will get out of it precisely what you put into it. There will be homework assignments throughout the semester, but they will not be graded. You may check your own solutions against the solutions that I will post, and I am happy to go over problems on the homework in my office hour.

Your final grade (pass or fail) will be based on a project that we will agree upon (see the next section for more details).

Final Project

The final project for the course will involve applying several of the principles that we discuss throughout the semester in the context of an actual research project. The easiest way to do this (I think) would be to pair it with a paper you are already writing for another course. Your specific circumstances may vary (please come and discuss your plans with me if they deviate significantly from what follows), but a satisfactory project might look something like this:

1) Comprehensive data collection plan
   a. What are the sources of your data?
   b. If you have multiple sources of data, how will they be combined?
   c. What types of transformations will you have to perform to have a dataset that is ready for analysis?

2) Computer code (in R or Stata or perhaps both) that takes the raw data and transforms it into a form that is ready for analysis
3) Comprehensive data analysis plan
4) Computer code (in R or Stata or both) that executes the data analysis plan
5) Computer code that transforms the analysis from 4) into tables and graphs that are ready to insert into a final paper.

Topics and preliminary schedule

Jan 24 - Feb 8 - Programming basics and best practices

* Commenting/Documentation
* Script/do files
* Replication
* Objects
* Functions
* Loops

Feb 15 – Feb 22 - Data management

* Recoding
* Merging
* Appending
* Indexing
* Cleaning

Mar 1 - Data Exploration

* Scatter plots
* Box plots
* Histograms/density plots

Mar 8 – Mar 15 - Visualization

* Axes
* The par() function
* Labeling
* Adding text
* Adding symbols

Mar 22 - Learning how to learn a new method

* The Journal of Statistical Software
* Online resources
* CRAN Documentation

Apr 5 - Basic mapping in R

* The maptools package
* Working with shapefiles
* A simple choropleth map
Apr 12 - Basic Text processing

*regular expressions
*the tm package

Apr 19 - Data collection

*Collecting data from the web

Apr 26 – May 10

Catch-up, review, final project presentations