Public Policy 567
Practicum in Data Analysis with Stata
Fall 2017 Syllabus

Instructor: Jonathan Hanson
jkhanson@umich.edu
4223 Weill Hall, 615-1496
Office Hours: Mon. 11:30–1:00, Thurs. 10:30–12:00, or by appointment

This purpose of this course is to help students become proficient users of Stata for data analysis in their future careers. Although some statistical concepts will be taught when necessary, the focus is to learn to utilize the capabilities of Stata for managing and manipulating data, producing sophisticated analysis, exploring results, creating graphical illustrations, and basic programming. This course assumes that students have completed a graduate-level course in statistics, but no previous experience with Stata is required.

Readings

Reference information for Stata is available from a wide range of sources, including the official Stata manuals, which are excellent and available online in pdf format. Since Stata has an enormous range of tools, no one person or book can serve as a comprehensive resource. Even highly-experienced users regularly search online or consult different resources to figure out how to perform various tasks. Learning to code is a process with trial and error.

For this course, I have developed a *Stata Practicum Handbook* that will serve as the main text. It is available on the Canvas site for the course. You can log into Canvas at `http://canvas.umich.edu` with your usual university credentials.

If you are looking for additional resources, a good, basic reference guide is the following book:

- Alan C. Acock, *A Gentle Introduction to Stata*, 5th edition, (Stata Press: 2016).

Other reading selections will be made available on the Canvas site.

Assignments and Grading

Your grade for this course will be determined by the following:

Problem sets	60%
Final Project	30%
Class Participation	10%

You will learn the material best when you use it. Problem sets will thus be assigned on a regular basis, accounting for 60% of the course grade. Problem sets will be submitted as electronic files to the course website on Canvas.

Since this is a lab course, participation means attending class and engaging in the activities that are part of the lab rather than other computer-related activities such as email or web-browsing. If you must miss class, please notify me in advance if possible so that we may discuss whether the absence can be excused. As a record of your participation in each class, I will ask you to upload the log file from your Stata session.

For the final project, each student will perform a data analysis on a topic of their choice and write up the results. The project is intended for you to utilize the data analysis tools learned in this course, and it should reflect the workflow practices learned in the class. I should be able to run your Stata do file and produce all the analysis that goes into your written report. Details will be provided in an assignment sheet during the early part of the semester. A proposal for this project is due on September 29, and a progress report is expected on November 3.

Some Computing Advice

When we are using Stata on the lab machines, it is efficient to have a working directory that will not disappear when you log out of your account. You can obtain such a directory on the university's Andrew File System (AFS) by making a request. See http://its.umich.edu/computing/backup-storage/afs and click the link to the AFS Self-Provisioning Tool. This is the same file space space that you can access as MFILE through a web browser. By request, you can make this space show up as a "home directory" mapped to M: when you log into a lab machine. You can then rely on a consistent location for your files when writing Stata code.

Another alternative, which I learned about only one day before the first class, is the Kumo Cloud Storage integration system. When you log into a university lab machine, you can setup a linkage to your U-M Box or U-M Google Drive accounts so that they will be mapped as a directory. You then then read or write to these directories just like you would with a local directory.

If you wish to purchase Stata software, consider GradPlan pricing (https://www.stata.com/order/gradplan-sites/). You will need at least Intercooled Stata. For this course, Stata SE may be needed for some datasets. Please note that you can use the university's Virtual Sites (see information at http://www.itcs.umich.edu/sites/labs/virtual.php) to access Stata when not on campus.

Academic Integrity

It is expected that students are familiar with the Ford School's expectations for academic integrity as described at http://fordschool.umich.edu/academics/expectations, which adhere to the academic integrity policies for Rackham Graduate School. Violations of these policies will be taken seriously.

Students with special needs

If you believe you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities to help us determine appropriate accommodations. I will treat any information you provide as private and confidential.

Inclusivity

Members of the Ford School community represent a rich variety of backgrounds and perspectives. We are committed to providing an atmosphere for learning that respects diversity. While working together to build this community we ask all members to:

- share their unique experiences, values and beliefs
- be open to the views of others
- honor the uniqueness of their colleagues
- appreciate the opportunity that we have to learn from each other in this community
- value one another's opinions and communicate in a respectful manner
- keep confidential discussions that the community has of a personal (or professional) nature
- use this opportunity together to discuss ways in which we can create an inclusive environment in Ford classes and across the UM community

Please refer to http://fordschool.umich.edu/academics/expectations for a full statement on the Ford School's academic expectations.

September 8: Introduction and Stata Overview

In this introductory class, we will review course policies, explore the layout of the program, discuss best practices for Stata workflows, and examine configuration issues.

- Associated reading: Acock, Chapter 1.

September 15 & 22: Variable and Dataset Management

These sessions will explore a wide range of data management issues in Stata: how Stata stores and displays data, converting between string and numeric data types, handling of dates, importing and exporting data, creating new variables with `gen` and `egen`, use of variable and value labels, recoding variables, and merging datasets.

- Associated reading: Acock, Chapters 2 and 3.

September 29 & October 6: Descriptives, Cross-Tabs, and Mean Comparisons

In these two sessions, we will explore Stata's tools for descriptive statistics, hypothesis tests involving means and proportions, cross-tabulation (i.e. contingency) tables, and χ^2 tests. Use of Stata's graphing capabilities associated with these tasks will be incorporated throughout. Additionally, we will learn how to access and use Stata's stored results as well as methods for exporting tables.

- Associated reading: Acock, Chapters 5-7.
- Submit proposal for final paper/project on September 29.
- Assignment 1 due on October 6.

October 13: Correlation and Linear Regression

In this session, we will have a thorough coverage of commands related to correlation and regression analysis. For regression, this will include post-estimation analysis with predicted values, illustrated results, and marginal effects. We will also learn tools for specification tests, hypotheses tests for coefficients, and interaction terms with marginal effects. Finally, we will use commands to export regression tables in formats that can be easily used in other applications.

- Associated reading: Acock, Chapters 8 and 10.1–10.4.

October 20: Regression Diagnostics

There is a wide variety of diagnostic tools to test for regression, including methods to identify influential cases, test for heteroskedasticity, and test for multicollinearity.

- Associated reading: Acock, Chapters 10.4–10.15.
- Assignment 2 due.

October 27: ANOVA and Commands for Survey Data

Analysis of Variance (ANOVA) can be used to test for a difference of means across multiple categories. A portion of this session will be devoted to that subject. The remaining portion of the session will be devoted to special commands for survey data. Analysis of survey data often requires special techniques that reflect the sampling frame for the survey. We learn the survey versions of various data analysis methods discussed previously.

November 3: Programming

Stata includes a programming language that one can use to create commands for lengthy or repetitive procedures. Many unofficial Stata commands are created in this way, and they can be installed as packages to Stata to expand the software's capabilities. This session will provide an introduction to programming in Stata.

- Final paper progress report due November 3.

November 10: Simulations

In this session, we make use of Stata's commands for random number generation to perform Monte Carlo simulation of scenarios involving public policy matters. We will also use Stata's functions to produce data that come from various probability distributions.

- Assignment 3 due

November 17: Dichotomous and Categorical Dependent Variables

When the dependent variable is dichotomous, linear regression is problematic in some key ways. We can instead use probit or logit analysis, which employ an S-shaped curve to estimate the probability that the dependent variable is equal to 1. Likewise, dependent variables that have nominal or ordinal categories require different estimation methods. This session will explore Stata's tools for probit, logit and multinomial analysis, including calculation of predicted effects and marginal effects.

- Associated reading: Acock, Chapter 11.

December 1: Time-Series and Panel Data

Time-series data and panel data violate the assumption of that the observations are independently and identically distributed. The values of variables are likely correlated across

observations. Since our prediction errors are not fully random in this scenario, our estimated standard errors are likely wrong. In this section, we learn Stata commands for estimating regressions using these kinds of data.

December 8: Miscellaneous Topics

This week is reserved for a range of topics that we did not have time for in previous weeks, such as displaying data with maps.

- Problem set 4 due.

Final projects are due Monday, December 18.