

Public Policy 567

Practicum in Data Analysis with Stata

Winter 2019 Syllabus

Instructor: Jonathan Hanson

jkhanson@umich.edu

4223 Weill Hall, 615-1496

Office Hours: Mon. 11:00–12:30, Thurs. 12:00–1:30, 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 February 1, and a progress report is expected on March 22.

Some Computing Advice

We are using Stata on lab machines, which is problematic since local files disappear when you log out. The university is trying to retire the Andrew File System (AFS), which had been the best way to use your university file space as a working directory. Unfortunately, this older system does not play well with current Windows networking and with Stata in particular. As the replacement, the university has adopted 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. Unfortunately, this system is slow, causing Stata to pause frequently during operation. In short, you have to choose the alternative that is least bad for you. First, work on a local directory and take the risk that you will lose everything if you forget to upload your files before you log out. Second, work on your Kumo-linked directory, suffering the slow performance but secure in the knowledge that your files will not disappear.

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.

January 18 & 25: Variable and Dataset Management

After an introduction to the course, 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 1, 2 and 3.
- Skim through: V. Orozco et al. 2018. "How to Make a Pie: Reproducible Research for Empirical Economics & Econometrics." Toulouse School of Economics, Working Paper TSE-933.

February 1 & 8: Descriptive Analysis and Basic Statistical Inference

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 February 1.
- Assignment 1 due on February 8.

February 15: 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.

February 22: 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.

March 1 & 8: Winter Break

March 15: 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.

March 22: 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 March 22.

March 29: 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.

April 5: 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.

April 12: Work Time for Final Projects

April 19: 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, parsing text, etc.

- Problem set 4 due.

Final projects are due Tuesday, April 30