Sociology 401/504: Advanced Social Statistics*

Matthew J. Salganik, Simone Zhang, and Ian Lundberg†

Class: TTh 9:30AM-11:50AM (Wallace Hall 165)
Lab: Fr 10:00AM-11:50AM (Wallace Hall 165).

Matthew J. Salganik
mjs3 at princeton dot edu, www.princeton.edu/∼mjs3/
Office hours: Tuesday 4:30-5:30 (Wallace Hall 145)

Simone Zhang, Preceptor
sxz at princeton dot edu
Office hours: M 4:30-6:30 (Wallace Hall 290) and Th 3-5 (Wallace Hall 190)

Ian Lundberg, Preceptor
ilundberg at princeton dot edu
Office hours: M 4:30-6:30 (Wallace Hall 290) and Th 3-5 (Wallace Hall 190)

This is a class designed for first year graduate students in the social sciences. It has a graduate course number (Soc504) and an undergraduate course number (Soc401), but it is the same class. The prerequisite is Soc500/400 or a similar introduction to statistics class which covers basic probability, regression, and causal inference and which covers coding in R. Please speak with the instructor if you have not taken Soc500/400 but are considering taking this course.

1 The Basics

1.1 Overview

This course is the second of the two-semester graduate-level social science statistics sequence. In this sequence, students will learn the statistical and computational principles necessary to perform modern, flexible, and creative analysis of quantitative social data. This course sequence will transform you from consumers of quantitative research to producers of it.

By the end of the semester, you will be able to:

- Conduct, interpret, and communicate results from analysis using generalized linear models.
- Conduct additional study of more advanced topics in quantitative methods

*This course relies heavily on materials developed by Brandon Stewart in previous iterations.
†Last Edited: January 31, 2018

- Build a solid, reproducible research pipeline to go from raw data to final paper.

In terms of statistical content SOC504/401 will cover maximum likelihood estimation, generalized linear models, and assorted topics helpful for data analysis.

Upon finishing the course sequence, students should be able to read an original scholarly article describing a new statistical technique, implement it in computer code, estimate the model with relevant data, understand and interpret the results, and explain the results to someone unfamiliar with statistics. The capstone project for the course sequence is a replication project in which students will replicate and extend a piece of scholarly work in the contemporary literature.

As this is a course sequence, it is natural to assume that the structure of the learning process will be the same. However this isn't always the case and the key differences from Soc500/400 are clearly denoted throughout as "New to Soc504/401."

1.2 Class and Lab

Formal instruction for the course is split into two pieces: class and precept/lab. The course meets two times a week and will cover the core statistical material. The lab meets once per week and will focus on practical computational skills. Both are an essential part of the learning process.

New to Soc504/401:
Class and lab will take on a slightly different role in Soc504/401. Reading before class will be a more important part of the learning process (although don't worry, we will help you learn how to read statistics effectively). Labs will also be somewhat more lecture driven and you will take on a bit more responsibility in teaching yourself the `R` code.

1.3 Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. Statistical methods is like a language and it will take time and dedication to master its vocabulary, its grammar, and its idioms. However like studying languages, statistics yields to daily practice and consistent effort.

The prerequisite is Soc500/400 or a similar introduction to statistics class which covers basic probability, regression, and causal inference and which covers coding in R. Please speak with the instructor if you have not taken Soc500/400 but are considering taking this course.

2 Materials

2.1 Computational Tools

The best way, and often the only way, to learn new statistical procedures is by doing. We will therefore continue to make extensive use of R as well as a number of companion packages. `R` is probably the most widely used statistical software. We recommend using R with RStudio. A basic foundation in using R is assumed.

2.2 Books

Required Note that the first title will be purchased through an online system described below.

- King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Cambridge University Press.[1]
- Fox, John. 2016 *Applied Regression Analysis and Generalized Linear Models. 3rd Edition.*

Suggested It is often helpful to see the same material in alternative ways. Thus here are some other texts you might consult.

- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.
- Blitzstein and Hwang. 2014. *Introduction to Probability.* (Digital version available free through the library).
- Morgan, Stephen L, and Christopher Winship. 2014. Counterfactuals and Causal Inference: Second Edition. Cambridge University Press. (available online for free through the library).
- Wickham, Hadley. *Advanced R* (available free online)

3 Assignments

There are three main types of assignments, each of which is described below.

1. **Preparing for class and lab:** For many classes and some labs there will be some reading that you must do before class. I expect you to come 100% prepared. I will not assign an unreasonable amount of reading and thus I won't spend valuable class time summarizing readings that you should have done before class.
2. **Weekly problem sets:** Learning data analysis takes practice. The problem sets are described below.
3. **Replication and extension project:** A co-authored paper. See below.

3.1 Readings

There are readings for each topic and they mostly cover the theory of the method along with some applications. Obviously, read the required readings, and feel free to read any others that pique your curiosity.

New to Soc504/401:
Reading will take on a more central role in Soc504/401 and preparing before class is essential. Reading statistics can be challenging at first. If you don't understand something, that's perfectly fine; we'll figure it out together and make sure no one is left behind.

[1] Why a political science book? The title here is somewhat unfortunate and a product of its time. The book is quite general to the social sciences.

3.2 Problem Sets

Methods are tools, and it is not very instructive to read a lot about hammers or watch someone else wield a hammer. You need to get your hands on a hammer or two. Thus, in this course, you will have homework on a weekly basis for the first part of the course. The assignments will be a mix of analytic problems, computer simulations, and data analysis.

Assignments should be completed in R Markdown which allows you to show both your answers and the code you used to arrive at them. Don't worry if you don't know R Markdown, we will show you how it works. Your wonderful preceptors will provide you with more detailed instructions before the first assignment is due.

Each week's homework will be made available on Blackboard starting Friday at noon and is due on Blackboard Thursday the following week (6.5 days later) at **11:59 pm**. Please bring a printed copy to precept or put one in the Soc 504 box in the mail room.

Solutions will also be available through Blackboard one hour after the homework is due. The homeworks will then be graded on a 50-point scale and returned to you within two weeks.

Looking at the solutions posted after the homework is an extremely important part of the learning process, so please keep up with the work!

Homework Extension Policy: You have a total of five late days that you can spend over the course of the semester no questions asked. Each late day grants a 24-hour extension. You may elect to use all your late days on one problem set or spread them across multiple problem sets. We will maintain a public spreadsheet that keeps track of the number of late days you have remaining at the start of each new problem set. We will drop your lowest problem set grade if you use zero late days.

After you have used all five late days, you will receive 50% credit for any subsequent late homework assignments you turn in by the last day of class, Thursday May 3rd.

Because we do not want to hold up the class we will not wait for everyone to submit their problem sets in order to post the solutions key. If you are turning your problem set in late you are on your honor to **not look at the solutions key** before submitting your work.

New to Soc504/401:
When submitting a problem set **late** (after solutions have been posted), please indicate the number of late days you are using and type the honor code statement "This problem set represents my own work in accordance with University regulations. - [Your name]" (This comes from Section 2.4.3 of *Rights, Rules, Responsibilities* [link].)

Code Conventions: Throughout the course, students will receive feedback on their code from the professor, the preceptor, and other students. Therefore, consistent code conventions are critical. We will explain how to automatically check your code style using RStudio and a package called lintr. Good coding style is an important way to increase the readability of your code (even by a future you!).

Collaboration Policy: We encourage students to work together on the assignments, but you should write your own solutions (this includes code). That is, no copy-and-paste from other people's

code. You would not copy-and-paste from someone else's paper, and you should treat code the same way. However, we strongly suggest that you make a solo effort at all the problems before consulting others.

Some problem sets will have "no collaboration" (NC) problems. Like in Soc500/400 you can use any resources with the exception of other human beings to help with these problems. The preceptors will however be willing to offer some assistance with the problem in office hours. It is okay if the NC problems aren't perfect- we understand that limiting collaboration will make this harder for some students, but it will also increase learning.

3.3 Help

We know that statistics can be challenging and help is available when you need it. We have made every effort to give you the tools you need to succeed in this course. Ultimately though it is your responsibility to put in the effort and seek out that help.

First, the readings provide ample sources of information and the suggested reading list contains many versions of the same material but presented from a different angle. Lab material and lecture slides will all be posted on Blackboard and can then be referenced.

For questions about the material and problem sets we will be using Piazza. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates, the preceptors, and the professor. Unless the question is of a personal nature or completely specific to you, you should not e-mail teaching staff; instead, you should post your questions on Piazza. The course staff will be monitoring the page, but we encourage you to help your classmates as well. I will post the link to the course page here at the start of class

3.4 Grading

Grades in the course will be assigned according to the following breakdown: 10% participation (including online and reviews), 40% problem sets, 50% final paper.

3.5 Replication and extension project

The main assignment is a research paper that applies some advanced method to, or develops one for, a substantive problem in your field of study. The goal of the paper is to write a publishable article. I know, it sounds hard, but that's only because you haven't learned some of the material we go over in class. There will be no final exam.

There will be a number of interim deadlines which we describe below. We may add others as the semester goes on.

Weekly updates: The final problem of every homework will ask you for updates on your replication project. We look forward to seeing the progress you make every week.

Paper Choice (February 15): You must choose a collaborator and three candidate papers to replicate by this date. As part of the submission of Homework 1, you will include a PDF copy of the three papers along with a brief paragraph explaining your choices.

Get data (March 1): You should acquire the replication data by this date. In our experience, this can sometimes be difficult.

Memo (March 29): By this date, you will turn in a replication memo with updates on your progress. This memo should summarize the replication and how you plan to extend the paper, and should outline a table of contents for the final paper. You will also upload all data, code, and information necessary to replicate the results of your analysis and reproduce your tables and figures. If your data are private or restricted, your memo should provide instructions for obtaining the data. We will invite each of you to a class folder on Google Drive to upload all of these materials.

Peer review (April 5): You will replicate the analyses of two other groups and write memos to the students, commenting on the replication, the readability of the code, and any ideas for extending the paper. You will upload your memo to the same Google Drive folder as the original replication. You will be evaluated based on how helpful, not how destructive, you are.

Draft poster (April 19): You will submit a preliminary draft of a poster summarizing your replication and extension.

Final poster (April 26): You will submit a final draft of your poster for us to print.

Poster Session (May 4): We will have a poster session in which students can share their results and get feedback from others in the class and the broader community. The time is tentative, but this will probably occur in the time when we ordinarily have precept (10am - 12pm). This is part of a broader graduate research day in the Sociology department which includes lightning talks by the second year Sociology students on their empirical papers.

Paper (May 15): The final version of the paper is due on Dean's Date, Tuesday, May 15. We will discuss the format for the paper more in depth in class but it will be loosely based on the submission format for *The Proceedings of the National Academy of Science* which is given here. This is a fairly rigid and short format that places a heavy focus on a concise presentation of findings. When you upload your paper to Blackboard, you will also upload all materials needed to produce your results (data, code, and any explanation files).

Paper Feedback (May 18): For graduate students, the final assignment of the class is to provide a short memo reviewing papers from two of your classmates. The memos will offer feedback in the style of a review and will help provide some guidance towards publication. This will be due Friday May 18 at 5PM. For undergraduates this will not be a required part of the class as it will be past Dean's Date. You are welcome to participate though if you would like; you just need to let us know ahead of time.

4 Course Outline

The following is a preliminary schedule of course topics. We may adjust the schedule due to comprehension, time, and interest. Readings will be announced in class and posted on Blackboard.

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Quantities of Interest and Binary Outcome Models
- Week 4: Generalized Linear Models, Probit/Logit
- Week 5: Categorical Analysis / Poisson Regression
- Week 6: Event Counts and Duration Modeling
- Week 7: Mixture Models and Expectation Maximization
- Week 8: Missing Data
- Week 9: Design-based sampling and applications
- Week 10: Design-based sampling and applications
- Week 11: Introduction to machine learning (supervised models)
- Week 12: Introduction to machine learning (unsupervised models)

5 Reading Schedule

- February 13 (Introduction):
  - Unifying Political Methodology, pg. 6-58
  - King, Gary. "Publication, Publication." [Link]
- February 15 (Likelihood):
  - Unifying Political Methodology, Chapter 4
- February 20:
  - King, Tomz, Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" *American Journal of Political Science*, Vol. 44, No. 2 (March, 2000): 341-355.
  - Hanmer and Kalkan (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science.*
  - Greenhill, Ward, and Sacks (2011). The Separation Plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science.*
- February 22 (Quantities of Interest):
  - King, Tomz, Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" *American Journal of Political Science*, Vol. 44, No. 2 (March, 2000): 341-355.

  - Hanmer and Kalkan (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science.*
  - Greenhill, Ward, and Sacks (2011). The Separation Plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science.*

- March 27 (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)

- March 29 (Expectation Maximization)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)

- April 3 (Missing Data)
  - King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* 95, 1 (March 2001): 49-69.
  - James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* 54, 2 (April, 2010): 561-581 (Optional)

- April 5 (Missing Data)
  - Blackwell, Matthew, James Honaker, and Gary King. 2014. "A Unified Approach to Measurement Error and Missing Data: Overview, Details and Extensions" *Sociological Methods and Research* (Optional)

- More to come . . .

6 Inspirations

The development of this course has been influenced by a number of people particularly: Brandon Stewart, Matt Blackwell, Jens Hainmueller, Kosuke Imai, Gary King, Teppei Yamamoto. Thanks to all of these excellent teachers for sharing their slides and syllabi with me.