Sociology 401/504: Advanced Social Statistics

Brandon Stewart, Rebecca Johnson and Ian Lundberg*

Class: MW 9:00AM-10:30AM (Wallace Hall 165)
Lab: Th 9:00AM-11:00AM (Wallace Hall 165).

Brandon Stewart
bms4@princeton.edu, scholar.princeton.edu/bstewart
Phone: 609-258-5094
Office Hours: When the door is open (Wallace Hall 149)

Rebecca Johnson, Preceptor
raj2@princeton.edu
Office hours: Alternating Fridays 1:30 - 3:30 and Tuesdays 7:30 - 9:30pm (Wallace Hall 190)

Ian Lundberg, Preceptor
ilundberg@princeton.edu
Office hours: Alternating Fridays 1:30 - 3:30 and Tuesdays 7:30 - 9:30pm (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 Soc400/500 or a similar introduction to statistics class which covers basic probability, regression and causal inference. Please speak with the instructor if you have not taken Soc400/500 but are considering taking it.

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:

*Last Edited: February 6, 2017

- Conduct, interpret, and communicate results from analysis using generalized linear models.
- Conduct additional study of more advanced topics in quantitative methods
- Build a solid, reproducible research pipeline to go from raw data to final paper.

In terms of statistical content Soc 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.

All students will have needed some previous study in linear regression, preferably Soc 500. However, other classes which cover the matrix approach to multiple regression are also acceptable. Come talk to the instructor if you are unclear whether you have sufficient prerequisites.

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.*

Required But Free

- Matloff, Norman. 2011. *The Art of R Programming: A Tour of Statistical Software.* No Starch Press. (Available free through the library).
- Blitzstein and Hwang. 2014. *Introduction to Probability*
- A variety of papers, book components will be assigned as well, available on the web.

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.
- 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 Project:** A co-authored paper. See below.

[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.

3.1 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 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.
To this end we will be using a new tool called *Perusall* for the King book *Unifying Political Methodology*. Perusall is a new ebook platform with collaborative annotation that allows you to post and answer questions directly in the text itself. This gives us the opportunity to answer questions outside of class in the text itself. So asking good questions not only helps you, it helps your classmates. If you know the answer to a question that another student posted, please make a contribution to the class and try to answer it!

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.

New to Soc504/401:
Each week's homework will be made available on Blackboard starting Wednesday at noon and is due on Blackboard Wednesday the following week (7.5 days later) at **11:59 pm**. Please bring a printed copy to precept Thursday morning or put one in the Soc 504 box in the mail room.

Solutions will also be available directly after lab through Blackboard. The homeworks will then be graded on a 50-point scale and returned to you within two weeks.

The problem sets including looking at the solutions key is an extremely important part of the learning process, so please keep up with the work!

You can have **one** no questions asked extension of one week on a problem set of your choosing. If you don't take an extension, we will drop your lowest grade (of any partially completed problem set). If all your problem sets are completed and with top-level grades (such that dropping the lowest wouldn't help you), we will add a 15 point grade bonus to your final exam. When submitting the work on which you claim the extension please include a note indicating the original date and that you are claiming your one extension; you do not need to explain why you are taking the extension. If you exceed the one-week extension period your grade will drop on the problem set by 10% per day down to 30%. You have until the beginning of reading period to submit a late problem set to get the 30% minimum.

After you use the one extension, any subsequent late homeworks will be penalized by 5 points (10%) off per day late, down to the 30% minimum. We will relax late penalties only in the case of documented emergencies.

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 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. All code written for this class should follow the Hadley Wickham's R Style Guide. Good coding style is an important way to increase the readability of your code (even by a future you!). We will explain how to automatically check your code style using RStudio and a package called lintr.

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 Soc400/500 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

The preceptors will be holding 2 hours of office hours each week. Brandon will explain his office hour policy in class.

New to Soc504/401:
You will also be able to post questions directly in the reading through Perusall. Seeking help on the NC problems will be described further in 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 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 odd others as the semester goes on.

Paper Choice: You must choose a co-author and three candidate papers to replicate by February 17 at 5pm, by which point you should submit via email a PDF copy of the paper to the instructor and both precepters along with a brief paragraph explaining your choices.

Get data: You should acquire the replication by March 1.

Memo 1 and Response: On March 15, you must turn in a replication memo with updates on your progress. This memo should include data and code to replicate the main tables and figures of the original paper you have chosen to replicate. You will e-mail both us and two other groups who have been selected to review your work. Your task for the following week is to replicate the other students' analyses and write memos to the students (with copies to us), commenting on the replication, the readability of the code, and any ideas for extending the paper. You will be evaluated based on how helpful, not how destructive, you are.

Memo 2 and Response: On March 29, you must turn in a draft of the paper with little text but with figures and tables, and a proposed table of contents for your paper, in a relatively polished form. You should include all the data and information necessary to replicate the results of your analysis and reproduce your tables and figures. You will e-mail these files to both us and the two other groups who have been selected to review your work. Your task for the following week is to replicate the other student's analysis and write a memo to this student (with a copy to us), pointing out ways to make the paper and the analysis better. You will be evaluated based on how helpful, not how destructive, you are.

Poster Session On May 3rd (time TBD) 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. This is part of a broader graduate research day in the Sociology department which includes lightning talks by the 2nd year Sociology students on their empirical papers. We will get the posters printed for you but they will be due **Thursday April 27th at 5 pm** to allow enough time for printing.

Paper The final version of the paper is due on Dean's Date Tuesday May 16, 5PM. 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.

Paper Feedback For graduate students final assignment of the class is that you will 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 19 at 5PM. For undergraduates this will not be a required part of the class as it will be past Dean's Date in order to allow you as much time as possible to work on your papers. 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.

- Week 1: Introduction and Theories of Inference
- Week 2: Maximum Likelihood Inference
- Week 3: Qois 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: Model Dependence and Matching
- Week 10: Mediation Analysis
- Week 11: Regularization and Hierarchical Models
- Week 12: Multilevel and Hierarchical Modeling

5 Inspirations

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