Sociology 500: Applied Social Science Statistics

Brandon Stewart, Clark Bernier and Elisha Cohen*

Class: 9:00am–10:30am Mondays and Wednesdays (Wallace Hall 165)
Precept: 1:00pm–3:00pm Thursdays (Location TBD).

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

Clark Bernier, Preceptor
cbernier@princeton.edu
Office hours: Tuesday 3:30pm-5:30pm (Wallace 190)

Elisha Cohen, Preceptor
eacohen1@princeton.edu
Office hours: Wednesday 5pm-7pm

1 The Basics

1.1 Overview

This course is the first of the two-semester sequence for Ph.D. students in Sociology. 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:

- Critically read, interpret and replicate the quantitative content of many articles in the quantitative social sciences
- Conduct, interpret, and communicate results from analysis using multiple regression (including dummy variables and interactions).
- Explain the limitations of observational data for making causal claims, and begin to use existing strategies for attempting to make causal claims from observational data.
- Write clean, reusable, and reliable `R` code.

*Last Edited: September 16, 2015

- Feel empowered working with data

More broadly by the end of the year, 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
- Build a solid, reproducible research pipeline to go from raw data to final paper.

In terms of statistical content Soc 500 covers basic probability, univariate inference and linear regression. Soc 504 will cover maximum likelihood estimation, generalized linear models and assorted topics.

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.

As an important part of the sequence, students learn how to make novel contribution to the scholarly literature. We will do this through a replication project in the second semester of the sequence. For those curious, a similar project was undertaken last year in Soc 504 which you can see here.

1.2 Class and Precept

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 precept meets once per week and will focus on practical computational skills. Both are an essential part of the learning process.

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.

Formally, the prerequisites vary for different types of students. For graduate students in the Sociology Department there are no course prerequisites. For anyone else, send Brandon an email. It is helpful but not essential if at some point you have taken calculus and have been exposed to basic matrix operations.

1.4 Online material

Here we will include a link to the full class website on Blackboard which will include detailed lecture notes, precept notes, a PDF version of this document, assignments, and other materials.

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 make extensive use of a flexible (open-source and free) statistical software program called

R as well as a number of companion packages. R is probably the most widely used statistical software. We recommend using R with RStudio. You will learn how to program in this class, if you do not know already.

2.2 Books

Required

- Gelman, Andrew and Hill, Jennifer. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. (We will reuse this in the second course in the sequence as well)

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*
- Imai, Kosuke 2015. *A First Course in Quantitative Social Science.* NOTE: this is being provided to us early as a favor by the author. Please do not distribute, share, or post in a public way.
- Hernn, Miguel A. and James M. Robins. 2012. *Causal Inference*. Forthcoming, Cambridge University Press. (Note that this book is still being written and you can find draft PDFs on the linked page above.)
- 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.
- Berksekas, Dimitri P. and John N. Tsitsiklis, *Introduction to Probability.* Athena Scientific. (Also available as lecture notes online.)
- Freedman, David, Pisani, Robert, and Purves, Roger. 2007. *Statistics.* W.W. Norton & Company. 4th edition.
- Morgan, Stephen L, and Christopher Winship. 2014. Counterfactuals and Causal Inference: Second Edition. Cambridge University Press. (although the first edition is also good).
- Ashenfelter, Orley, Levine, Philip, and Zimmerman, David. 2003. *Statistics and Econometrics: Methods and Applications.* John Wiley & Sons.
- Wooldridge, Jeffrey M. *Introductory Econometrics.* New York: South-Western. 5th edition. (earlier editions are fine)
- Wickham, Hadley. *Advanced R* (available free online)

3 Assignments

There are three main types of assignments:

1. **Preparing for class and precept:** For many classes and some precepts 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. **Final exam:** A cumulative take-home final exam will conclude the semester.

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. In addition, though, engage with the readings: take notes, write down your impressions or confusions, talk with your classmates, preferably through Piazza (see below for more details). All of your classes should be pushing your research forward and you will be more creative the more you actively read.

3.2 Problem Sets

Methods are tools and it isnt 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. 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 Wednesday at noon and is due Thursday the following week (8 days later) at the start of precept. Solutions will also be available directly after precept through Blackboard. The homeworks will then be graded on a $(+, \checkmark, -)$ basis (including half grades between these categories) and returned to you within two weeks. No late homework will be accepted except in the case of a documented emergency.

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!

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. This policy applies only to homework assignments and not to the take-home final.

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 o 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. Precept 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

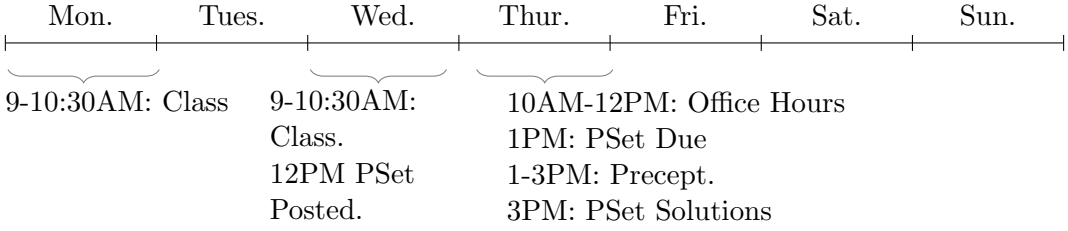
The preceptors will be holding 2 hours of office hours each week. Brandon will explain his office hour policy in class.

3.4 Grading

Final grades will be a weighted average of the final exam (30%) and the weekly problem sets (70%). We reserve the right to provide some bonus credit for active *participation* inside and out of class. For example a student who actively assists their classmates on piazza by answering questions or who engages productively in class might be entitled to a small bonus.

3.5 Weekly Schedule

The timeline below gives the outline of the weekly schedule. We will fill in as more times are set.



4 Course Outline

The following is a preliminary schedule of course topics. We may adjust the schedule due to comprehension, time, and interest. Please note also that readings are subject to change, in particular

you should expect that we will add individual articles for discussion.

4.1 Introduction - Sept 16

- Course Details, Outline and Requirements
- What are the goals of the course?
- What are the goals of this sequence?

Reading

- Gelman and Hill, Ch. 1
- Imai Ch. 1

4.2 Basic Probability and Random Variables - Sept 21, 23

- The basics of probability
- Marginal, joint, and conditional probability
- Law of total probability
- Independence
- Random Variables and Functions of Random Variables

Reading

- Gelman and Hill, Ch. 2.1
- Blitzstein and Hwang 1-1.3, 1.8, 2-2.5, 3-3.2, 3.10, 4-4.2, 4.4-4.6
- Optional: Imai Ch. 6 (if you find Blitzstein and Hwang too hard)
- Optional: D'Amour, Stewart and Zemplenyi 2012 (if you are unfamiliar with integrals)

4.3 Multiple Random Variables - Sept 28, 30

- Joint and conditional distributions
- Conditional expectation
- Covariance, correlation, and independence

Reading

- Blitzstein and Hwang 7.0-7.3
- Optional: Review Imai Ch. 6

4.4 Inferences about a Single Variable - Oct. 5, 7

- Populations, samples, estimation
- Small samples, large samples, and asymptopia
- Properties of estimators
- Hypothesis testing, confidence intervals
- Understanding the concerns about p-values part 1

Reading

- Gelman and Hill 2.2-2.6, 7.1

4.5 Regression, Causality, and the Statistical Model - Oct. 12, 14

- Potential outcomes and causal inference
- Difference in means
- Nonparametric regression
- Parametric models and linear regression
- Bias-variance tradeoff

Reading

- Momentous Sprint at the 2156 Olympics, by Andrew J. Tatem, Carlos A. Guerra, Peter M. Atkinson, and Simon I. Hay, Nature 2004.
- Gelman and Hill Ch. 9
- Imai Ch. 2

4.6 Simple Linear Regression in Scalar Form - Oct. 19, 21

- Mechanics of Ordinary Least Squares
- Assumptions of the linear model
- Properties of least squares
- Inference with regression

Reading

- Gelman and Hill Chapters 3 and 4, Section 7.1
- Imai 4.2

4.7 Linear Regression with Two Regressors - Oct. 26, 28

- Mechanics of regression with two regressors
- Omitted variables and multicollinearity
- Dummy variables, interactions, and polynomials

Reading

- Review Gelman and Hill Ch. 3 and 4, Read Section 7.2

4.8 Multiple Linear Regression- Nov 9, 11

- Matrix algebra and mechanics of multiple linear regression
- Inference in a multiple linear regression model

Reading

- Imai 4.3-4.3.3
- Fox *Applied Regression* , Ch. 9, Sections: 9.1 - 9.2. (skip sections 9.1.1 and 9.1.2, posted on blackboard)
- Optional: Fox *Applied Regression* Appendix B.1.1 and B.1.2 (if you need a review of matrix algebra)

4.9 Statistical Inference for Least Squares and Star Gazing- Nov 16, 18

- Properties of the Least Squares Estimator
- Understanding the concerns about p-values part 2
- Visualizing Results

Reading

- Nunzo, R. (2014) Scientific method: Statistical errors *Nature*.
- Leek and Peng (2015) Statistics: P values are just the tip of the iceberg *Nature*
- Cohen, J. (1994). The earth is round ($p < .05$) *American Psychologist*.
- Simmons, J. et al. (2014) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*.
- Optional: King, G. Tomz, M., and Wittenberg, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science*.
- Optional: Ward, M.D., Greenhill, B.D., and Bakke, K.M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research.*

4.10 Diagnosing and Fixing Problems- Nov 23, 30, Dec. 2

- Functional form
- Model fit, outliers, and influential observations
- Heteroskedasticity and non-Normal errors
- Measurement error

Reading

- Gelman and Hill Ch. 25

4.11 Panel Data and Robust Inference- Dec. 7,9

- Fixed effects
- Random effects
- Clustered standard errors
- Hierarchical models

Reading

- Gelman and Hill Ch. 11

4.12 Causality Revisited- Dec. 14, 16

- The Assumption of No Unmeasured Confounding
- Choosing Conditioning Variables
- Regression Based Estimation (Additive and Interactive)
- Causal Questions Not Addressed in the Course

Reading

- Hernn and Robins Chapters 1-3
- Review: Gelman and Hill Ch. 9

5 Inspirations

The development of this course has been influenced by a number of people particularly: Adam Glynn, Kevin Quinn, Matt Salganik, Gary King and Matt Blackwell. Thanks to all of these excellent teachers for sharing their slides and syllabi with me.