Sociology 500: Applied Social Science Statistics

Matthew J. Salganik, Cambria Naslund, and Lai Wei*

Class: 1:30pm–2:50pm Mondays and Wednesdays (Room: Wallace 002)
Precept: 10:00am–11:50am Thursdays (Room: Wallace 165)

Matthew J. Salganik
mjs3 [at] princeton [dot] edu
Office hours: Generally Wednesday 3-4pm, but see `https://piazza.com/class/k0cxksxy59q4y6?cid=23`, Wallace 145

Cambria Naslund, Preceptor
cnaslund [at] princeton [dot] edu
Office hours: Monday 3:30-4:30pm, Wallace 290 & Wednesday noon-1pm, Wallace 290

Lai Wei, Preceptor
laiw [at] princeton [dot] edu
Office hours: Tuesday 10am-noon, Wallace 125

1 The Basics

1.1 Course Goals

This is the first course in the Department of Sociology's two-course graduate statistics sequence. Starting from only basic math, we build up a foundation for linear regression and its application to causal inference. Students will learn the statistical and computational principles necessary to perform modern, flexible, and creative analysis of quantitative social data.

By the end of this 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.
- 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 in the tidyverse style.
- Feel empowered working with data

*Last Edited: September 24, 2019

The second course in the course in the sequence, SOC 504, will be offered in the spring. The overarching goal of the two-course sequence is to move you from being consumers of quantitative research to producers of it. The capstone of the two-course sequence is the replication and extension project. In this project, completed in Sociology 504, you and a partner will choose a paper of interest, reproduce the results and then extend them to make something new. The projects are presented at Graduate Research Day in the Spring during a poster session and written up in a paper.

Upon finishing the two course sequence, you 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, interpret the results, and explain the results to someone unfamiliar with statistics. Beyond the two-course sequence, we encourage you to participate in the broader statistical life at Princeton including the Sociology Statistics Reading Group (`https://scholar.princeton.edu/bstewart/sociology-statistics-reading-group`) and the Quantitative Social Science Colloquium (`https://q-aps.princeton.edu/book/QSS-seminar`).

This course will require a lot of hard work from all of us; however, we have structured the class to provide you the maximal return on every hour of work you put in. As you read through this syllabus you will find numerous avenues for seeking help. If you are willing to put in the time, we are always happy to help. Please don't be shy about telling us where you need support.

1.2 Class and Precept

Formal instruction for the course is split into two pieces: class and precept/lab. We have lecture twice a week and will typically focus on statistical material. The precept meets once per week and will typically 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. Learning statistical methods is like learning a new language, and it will take time and dedication to master its vocabulary, its grammar, and its idioms. However like studying languages, statistics and programming yield to daily practice and consistent effort.

We intentionally have no formal pre-requisites. Beyond high-school level algebra, it is helpful to have some familiarity with univariate calculus (essentially knowing what derivatives and integrals are in principle even if you forget how to do the mechanics) and basic matrix operations (matrix multiplications and inverses). If these concepts are unfamiliar, you might want to review the materials from our summer methods camp (`http://pusocmethodscamp.org/`). All the other things you need to know will be covered at least briefly in class (e.g., probability).

Even if you have seen some of the materials in class before (e.g., you had an undergraduate class on linear regression), you will likely find a lot to learn here. By rebuilding the foundations of linear regression from scratch, we help to ensure that everyone is on the same page, can more deeply appreciate the intricacies of these methods, and can have a solid foundation for learning more advanced methods. If you are concerned that you may have already covered the material before, come talk to the instructor.

2 Materials

2.1 Computational Tools

The best way, and often the only way, to learn about data analysis and new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical software program called R, RStudio, and a number of companion packages in the tidyverse style. Problem sets and the final exam will be completed in R Markdown. You will learn how to program in this class, if you do not know already.

2.2 Readings

This class uses extremely detailed lecture slides (it is not uncommon to have 100 slides in a week). We encourage you to think of these as the primary reading in the class. We will also provide accounts of the material from other sources as well. We have found that it is often helpful to see the same material from multiple difference sources.

We will use a collection of books, rather than a single book. This approach has advantages and disadvantages. The disadvantage of this approach is that it can be confusing to switch between authors who emphasize different things and use different notation. However, the advantage of this approach is that you can see the same material presented in different ways and you get practice switching back and forth between different notation (which is a skill that you will need to develop in order to continue your education beyond this class).

The required readings will be drawn from these books:

- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. (available online through JSTOR)
- Aronow and Miller. 2019. *Foundations of Agnostic Statistics.* (available online through the library)
- Blitzstein and Hwang. 2014. *Introduction to Probability* (available online through the library)
- Healy, Kieran. 2018. *Data Visualization: A Practical Introduction.* Princeton University Press. (available online for everyone)
- Hernán, Miguel A. and James M. Robins. Forthcoming. *Causal Inference*. Boca Raton: Chapman & Hall/CRC. (Note that this book is still being written and you can find draft PDFs on the linked page above.)
- Morgan, Stephen L, and Christopher Winship. 2014. Counterfactuals and Causal Inference: Second Edition. Cambridge University Press. (available online through the library)

In addition to materials from these books, there are will also be articles that are assigned.

2.3 Optional readings

We recommend these books for students who want to see the statistical material presented in a different way:

- Imai, Kosuke. 2017. *A First Course in Quantitative Social Science.*
- Fox, John. 2016 *Applied Regression Analysis and Generalized Linear Models. 3rd Edition.*

The Imai book is excellent but geared at an introductory undergraduate audience and thus doesn't generally have sufficient depth for this class. However, students in the past have found it to be a

useful resource for starting out in a particular area. The Fox book is one that we have used in prior iterations of this class, but ultimately it can only support a small fraction of the material.

We recommend these books for students who want to see the programming material presented in a different way:

- Grolemund and Wickham. 2017. *R for Data Science* (available online for everyone)

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. We don't assign that much reading but we do assume that you have read it.
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 Preparing for Class and Precept

There are readings for each topic and they generally cover the mathematical underpinnings. Reading statistical work can be challenging the first time you do it. There will be a temptation to skip over all the math - don't! The math is often where the action is in statistical work. Read carefully and go line by line making sure you understand.

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, and post questions on Piazza (see below). 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

Statistical methods are tools and it isn't 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 problem sets including looking at the solutions key is an extremely important part of the learning process, so please keep up with the work!

Problem sets will be graded from 0-50 points. We also reserve the right to add bonus points for aesthetics including presentable graphs, clear code, nice formatting and well written answers.

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 comparable 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. 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 in late you are on your honor to *not look at the solutions* before submitting your work and you are required to explicitly write on your assignment that you have not looked at the solutions. 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.

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. Good coding style is an important way to increase the readability of your code (even by a future you!). We strongly recommend you follow the code conventions developed by Hadley Wickham and implemented in the package lintr, which is built into R Studio.

If you would like to follow some other set of coding conventions, please contact the instructor.

Collaboration Policy: Unless otherwise stated, 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'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.

There will be up to two problem sets which will not allow collaboration (sort of like take-home midterms).

3.3 Final Exam

The final assessment of the class is a take-home exam. The exam is "open-book" in the sense that you can use the slides, your notes, books, and internet resources to answer the questions. However, the final exam must be completed *by yourself*. The exam will be available during the entire period allocated for take-home exams by the university. It will be approximately the length of a long problem set (although we caution that it might take longer if you are used to collaborating on the problem sets). We encourage you to start early. Before the final exam we will distribute a practice final that will help you in your preparations.

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.

4 How to Learn in this Course

If you find this course challenging, you are not alone. Statistics can be challenging and we cover a lot of ground. However, I am confident that you can handle it. In this section of the syllabus I'm going to provide details on some of the forms of support that we offer in this class and pull back the curtain a bit on the pedagogical design.

> Your primary responsibilities in this class are to *work hard* and *communicate* with us about what you need. You can't learn if you aren't putting in the time. We can't help if we don't know there is a problem.

The course is designed to provide every tool we can think of to help you learn the material. If you are willing to put in the time, we want to ensure that time is used as effectively as possible.

4.1 Resources for Getting Help

There are a few main sources of support in the class.

1. Class and Precept
   We strongly encourage you to be an active participant in class and precept. Ask questions during class if you don't understand something that is happening. You don't even have to have a specific question: just raise your hand and let me know that you aren't following what is happening. I'm always happy to stop and go back.

2. Daily Feedback
   After every class I will pass around a note card and ask you to write down something about class. You can write something you liked or didn't like. Something you want to understand better or want to hear more about it. Maybe you want to know how a piece of material connects to the broader goals of the class. You can even just draw a smiley face. I will address the questions either in class or on Piazza.

3. Readings and Slides
   If you are studying alone and you hit something you don't understand, your first instinct should be to study the readings and slides. There is a lot of material in the slides and they are intended to be reviewed multiple times, not just seen once during lecture.

4. Piazza
   Piazza is a classroom discussion board where you can post questions about the material. 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. A big part of why we use Piazza is because reading other people's questions can be really helpful for bolstering your understanding of the material.

5. Preceptor Office Hours
   Cambria and Lai will both offer two hours of office hours each week. Preceptor office hours are often really useful for getting help with the problem sets.

6. Instructor Office Hours
   I will have office hours for one hour each week. You can stop by and talk to me about any aspect of the class. If you need to schedule a particular time to see me, please reach out by email and we will find a time that works.

7. Problem Set Solutions
   As soon as the problem sets are due the solution is posted. I know it is really tempting to just turn your focus to the next problem set, but if you were at all unsure about something, I highly encourage you to check the solutions. The class is extremely cumulative and it will absolutely help you out to lock down core concepts. Even if a concept in (for example) week 3 seems unrelated to week 4, it may well come back in week 5 or 6.

8. Final Exam Prep
   We will host a review session for the final exam sometime in January during the reading period. We also distribute a practice final exam during the winter break.

9. External Consulting Services
   Princeton offers numerous statistical consulting services. Q-APS offers statistical and formal theory consulting (`https://q-aps.princeton.edu`). This isn't really ideal for problem set help, but can be useful if you just need help understanding a broader concept.

10. Individual Tutoring
    In circumstances where it is deemed necessary, the department has agreed to pay for individual and/or small group tutoring. This would be aimed at helping with basic programming in R and foundational concepts in the course. If you believe this would be beneficial to you, please contact the instructor.

This is a lot of resources but if you can think of something else that would be useful to you, we encourage you to come talk to us. Again, if you are willing to put in the time, we can get you a form of support that matches your needs.

4.2 How is the Course Designed?

At a high level, this course builds up the infrastructure of linear regression and causal inference from the basics of probability. The first four weeks focus on foundation elements: probability, random variables and the basics of statistical inference. The second four weeks covers linear regression and its variants. The final four weeks are devoted to causal identification and estimation.

Each week covers a specific topic with the two lectures in a given week usually closely connected and included in a common slide deck. Lecture will often go by really quickly and you may hear a lot of things for the first time. It is designed to fill you in on the core statistical ideas animating the week's topics. The lecture won't focus on code, it will focus on the underlying logic and the applications of the ideas to social science research. Precept will review core ideas of lecture and teach you the programming tools necessary to implement the things shown in lecture. The code shown in precept is very closely tied to what you will need to complete the problem sets.

The problem sets are where I expect the majority of the learning will be solidified. These assignments are challenging and time consuming, but it is only through carefully engaging with the material that you will cement your understanding of it. If at the end of lecture you feel like you don't have a good handle on the material — that's to be expected. If after the problem set has been submitted you still feel uneasy with the material, you should come to office hours and talk to one of us about it.

Finally, there is a strong focus on the class on understanding why things work rather than just applying them. For example, we will often program up our own functions for things `R` has built in functionality for. Why do we do that? By programming it ourselves or deriving a known result, we force ourselves to really understand the underlying mechanics. This not only improves our understanding of statistical analysis but it also helps learn new things in the future. Our goal isn't just for you to learn this material, it is prepare you to teach yourself new material in the future.

4.3 Advice from Prior Generations of Students

Each year I ask students to provide advice to future generations of students. Here is some advice from prior students responding to "What advice would you give to another student considering taking this course?" I think the advice is great and it may be helpful coming from other students.

- Be ready to spend a lot of time
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're treading water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.
- It will be hard but you will learn so much.
- This course is very challenging but greatly contributed to my understanding of social statistics. If you're truly invested in the subject and willing to put in the work (more than you expect possibly), it will be one of the best courses you've taken.
- This is a course where you will learn a lot and spend most of your time doing the psets. I highly recommend office hours for clarification as lecture covers a lot of material.

4.4 A note to everyone that is not a first year PhD student in the Department of Sociology

This course is completely designed to provide training to first year PhD students in the Department of Sociology. That means that some of the topics covered—and the way that these topics are covered—may not be optimal for undergraduates or PhD students from other departments. Further, undergraduates may find that the course has a different style and pace than the courses they have taken in the past. Finally, I do not know who will be able to enroll in SOC 504 in the spring.

5 Course Outline

Readings are subject to change as the semester goes on. Check the slides for updated reading lists. The required reading is often towards the middle of the road in difficulty and reflects as closely as possible the content I cover in the class. We will talk more in class about the other readings. If you like reading as a mode of learning and these aren't meeting your needs, please come talk to me. There are many, many books covering this same material and no single reading is going to be write for all students. We will find something that fits your needs.

Week 1: Introduction and Probability - Sept 11

- Course Details, Outline and Requirements
- Probability Basics
- Sample Spaces, Events, Law of Total Probability, Bayes Rule

Reading

- Blitzstein and Hwang, Chapter 1 (Probability and counting)
- Optional: Imai, Chapter 6 (probability)
- Optional: Aronow and Miller, Chapter 1

Week 2: Random Variables - Sept 16, 18

- Random Variables
- Marginal, joint, and conditional distributions
- Expectations, Conditional Expectations
- Covariance, correlation, and independence

Reading

- Blitzstein and Hwang, Chapter 2, 3-3.2 (random variables), 4-4.2 (expectation), 4.4-4.6 (indicator rv, LOTUS, variance), 5.1 - 5.4 (continuous random variables), 7.0-7.3 (joint distributions), Chapter 9 (Conditional expectation)
- Optional: Imai Chapter 6 (probability)
- Optional: Aronow and Miller, Chapter 2

Week 3: Learning from Random Samples - Sept 23, 25

- Populations, samples, estimation
- Point estimation
- Properties of estimators
- Interval Estimation

Reading

- Aronow and Miller Chapter 3.1-3.2.6 (IID Random variables and estimation), 3.4.1 (confidence intervals)

Week 4: Testing and Regression - Sept 30, Oct 2

- Hypothesis testing
- Nonparametric regression
- Parametric models and linear regression
- Bias-variance tradeoff
- Regression as a predictive model

Reading

- Aronow and Miller, Chapter 3.4.2 (testing)
- Aronow and Miller, Chapter 4.1.1 (bivariate regression)
- "Momentous Sprint at the 2156 Olympics" by Andrew J. Tatem, Carlos A. Guerra, Peter M. Atkinson, and Simon I. Hay, Nature 2004.
- Optional: Imai Ch 2

Week 5: Simple Linear Regression in Scalar Form - Oct. 7, 9

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

Reading

- Aronow and Miller 4.1.2 (OLS Regression)
- Optional: Imai 4.2

Week 6: Linear Regression with Two Regressors - Oct. 14, 16

- Mechanics of regression with two regressors
- Simpson's Paradox
- Omitted variables and multicollinearity
- Dummy variables, interactions, and polynomials

Reading

- Optional: Fox Ch 5-7

Week 7: Multiple Linear Regression- Oct. 21, 23

- Matrix algebra and mechanics of multiple linear regression
- Inference in a multiple linear regression model
- Concerns about p-values and multiple testing
- Bootstrap

Reading

- Aronow and Miller 4.1.3 (Regression with Matrix Algebra)
- Optional: Fox Chapter 9.1-9.4 skip 9.1.1-9.1.2 (Linear Models in Matrix Form)
- Optional Fox Chapter 10 (Geometry of Regression)
- Optional: Imai Chapter 4.3-4.3.3

Fall Break

Week ≈8: What Can Go Wrong and How to Fix It- Nov 4, 6, 11

- Diagnostics with Residuals
- Unusual and Influential Data → Robust Estimation (Day 1)
- Nonlinearity → Generalized Additive Models (Day 2)
- Unusual Errors → Sandwich Standard Errors (Day 3)

Reading

- Optional: Fox Chapters 11-13, 19
- Optional: Fox Chapter 19 Robust Regression
- Optional: King and Roberts "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis*, 2, 23: 159–179.
- Optional: Aronow and Miller Chapters 4.2-4.4 (Inference, Clustering, Nonlinearity)
- Optional: Angrist and Pishke Chapter 8 (Nonstandard Standard Error Issues)

Week ≈9: Regression in the Social Sciences and An Introduction to Causal Inference- Nov 13, 18

- Presenting Results and Making a Case
- Visualization and Quantities of Interest
- Frameworks for Causal Inference (Potential Outcomes and Causal Graphs)

Reading

- Healy Chapter 1: Look at your Data
- Morgan and Winship Chapter 1: Causality and Empirical Research in the Social Sciences
- Morgan and Winship Chapter 13.1: Objections to Adoption of the Counterfactual Approach
- Angrist and Pishke Chapters 1-2
- Hernan and Robins (2016) Chapter 1: A definition of a causal effect
- Optional: Samii, Cyrus. 2016. "Causal Empiricism in quantitative research. *The Journal of Politics.*
- Optional: Aronow, Peter M. and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science.*
- Optional: Morgan and Winship (2015) Chapter 2 (Potential Outcomes), Chapter 3 (Causal Graphs)

Week 10: Causality With Measured Confounding- Nov 20, 25

- Review of Causal Framework
- The Experimental ideal
- Graphical Models of Causal Effects
- The Assumption of No Unmeasured Confounding
- Choosing Conditioning Variables
- Colliders and Back-Door Criterion

Reading

- Angrist and Pishke Chapter 2 (The Experimental Ideal) Chapter 3 (Regression and Causality)
- Morgan and Winship Chapters 3-4 (Causal Graphs and Conditioning Estimators)
- Hernan and Robins Chapter 3 Observational Studies
- Optional: Elwert and Winship (2014) "Endogenous selection bias: The problem of conditioning on a collider variable" *Annual Review of Sociology*
- Optional: Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects

Thanksgiving

Week 11: Unmeasured Confounding and Instrumental Variables- Dec 2, 4

- The Assumption of No Unmeasured Confounding
- Natural Experiments
- Classical Approach to Instrumental Variables
- Modern Approach to Instrumental Variables
- Regression Discontinuity

Reading

- Angrist and Pishke Chapter 4 Instrumental Variables
- Angrist and Pishke Chapter 6 Regression Discontinuity
- Morgan and Winship Chapter 9 Instrumental Variable Estimators of Causal Effects
- Optional: Hernan and Robins Chapter 16 Instrumental Variable Estimation

Week 12: Repeated Observations and Panel Data- Dec 9, 11

- Fixed effects
- Panel Data and Causal Inference
- Difference-in-Differences

Reading

- Angrist and Pishke Chapter 5 Parallel Worlds: Fixed Effects, Differences-in-Differences and Panel Data
- Optional: Imai and Kim "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data"
- Optional: Angrist and Pishke Chapter 6 Regression Discontinuity Designs
- Optional: Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects

6 Inspirations

The development of that course was in turn influenced by a number of people particularly: Matt Blackwell, Dalton Conley, Adam Glynn, Justin Grimmer, Jens Hainmueller, Erin Hartman, Chad

Hazlett, Gary King, Kosuke Imai, Kevin Quinn, Brandon Stewart, and Teppei Yamamoto. I am grateful to everyone who has contributed to these materials, directly or indirectly. I am also grateful to generations of past preceptors who have had a huge influence on the direction the class has gone including Clark Bernier, Elisha Cohen, Ian Lundberg and Simone Zhang.