

Sociology 504: Advanced Social Statistics

Sam Trejo & Luyin Zhang
Class: Tuesday/Thursday 10:30 a.m.-12:00 p.m.
Lab: Friday 10:00 a.m.-12:00 p.m.
Wallace Hall 165

Professor Sam Trejo

Office: 187 Wallace Hall
Office Hours: Wednesdays 3:45 - 4:45pm
Email: samtrejo@princeton.edu
www.samtrejo.com

Preceptor Luyin Zhang

Office: 225 Wallace Hall
Office Hours: TBD
Email: luyin.zhang@princeton.edu

Note: this course is designed for first-year doctoral students in the social sciences. The prerequisite is Sociology 500 – which covers introductory probability, multivariate linear regression, and the foundations of causal inference – or a similar statistics course. Please speak with the instructor if you have not taken Sociology 500 but are considering enrolling in Sociology 504.

“Sociology is not like physics. Nothing but physics is like physics, because any understanding of the world that is like the physicist’s understanding becomes part of physics...”

– OTIS DUDLEY DUNCAN
NOTES ON SOCIAL MEASUREMENT, 1984

Course Description

This course is the second class in the required first-year statistics sequence for doctoral students in the Department of Sociology. The overarching goal of the two-course sequence is to help students grow from consumers to producers of quantitative social research. Students learn the statistical and computational principles necessary to perform modern and innovative analysis of quantitative social data. The sequence's capstone is a replication project in which students choose a published work of social science, reproduce (and extend) the paper's key results, and then present their findings via a poster session at the Department of Sociology's annual graduate research day on April 25th.

In terms of statistical content, Sociology 504 is divided into two modules. The first module – **Modeling Discrete Outcomes** – concerns generalized linear models and their estimation using maximum likelihood approaches. We cover a variety of strategies for modeling dichotomous, ordinal, categorical, and count outcomes. The second module – **Applied Causal Inference** – focuses on experimental and quasi-experimental methods utilizing the potential outcomes framework (also known as the Neyman–Rubin causal model). We cover randomized experiments, instrumental variables, regression discontinuity designs, matching, difference-in-differences, and synthetic control methods.

Formal instruction for the course is split into lecture (Tuesday/Thursday) and lab (Friday); both are essential parts of the learning process. The lecture covers the core statistical material, whereas the lab focuses on practical computational skills. By the end of the Spring semester, students should be able to read an original scholarly article describing a new statistical technique, implement the model with relevant data using statistical software, interpret the results, and explain the findings to someone unfamiliar with statistics. This course will require a lot of hard work, but the payoff is well worth it; if a student is willing to put in the time, we are always happy to help.

Assignments & Grading

Grades in the course will be assigned according to the following breakdown:

Participation	10%
Problem Sets (5)	50%
Replication & Extension Project	40%

Participation

Students are expected to attend and participate in all class sessions. Please email both the instructor and the preceptor in advance explaining your situation in the event that you need an excused absence.

Problem Sets

This is a hands-on course in statistical analysis; as such, students will complete a problem set that involves the analysis of empirical data roughly every other week. Students will work in groups – assigned at the beginning of that semester – on all problem sets.

Students will be randomly selected to individually present their answers, making collaboration essential. A single problem set from each group is due by 11:59pm on the day before they will be discussed in class. Optional bonus questions will be submitted individually.

Problem sets that display a concerted effort to fully address each question will receive a grade of at least 80%, with the final 20% being awarded for thoroughness, clarity, and the appropriate use of the methods. Late problem sets will be penalized 10% off per day late (except in the case of documented emergencies); if you turn in your problem set late, you are on your honor to **not look at the solution keys** before submitting your work.

Due Dates

- Problem Set #1 (February 12)
- Problem Set #2 (March 3)
- Problem Set #3 (March 24)
- Problem Set #4 (April 7)
- Problem Set #5 (April 21)

Replication & Extension Project

The primary assignment in this course is a research paper – written individually or in pairs – that applies an advanced statistical method to a substantive problem in your field of study. This assignment is structured via the replication and extension of an already published paper. Ultimately, the goal is for each student to produce a publishable article. This assignment serves in place of a final exam. There will be a number of interim deadlines, which are listed below and described further on the replication and extension project handout.

Deadlines

- Paper Selection Memo Due (February 5)
- Data Acquisition Deadline (February 24)
- Replication Memo Due (March 17)
- Peer Feedback Due (March 31)
- Practice Poster Presentations (April 17)
- Department Poster Session (April 25)
- Final Paper Due (May 15)
- Peer Review Reports Due (May 18)

Laptop Use

A growing body of evidence suggests that the use of laptops, tablets, and phones in classrooms tends to be detrimental to learning. In general, we discourage their use on lecture days. However, if you want to use a device during class, we ask that you contact us outside of class to make this request. Those who choose to use their laptops will be asked to sit in the back of the room so as to provide the least distraction to other students. For more context on this policy, see [this video](#).

Course

Calendar

TUESDAY		THURSDAY	
Jan 28th Introduction •	1	30th Logistic Regression & the Probit Model •	2
Feb 4th Maximum Likelihood Estimation I •	3	6th Maximum Likelihood Estimation II •	4
11th Generalized Linear Models •	5	13th Discussion of Problem Set #1 •	6
18th Ordinal, Categorical, & Count Models I •	7	20th Ordinal, Categorical, & Count Models II •	8
25th Multi-Level Models & Regularization •	9	27th Additional Topics •	10
Mar 4th Discussion of Problem Set #2 •	11	6th The Potential Outcomes Framework •	12
11th <i>-Spring Recess-</i>		13th <i>-Spring Recess-</i>	
18th Randomized Experiments •	13	20th Instrumental Variables •	14
25th Discussion of Problem Set #3 •	15	27th Regression Discontinuity Designs •	16
Apr 1st Matching Estimators •	17	3rd Mediation Analysis •	18
8th Discussion of Problem Set #4 •	19	10th Difference-in-Differences •	20
15th Synthetic Control Methods •	21	17th Practice Poster Presentations	22
22nd Discussion of Problem Set #5 •	23	24th Additional Topics •	24

Course Readings

Part I: Modeling Discrete Outcomes •

- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. King, Gary. (Cambridge University Press, 1989).
- *Statistical Methods for Categorical Data Analysis (2nd Edition)*. Daniel Powers and Yu Xie (Emerald Publishing, 2008).
- *Applied Regression Analysis and Generalized Linear Models (3rd Edition)*. John Fox. (SAGE Publications, 2015).

Part II: Applied Causal Inference •

- *Mostly Harmless Econometrics*. Joshua D. Angrist and Jörn-Steffen Pischke. (Princeton University Press, 2009).
- *Mastering 'Metrics*. Joshua D. Angrist and Jörn-Steffen Pischke. (Princeton University Press, 2014).
- *Counterfactuals and Causal Inference (2nd Edition)*. Stephen L. Morgan and Christopher Winship. (Cambridge University Press, 2014).

REQUIRED

OPTIONAL

Introduction

- *Statistical Methods for Categorical Data Analysis Chapter 1*
- *Publication, Publication*. Gary King. (PS: Political Science & Politics, 2006).

Logistic Regression & the Probit Model

- *Applied Regression Analysis and Generalized Linear Models Chapter 14.1*
- *Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects From Limited Dependent Variable Models*. Michael J. Hanmer and Kerem O. Kalkan. (American Journal of Political Science, 2013).

Maximum Likelihood Estimation

- *Unifying Political Methodology Chapters 2 & 4*
- *The Epic Story of Maximum Likelihood*. Stephen M. Stigler. (Statistical Science, 2007).

Generalized Linear Models

- *Applied Regression Analysis and Generalized Linear Models Chapter 15.1*
- *A Model of Text for Experimentation in the Social Sciences*. Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoidi. (Journal of the American Statistical Association, 2016).

Ordinal, Categorical, & Count Models

- *Applied Regression Analysis and Generalized Linear Models Chapter 14.2*

Multi-Level Models & Regularization

- *Estimation in Parallel Randomized Experiments*. Donald B. Rubin. (Journal of Educational Statistics, 1981).
- *Machine Learning for Sociology*. Mario Molina and Filiz Garip. (Annual Review of Sociology, 2019).

The Potential Outcomes Framework

- *Mostly Harmless Econometrics* Chapter 1
- *Statistics and Causal Inference*. Paul W. Holland. (Journal of the American Statistical Association, 1986).

Randomized Experiments

- *Mastering Metrics* Chapter 1 or *Mostly Harmless Econometrics* Chapter 2
- *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*. Matthew J. Salganik, Peter S. Dodds, and Duncan J. Watts. (Science, 2006).
- *Double Jeopardy: Teacher Biases, Racialized Organizations, and the Production of Racial/Ethnic Disparities in School Discipline*. Jayanti Owens. (American Sociological Review, 2022).
- *My School District Isn't Segregated: Experimental Evidence on the Effect of Information on Parental Preferences Regarding School Segregation*. Marissa Thompson and Sam Trejo. (Sociology of Education, 2024).

Instrumental Variables

- *Mastering Metrics* Chapter 3 or *Mostly Harmless Econometrics* Chapter 4
- *Handle With Care: a Sociologist's Guide to Causal Inference With Instrumental Variables*. Chris Felton and Brandon M. Stewart. (Sociological Methods & Research, 2022).
- *Pounds That Kill: The External Costs of Vehicle Weight*. Michael L. Anderson and Maximilian Auffhammer. (Review of Economic Studies, 2014).
- *Community and the Crime Decline: the Causal Effect of Local Nonprofits on Violent Crime*. Patrick Sharkey, Gerard Torrats-Espinosa, and Delaram Takyar. (American Sociological Review, 2017).
- *The Effect of Violent Crime on Economic Mobility*. Patrick Sharkey and Gerard Torrats-Espinosa. (Journal of Urban Economics, 2017).
- *The Effects of Active and Passive Leisure on Cognition in Children: Evidence From Exogenous Variation in Weather*. Thomas Laidley and Dalton Conley. (Social Forces, 2018).

Regression Discontinuity Designs

- *Mastering Metrics* Chapter 4 or *Mostly Harmless Econometrics* Chapter 6
- *Do Better Schools Matter? Parental Valuation of Elementary Education*. Sandra E. Black. (Quarterly Journal of Economics, 1999).
- *Comparing Inference Approaches for Rd Designs: a Reexamination of the Effect of Head Start on Child Mortality*. Matias D. Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. (Journal of Policy Analysis and Management, 2017).

Mediation Analysis

- *Advances in Mediation Analysis*. King Makovi and Christopher Winship. (Research Handbook on Analytical Sociology, 2021).
- *Why Does Parental Divorce Lower Children’s Educational Attainment? A Causal Mediation Analysis*. Jennie E. Brand, Ravaris Moore, Xi Song, and Yu Xie. (Sociological Science, 2019).

Matching Estimators

- *Counterfactuals and Causal Inference Chapter 5*
- *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score*. Paul R. Rosenbaum and Donald B. Rubin. (Journal of the American Statistical Association, 1984).
- *Why Propensity Scores Should Not Be Used for Matching*. Gary King and Richard Nielsen. (Political Analysis, 2019).
- *Adjusting for Confounding with Text Matching*. Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. (American Journal of Political Science, 2020).

Difference-in-Differences

- *Mastering Metrics Chapter 5 or Mostly Harmless Econometrics Chapter 5*
- *Prenatal Exposure to an Acute Stressor and Children’s Cognitive Outcomes*. Florencia Torche. (Demography, 2018).
- *Local Exposure to School Shootings and Youth Antidepressant Use*. Maya Rossin-Slater, Molly Schnell, Hannes Schwandt, Sam Trejo, and Lindsey Uniat. (Proceedings of the National Academy of Sciences, 2020).
- *Simple Approaches to Nonlinear Difference-in-Differences with Panel Data*. Jeffrey M. Wooldridge. (The Econometrics Journal).

Synthetic Control Methods

- *Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects*. Alberto Abadie. (Journal of Economic Literature, 2021).
- *The Effects of the Flint Water Crisis on the Educational Outcomes of School-Age Children*. Sam Trejo, Gloria Yeomans-Maldonado, and Brian Jacob. (Science Advances, 2024).