Spring 2020 | Wed./Fri., 2-3:15pm | Room: 12 Waverly Place L120 | Credits: 4

DS-UA 201: Causal Inference

Anton Strezhnev

Office: Center for Data Science 615, 6th Floor, 60 5th Avenue
Office Hours: Thursdays 1pm-3pm or schedule an appointment by e-mail
as6672@nyu.edu
http://www.antonstrezhnev.com

Teaching Assistant: Sidak Yntiso
Office Hours: Mondays 10am–12pm (7th Floor, Center for Data Science, 60 5th Avenue)
sgy210@nyu.edu

Course Overview

We often want to know the relationship between cause and effect. Almost every domain has significant causal research questions that can drive decisionmaking. Labor economists want to know whether job training programs successfully increase participants' wages. Epidemiologists want to know whether a particular medical treatment improves quality of life. Advertisers want to know whether a marketing campaign is effective at boosting sales. You've probably heard that "correlation does not imply causation." But that raises the question: What exactly is causation and how can it be determined whether an observed relationship is truly causal?

This course will teach you the fundamentals of how to reason about causality and make causal determinations using empirical data. It will begin by introducing the counterfactual framework of causal inference and then discuss a variety of approaches, starting with the most basic experimental designs to more complex observational methods, for making inferences about causal relationships from the data. For each approach, we will discuss the necessary assumptions that a researcher needs to make about the process that generated the data, how to assess whether these assumptions are reasonable, and finally how to interpret the quantity being estimated.

This course will involve combination of lectures, sections and problem sets. Lectures will focus on introducing the core theoretical concepts being taught in this course.

Sections will emphasize application and discuss how to implement various causal inference techniques with real data sets. Problem sets will contain a mixture of both theoretical and applied questions and serve as a way of reinforcing key concepts and allowing students to assess their progress and understanding throughout the course.

As a part of this course, you will be introduced to statistical programming using the R programming language. This is a free and open source language for statistical computing that is used extensively for data analysis in both academia and industry. No prior experience in programming is necessary and we recognize that students will come in with a variety of backgrounds and different levels of experience in programming. This course is designed to emphasize learning by doing and will teach statistical programming with the aim of preparing students to analyze actual data.

Prerequisites

DS-UA 111 (Data Science for Everyone) is a great introduction to probability, statistical inference and programming and is recommended for taking this course. However, because introductory statistics is taught in a variety of ways by a variety of disciplines, we are very flexible in allowing students with other backgrounds in statistics to take this course. Please contact the instructor, Anton Strezhnev (as6672@nyu.edu), if you are interested in enrolling but do not have DS-UA 111 as a prerequisite.

In general, the necessary prior knowledge of statistics required for success in this course is very minimal – if you have a general familiarity with linear regression, you are more than ready for this class. The first few weeks will incorporate a review of the most important concepts (e.g. probability, random sampling, conditional averages) and we will include refreshers whenever additional concepts are introduced throughout the course. The focus of this course is on developing students' ability to reason systematically about causal relationships. We believe that both students with significant experience in data analysis and descriptive statistical inference and those with less prior background will benefit from and be able to succeed in this course.

Logistics

Lectures: Wednesdays, Fridays from 2pm-3:15pm – Location: 12 Waverly Place, Room L120
Sections: Monday: 2pm-2:50pm - 60 5th Avenue, Room C10, Wednesday: 11am-11:50am - 60 5th Avenue, Room C12

You may choose to attend one of either of the two section times each week. We strongly recommend that you try to attend the sections regularly as they comprise a significant element of the course instruction.

Lecture slides will be made available on the course website after each lecture.

We will use Piazza as a course discussion platform and to post announcements. See the NYU Classes course website for the registration link.

Textbooks

The following text is required for the class:

- Imai, Kosuke. *Quantitative Social Science: An Introduction.* Princeton University Press. 2017.

This textbook is designed to introduce students to both statistical computing and causal inference through a variety of applied examples and exercises. We hope that it will be useful to you as a reference even after the course is over. We will primarily be focusing on chapters 1-3 with occasional excerpts assigned from later chapters.

Excerpts from other books and articles will be assigned as readings and posted as PDFs on the course website. In general, you should expect to read about one to two chapters from some of the textbooks. Later weeks will replace some textbook chapters with academic papers. You will often find that the textbook readings for a given week address similar topics

You may find the following books useful, and excerpts from some will be assigned as part of the class. However, these are not required for the course:

- Angrist, Joshua D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. 2009.

- Imbens, Guido W. and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press. 2010.

- Hernán, Miguel A. and James M. Robins. *Causal Inference: What If.* Chapman & Hall/CRC. 2020. (PDF available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/)

Requirements

Students' final grades are based on three components:

- **Problem sets** (35% of the course grade). Students will complete a total of five problem sets throughout the semester. Problem sets will primarily cover topics from the lecture and section for that week and the previous week. Problem sets will be graded on a (+/✓/-) scale with a + awarded for near-perfect work, a ✓awarded for generally good work with clear effort shown but with some errors, and a - awarded for significantly incomplete or incorrect work with major conceptual errors and little effort shown. Problem sets are designed to be somewhat more challenging than both the midterm and final exams and we do not expect students to perform perfectly on each problem set. Problem sets will be assigned on Wednesdays and will be due on the following Thursday (by 11:59pm).

  - *Collaboration policy*: We strongly encourage collaboration between students on the problem sets and highly recommend that students discuss problems with each other either in person or via the course's online discussion board. However, each student is expected to submit their own write-up of the answers and any relevant code. **Students may not copy each other's answers, including any R code**. Any sharing or copying of assignments is considered cheating and will result in an F in the course. A second cheating incident will, by CAS rules, result in a one-semester suspension from the College.
  - *Office hours and online discussion*: Students should feel free to discuss any questions about the problem sets with the teaching staff during sections and office hours. We also strongly encourage students to post questions about both the problem sets and the assigned readings on the course discussion board (Piazza) and respond to other students' questions. Responding to other students' questions will contribute to your participation grade.
  - *Submission guidelines*: Problem sets will be distributed as `PDF` and `Rmarkdown` files (`.Rmd`). You should submit your answers and any relevant R code in the same format: including an `Rmarkdown` file (`.Rmd` extension) and a corresponding compiled `.pdf` file as your submission. `Rmarkdown` combines the text formatting syntax of Markdown markup language with the ability to embed and execute chunks of `R` code directly into a text document. This allows you to present your code, graphical output, and discussion/write-up all in the same document. We recommend that you edit the distributed `Rmarkdown` file for each problem set directly.

- **Take-home midterm and final exams** (25% and 30% of the course grade respectively). The take-home midterm and final are similar in structure to the problem sets and are designed to evaluate your knowledge of the course material. Unlike the problem sets, students are not permitted to collaborate with other students. The teaching staff will answer any clarifying questions on the Piazza discussion board for this course. Details on when the midterm and final will be posted and when they are due will be given later in the semester.

- **Participation** (10% of the course grade). We expect students to take an active role in learning in both lecture and section. Engagement with the teaching staff by asking and answering questions will contribute to this grade. Students can also earn participation credit by interacting with their classmates on the Piazza discussion board.

Computing

This course will also serve as an introduction to statistical computing using the `R` programming language. This is a free and open source programming language that is available for nearly all computing platforms. You should download and install it from `http://www.r-project.org`. Unless you have strong preferences for a specific coding environment, we recommend that you use the free RStudio Desktop Integrated Development Environment (IDE) which you can download from `https://rstudio.com/products/rstudio/download/#download`. In addition to being a great and simple to use environment for editing code, `RStudio` makes it very easy to write and compile `Rmarkdown` documents: the format in which problem sets will be distributed. In addition to base `R`, we will introduce students to data management and cleaning via the tidyverse set of packages along with basic graphics and visualization using ggplot2.

Schedule

A schedule of topics and readings is provided below. All readings (aside from those from the Imai text) will be posted on the course website. Each week has 2 lectures and 1 section. Some topics will span multiple weeks. This schedule is subject to change depending on time, student interest, and how the class feels about the course's pacing.

Week 1: Introduction (January 29)

- Course Introduction, Requirements, Outline

Readings

- Probability Review: Imai, Chapter 6

Statistical Review (January 31)

- Random variables and representing uncertainty
- Estimation and conditional mean functions

Readings

- Imai, Chapter 1.3 (pp. 10-27)
- Imai, Chapter 2.1 - 2.2, (pp 32-46)

Week 2: The Potential Outcomes Model (February 5 – February 7)

- Counterfactual reasoning and the "Fundamental Problem of Causal Inference"
- Estimands and causal quantities of interest
- Causal identification versus estimation.

Readings

- Review of Estimation: Imai, Chapter 7
- Imbens and Rubin, Chapter 1 (pp. 3-22)
- Imai, Chapter 2.3 (pp. 46-48)
- Hernán and Robins. Chapter 1 (pp. 3-12)

Problem Set 1 Assigned: Feb 5, Due Feb 13

Week 3/4: Randomized Experiments (February 12 – February 21)

- Why randomization allows identification of causal effects
- Estimation and inference for average treatment effects
- How to improve experimental designs via stratification
- Randomization inference and permutation testing

Readings

- Imai, Chapter 2.4 (pp. 48-54)
- Angrist and Pischke, Chapter 2, "The Experimental Ideal" (pp. 11-24)
- Athey and Imbens, "The Econometrics of Randomized Experiments," *Handbook of economic field experiments.* Vol. 1. North-Holland, 2017. 73-140.
- Bowers, Jake, and Costas Panagopoulos. "Fisher's randomization mode of statistical inference, then and now." (2011).

Problem Set 2 Assigned: Feb 19, Due Feb 27

Week 5: Selection on observables, Part 1 (February 26 – February 28)

- What to do when random assignment of treatment is not possible – common challenges of observational designs
- Assumptions behind “no unobserved confounding” designs
- Representing assumptions using graphical models
- Covariate adjustment via subclassification

Readings

- Imai, Chapter 2.5.1 - 2.5.2
- Cunningham, The Causal Inference Mixtape, Chapter 4 - Directed Acyclical Graphs.
- Hernán and Robins, 2.2-2.3 (pp. 17-20), Chapter 3
- Morgan and Winship, Chapter 5.1 - 5.3

Week 6/7: Selection on observables, Part 2 (March 4 – March 13)

- Sources of bias in observational designs
- Alternative estimation strategies: inverse propensity of treatment weighting, matching, regression

Readings

- Hernán and Robins. Chapter 2.4
- Morgan and Winship, Chapter 5.4
- Morgan and Winship, Chapter 6
- Ho et. al. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” Political Analysis, Vol. 15: 199-236.
- Samii, Cyrus, and Peter M. Aronow. “On equivalencies between design-based and regression-based variance estimators for randomized experiments.” Statistics & Probability Letters 82.2 (2012): 365-370.
- Aronow, Peter M., and Cyrus Samii. ”Does regression produce representative estimates of causal effects?.” American Journal of Political Science 60.1 (2016): 250-267.

Take-home Midterm Assigned: March 4, Due March 12

Spring Break (March 16 – March 22)

Week 8: Fixed Effects estimators (March 25 - March 27)

- How to use repeated observations over time to deal with time-invariant confounding.
- Within-unit vs. between-unit designs.
- Working with panel data

Readings

- Imai, Chapter 2.5.3
- Angrist and Pischke, Chapter 5.1
- Imai and Kim, "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" American Journal of Political Science, Vol. 63, No. 2, April 2019.

Problem Set 3 Assigned: March 25, Due April 2

Week 9: Differences-in-diffferences (April 1 – April 10)

- Weakening "selection on observables" by studying changes over time.
- Assumptions behind the "differences-in-differences" strategy – parallel trends
- Estimation and diagnostics for the identification assumptions.
- Pitfalls and challenges when units initiate treatment at different times.

Readings

- Angrist and Pischke, Chapter 5.2-5.4
- Cunningham, The Causal Inference Mixtape, Chapter 10 - Differences-in-differences
- Goodman-Bacon (2019), "Difference-in-Differences with Variation in Treatment Timing", Working Paper.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119.1 (2004): 249-275.

April 17 – NO CLASS

Week 10/11: Instrumental Variables (April 15, April 22, April 24)

- Estimating effects under unobserved confounding using exogenous variation in treatment induced by an "instrument"
- Assumptions behind the instrumental variable strategy – exogeneity, relevance, "exclusion restriction"
- Estimation via the Wald Estimator and Two-Stage Least Squares
- Interpreting the IV estimand – Local Average Treatment Effect
- What makes a good instrument?

Readings

- Angrist, Imbens and Rubin (1996) "Identification of causal effects using instrumental variables." Journal of the American Statistical Association, 91:434, 444-455
- Angrist, Joshua D., and Alan B. Krueger. "Instrumental variables and the search for identification: From supply and demand to natural experiments." Journal of Economic perspectives 15.4 (2001): 69-85.
- Cunningham, The Causal Inference Mixtape, Chapter 8 - Instrumental Variables

Problem Set 4 Assigned: April 15, Due April 23

Week 12: Regression Discontinuity Designs (April 29 – May 1)

- Estimating effects under unobserved confounding using quasi-random assignment at a cutpoint.
- Common applications: Elections, test scores
- Estimation and sensitivity to modeling assumptions.

Readings

- Angrist and Pischke, Chapter 6
- Imbens, Guido W., and Thomas Lemieux. "Regression discontinuity designs: A guide to practice." Journal of econometrics 142.2 (2008): 615-635.

Problem Set 5 Assigned: April 29, Due May 7

Week 13: Causal Inference in Industry (May 6 - May 8)

- Applications of causal inference in the “data science” profession.
- Translating concepts between academia and industry (e.g. A/B testing)
- Practical challenges in designing and implementing experiments at scale.
- How to get people to care about causation.

Readings

- Duflo, Esther, Rachel Glennerster, and Michael Kremer. ”Using randomization in development economics research: A toolkit.” Handbook of development economics 4 (2007): 3895-3962.
- Jones, Jason J., et al. ”Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election.” PloS one 12.4 (2017).
- Varian, Hal R. ”Causal inference in economics and marketing.” Proceedings of the National Academy of Sciences 113.27 (2016): 7310-7315.

Take-home Final Assigned: May 11, Due May 19

Moses Statement

Disability Disclosure Statement: Academic accommodations are available for students with disabilities. The Moses Center website is `www.nyu.edu.csd`. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.