Stat286/Gov2003: Causal Inference with Applications

Kosuke Imai

Professor of Government and of Statistics
Harvard University

Fall 2023

Substantive questions in empirical scientific and policy research are often causal. Does voter outreach increase turnout? Are job training programs effective? Can a universal health insurance program improve people's health? This class will introduce students to both theory and applications of causal inference. As theoretical frameworks, we will discuss potential outcomes, causal graphs, randomization and model-based inference, sensitivity analysis, and partial identification. We will also cover various methodological tools including randomized experiments, regression discontinuity designs, matching, regression, instrumental variables, difference-in-differences, and dynamic causal models. The course will draw upon examples from political science, economics, education, public health, and other disciplines.

1 Contact Information

Instructor
Name: Kosuke Imai
Email: Imai@Harvard.Edu
URL: https://Imai.Fas.Harvard.Edu
Office: CGIS K306
Office hours: Wednesdays 1:30pm – 3:00pm (sign up at https://tinyurl.com/KosukeOfficeHours) or by an appointment

Teaching Fellows

Name:	Sooahn Shin (Gov2003)	Kyla Chasalow (Stat286)
Office hours:	Friday 3:00-4:30	Thursday 4:00-5:30
Section:	Friday 2:00-3:00	Thursday 3:00-4:00
Location:	CGIS Knafel K050 (Tentative)	Science Center 705
Email:	sooahnshin@g.harvard.edu	kyla_chasalow@g.harvard.edu

2 Prerequisites

This course assumes the solid knowledge of

- probability and statistical theory (based on calculus)
- linear models (based on matrix algebra)
- data analysis using R

at the level of either (1) Stat110, Stat111, and Stat139, or (2) Gov2001 and Gov2002 in the new sequence

3 What's the Difference between Gov2003 and Stat286?

The main differences between the two courses will be the following:

- Some of the exercise questions will be different between the two courses with those for Stat286 being more methodologically advanced and those for Gov2003 being more focused upon empirical applications.
- The TF sections will be separately run though students are welcome to attend both.
- The grading will be done separately for each course.
- You will be working with other students in the same course. Gov2003 tends to be taken by graduate students from Government, Public Policy, Sociology, and related disciplines while Stat286 tends to attract graduate and undergraduate students from Statistics, Biostatistics, and related disciplines.

4 The Instructional Tools

We use a variety of instructional tools to run this course.

- **Course website** (`https://canvas.harvard.edu/courses/123396`): This is the entry point to all the course materials. The links to the review questions and problem sets will be posted here too.
- **Course calendar** (`https://bit.ly/stat286gov2003calendar`): For your convenience, class meetings and TF sections as well as various deadlines are made available through this Google calendar.
- **Perusall** (https://app.perusall.com/courses/gov-2003-stat-286-causal-inference-with-applications-30216009): This platform will host lecture videos and lecture slides. You can also ask questions about their contents by annotating relevant parts of the lecture slides and videos rather than directly emailing an instructional staff. We encourage you to watch their Getting Started video and learn about how to ask questions if you are new to Perusall. You can enter the Perusall course site from Canvas.
- **Ed** (https://edstem.org/us/courses/41774): This platform will host all assignments and their solutions. Questions about assignments should be posted here rather than directly emailing an instructional staff. You may find this user guide helpful to orient yourself to the platform.
- **Gradescope** (`https://www.gradescope.com/courses/560236`): This is where you will submit all of your assignments. The grades for the assignments will be available here as well. There is a student guide you can check for any questions about the workflow.

5 The Structure of the Course

This course is structured to make productive use of in-person class meetings where we learn from one another through collaboration. Each student may choose no more than one student as a collaborator for any of the assignments for this course (for the take-home midterm and final exams, you are required to complete on your own). Each of you, however, must write up your own solutions and make a separate submission. **Under no circumstances may you copy someone else's answer**

including computer code, mathematical derivation, and substantive interpretation. Please do not forget to write down your partner's name on your solution to indicate collaboration.

The course is divided into 10 modules. Each module is centered around a particular theme, and is roughly based on the following basic structure.

- **Pre-module assignments**: Watch video lectures available at Perusall *before* the class meetings of each module. *We will assume that you have watched these videos and asked questions you have about them at Perusall.* We also strongly recommend that you use the reading assignments for better understandings of the materials covered in each module.

- **Review Questions**: These questions are designed to better understand the materials presented in the pre-module video lectures. You may also find the assigned readings useful solving Review Questions. You will be working with another student on Review Questions during the class meetings of each module. Review Questions are not graded, and you are not required to submit your answers. The solutions of these Review Questions wil be made available before the next class meeting.

- **Class Meetings**: CGIS South S020, Mondays and Wednesdays 10:30am – 11:45am. For each module, the first class meeting will begin with a short lecture. We assume that students have watched the assigned the video lecture already. Then, you will spend some time working together with your partner on Review Questions with the help of the instructional team. The meeting will end with another short lecture. Before the second class meeting, you should solve Review Questions though no submission is required. During the second class meeting, the instructor will give a lecture on additional materials of the module.

- **Problem Sets**: There is one problem set for each module. In collaboration with your partner, you are required to complete the selected questions from Problem Set and submit your own answers to the selected questions. Different questions may be selected for Gov2003 and Stat286 students.

- **TF section**: The TF sections will review the course materials and provide some guidance to solving Problem Sets. There will be separate sections for Gov2003 and Stat286. The section attendance is optional, but students can also attend both sections.

6 Course Requirements

The final grade is based on the following components:

- **Class participation** (10% of the course grade): We evaluate the level of your engagement during the in-person class meetings and TF sections as well as Perusall and Ed online discussions.

- **Problem Sets** (20% of the course grade): You and your partner should choose a total of 6 modules out of 10 modules (3 modules before the midterm and 3 after the midterm), for which you are to submit your answers to the selected questions from Problem Sets. Each module will be weighted equally. If you decide to submit the solutions to additional exercises, we will take your best scores among those submitted. The following rules apply to all problem sets:

  - *Collaboration policy.* Students are allowed to discuss the problem sets with the instructional staff and any other students in class. However, you and your partner are required

to write up your own answers. **Under no circumstances may you copy someone else's answer including computer code, mathematical derivation, and substantive interpretation.**

- *Genarative AI policy.* ChatGPT and other generative AI should not be used when working on problem sets except for the purpose of debugging your computer code.
- *Online help and office hours.* You are strongly encouraged to reach out to the instructional staff through Perusall, Ed, and office hours about any questions you might have about the course materials. Students should also feel free to ask questions and answer the questions posed by others at Perusall and Ed, which will count towards class participation.
- *Submission policy.* All answers including the math and computer code are encouraged to be incorporated into the Rmarkdown file provided by the instructional staff. Handwritten math answers are acceptable once merged into one pdf document. For the exercises, each student should submit the pdf file electronically to Gradescope. Once you upload the PDF file, you will see a list of the questions in the assignment and thumbnails of your file. For each assigned question, click the PDF page(s) that contains your answer. No late submission will be accepted unless you obtain a prior approval from the instructor.

- **Mid-term and final take-home exams** (70% of the course grade): The midterm and final will be 24-hour take-home exams and will consist of both analytical and empirical questions. They cover the first and second half of the materials, respectively, and are equally weighted. You should not consult with anyone including the instructional team. Clarifying questions can be asked by emailing the instructional team. Detailed instructions will be given later in the semester.

7 Textbook

The required textbook for this course is,

- Imbens, Guido W. and Rubin, Donald B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

The purchasing link for Harvard Coop is here. In addition, you may find the following books useful,

- Angrist, Joshua D. and Pischke, Jörn-Steffen. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.
- Hernán M. A., Robins James M. (2020). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming. Available free at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

8 Course Plan

This course intends to provide a brief introduction to the following 10 topics in causal inference. The problem set for each module is due by 10 am on the day listed.

Module 1 Introduction

Class 1 (9/6) Overview of the course

Video lecture Potential outcomes

Reading Imbens & Rubin, Chapter 1 (Optional Chapter 2)

Class 2 (9/11) Potential outcomes

Problem Set 1 September 18

Module 2 Permutation Test

Video lecture Permutation test

Reading Imbens & Rubin, Chapter 5 (Optional Chapter 4)

Class 1 (9/13) Inverting permutation tests

Class 2 (9/18) Conditional randomization tests

Problem Set 2 September 25

Module 3 Inference for Average Treatment Effects

Video lecture Average treatment effects

Reading Imbens & Rubin, Chapters 6, 9 (Skip 9.6–9.7), and 10 (Skip 10.6–10.7)

Class 1 (9/20) Stratified experiments

Class 2 (9/25) Matched-pair experiments

Problem Set 3 October 2

Module 4 Linear Regression and Randomized Experiments

Video lecture Simple linear regression

Reading Imbens & Rubin, Chapters 7, 9 (9.6–9.7), and 10 (10.6–10.7)

Class 1 (9/27) Cluster randomized trials

Class 2 (10/2) Covariate adjustment in randomized experiment

Problem Set 4 October 9

Module 5 Instrumental Variables

Video lecture Noncompliance in randomized experiments

Reading Imbens & Rubin, Chapters 23 and 24

Class 1 (10/4) Instrumental variables

Class 2 (10/11) Instrumental variables in observational studies

Problem Set 5 October 18

Midterm Week

Review Session 1 October 16

Review Session 2 October 18

Midterm Released after Review Session 2, 24-hour take-home submitted by 10 am on October 23

Module 6 Regression Discontinuity Designs

Video lecture Sharp regression discontinuity designs

READING Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Jounal of Econometrics*, 142(2):615–635, February 2008. doi: 10.1016/j.jeconom.2007.05.001.

CLASS 1 (10/23) Diagnostics of regression discontinuity design

CLASS 2 (10/25) Fuzzy and other regression discontinuity designs

PROBLEM SET 6 November 1

Module 7 Observational Studies

VIDEO LECTURE Regression with observational data

READING Imbens & Rubin, Chapters 21 and 22

CLASS 1 (10/30) Sensitivity analysis

CLASS 2 (11/1) Partial identification

PROBLEM SET 7 November 8

VIDEO LECTURE Directed acyclic graphs and backdoor criterion

READING Felix Elwert. *Handbook of Causal Analysis for Social Research (ed. Stephen L. Morgan)*, chapter 13. Graphical Causal Models, pages 245–273. Springer, Dordrecht, 2013. ISBN 9789400760936.

CLASS 3 (11/6) do-calculus and frontdoor criterion

Module 8 Matching and Weighting

VIDEO LECTURE Matching and weighting methods

READING Imbens & Rubin, Chapters 13, 15, and 18 (Optional Chapters 12, 14, and 19)

Jared K. Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, October 2004. doi: 10.1002/sim.1903.

CLASS 1 (11/8) Optimal matching and weighting methods

CLASS 2 (11/13) Semiparametric causal inference

PROBLEM SET 8 November 20

Module 9 Causal Mechanisms and Heterogeneity

VIDEO LECTURE Controlled Direct Effects, Natural Direct and Indirect Effects

READING
- STAT286

  Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, February 2010. doi: 10.1214/10-STS321.
- GOV2003

  Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789, November 2011. doi: 10.1017/S0003055411000414.

CLASS 1 (11/15) Causal mediation analysis

VIDEO LECTURE Causal heterogeneity

Class 2 (11/20) Individual treatment rules

Problem Set 9 November 29

Module 10 Fixed Effects, Difference-in-Differences, and Synthetic Control Method

Video lecture Difference-in-differences and fixed effects regressions

Reading Angrist & Pischke, Chapter 5

Class 1 (11/27) Nonlinear difference-in-differences design

Class 2 (11/29) Matching and weighting methods for panel data

Problem Set 10 December 6

Video lecture Synthetic control method

Reading Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, June 2010. doi: 10.1198/jasa.2009.ap08746.

Class 3 (12/4) Difference-in-differences, regression, and synthetic control

Final Exam release date TBA