Info 6751: Causal Inference in Observational Settings. Fall 2022

Ian Lundberg, ilundberg@cornell.edu, ianlundberg.org

Lecture
TTh 9:40–10:55am
Gates 310
Streaming in Bloomberg 398

Office Hours
TTh 11am-12pm or by appointment:
calendly.com/ianlundberg/office-hours
Gates 223 or Zoom: cornell.zoom.us/my/ilundberg
Questions can also be posted on Ed in Canvas

Course description. Causal claims play a central role in both science and policy. Scientists want to understand how outcomes respond to inputs, and policymakers want to intervene on inputs to produce desired outcomes. This course will equip students to conduct causal research, with an emphasis on observational settings. The course focuses on (1) defining a causal question, (2) defending causal assumptions, and (3) estimating the target quantity. These steps will be discussed in the potential outcomes framework and the structural causal modeling framework. A theme of the course is that one should design an observational study to mimic a hypothetical randomized experiment. Students will leave the course prepared to evaluate the credibility of causal claims, answer causal questions in their own research, and engage with new methods for causal inference.

Learning goals. Students will learn to

- evaluate the credibility of causal claims
- answer causal questions in their own research
- engage with new methods for causal inference

Who should take this course? The course is designed to support the development of causal research projects. Thus, it is a good fit for PhD students in information science, computer science, statistics and data science, and the social sciences. Master's students are welcome to enroll and undergraduates are welcome with permission from me; these students should be excited about engaging with ideas for research projects.

Prerequisite. Familiarity with basic probability and statistics (e.g., random variables, expectation, confidence intervals). Email me if you are unsure.

Instructional format. Lecture, with streaming from Ithaca to Cornell Tech. In addition, some sessions will involve hands-on group exercises and discussion of students' research proposals.

Course readings. Readings will be available online for free. See Schedule of Topics. Lecture slides will also be posted on the course website. Many readings from:

> Hernán, M.A., and J.M. Robins. 2020. *Causal Inference: What If?* Boca Raton: Chapman & Hall / CRC. PDF available at hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Statistical software. You can use any statistical software you prefer. I use R and will best be able to support you in R. If you are fluent in another software (e.g., Python), you are welcome to use that. The focus of this course is on conceptual ideas, not a programming language.

Typesetting. While typesetting in LaTeX is a useful skill, it is not required. You may handwrite any assignment and upload a scanned copy. Whether you typeset or handwrite will not affect your grade.

Grading. Letter grade or Satisfactory / Unsatisfactory. Grades will be determined by:

1) Problem sets	50% (your two lowest problem set scores will be dropped)
2) Ideas for the research proposal	10%
3) Final research proposal	30%
4) Feedback to two peers on their proposals	10%

For details, see Assignments.

Late work. Late work will be accepted only under exceptional circumstances. Talk to me if needed.

Academic integrity. Each student in this course is expected to abide by the Cornell University Code of Academic Integrity. Any work submitted by a student in this course for academic credit must be the student's own work.

Students with disabilities.[1] Your access in this course is important to me. Please request your accommodation letter early in the semester, or as soon as you become registered with Student Disability Services (SDS), so that we have adequate time to arrange your approved academic accommodations.

- Once SDS approves your accommodation letter, it will be emailed to both you and me. Please follow up with me to discuss the necessary logistics of your accommodations.
- If you experience any access barriers in this course, such as with printed content, graphics, online materials, or any communication barriers; reach out to me or SDS right away.
- If you need an immediate accommodation, please speak with me after class or send an email message to me and to SDS at sds_cu@cornell.edu.
- If you have, or think you may have a disability, please contact Student Disability Services for a confidential discussion: sds_cu@cornell.edu, 607-254-4545, sds.cornell.edu.

Assignments

Assignments consist of problem sets, a final research proposal, and feedback to two peers on their proposals. Assignments are due on the course website in PDF form. All assignments are due on Mondays at 5pm.

1) Problem sets. Due every Monday at 5pm, beginning Aug 29.
The first problem set will be released Aug 22 and due Aug 29. The final problem set will be released Nov 28 and due Dec 5. Problem sets are intentionally brief and focus on key concepts. The answer key will be posted each week after the deadline, and I encourage you to review what you might have missed.

Suggested workflow for readings and problem sets.

Monday	Problem set released
Tuesday morning	Lecture relevant to Problem Set Part 1.
Tuesday afternoon (suggestion)	Read the material related to Tuesday's lecture.
	Complete Problem Set Part 1.
Thursday morning	Lecture relevant to Problem Set Part 2.
Thursday afternoon (suggestion)	Read the material related to Thursday's lecture.
	Complete Problem Set Part 2.
Monday, 5pm	Deadline to submit the problem set.

This workflow is designed to cover material in this order: lecture, reading, problem set. The reason for this order is because lecture will help you allocate time efficiently on the readings and problem sets.

Problem sets in November will be especially abbreviated to allow time to work on the final proposal.

2) Ideas for the research proposal. Due 5pm on Monday Oct 31.
Write 1–3 rough ideas for the research proposal (below). Whatever you write should be less than 1 page total. This is a chance for feedback—I will provide comments on the ideas within one week.

3) Final research proposal. Due 5pm on Monday, Nov 21 (preceding Thanksgiving).
The culminating assignment is a proposal to study one causal question relevant to your own research.

[1] This statement is based on guidelines from Student Disability Services.

Length: 1,000 word limit, excluding references. To count words in a PDF, I suggest using this word counter and checking the box to exclude numbers: montereylanguages.com/pdf-word-count-online-free-tool.html.

Formatting: I suggest embedding any figures and tables in the text. Your proposal must include the following:

1. *Define the causal estimand.* A mathematical definition of the causal estimand, using either potential outcomes or structural causal models. There may be several causal estimands, but one well-specified estimand is preferable to many incompletely-specified estimands.
2. *Motivate the estimand.* Motivation for why one should care about that estimand. This might involve existing theories and past work, and/or a conceptual argument that you want to make.
3. *Identify the estimand.* A mathematical statement of identification assumptions, using either potential outcomes or structural causal models.
4. *Define the data you would want.* Tell us about the data you would use to answer the question.
    - Note: The project does not need to be feasible today. You do not need to possess the data. However, the data should exist or be possible to collect. It should be the case that you could carry out the project if given the right resources. If you would need $100,000 to collect data, simply say so. If you would need access to data held privately by a company, say so. Explain exactly what the data would be. A good project may also involve no costs and data that you already possess. The goal of the proposal is to motivate and define an important question and outline how you would answer that question. The degree of difficulty in carrying out the project will not be considered in grading.
5. *Define an estimator.* Tell us how you would use the data to estimate the causal estimand. What statistical model or estimation algorithm would you use? What statistical assumptions (e.g., functional form) does it entail?

Proposals will be graded on the clarity of the components above.

4) Feedback on proposals. Due at 5pm on Monday, Dec 5.
You will provide written feedback to two peers on their proposals. The written feedback to each author should be between half a page and one page. You should start by summarizing the author's proposal. Then, consider the evaluation criteria above. Offer suggestions for improvement, and also emphasize what you see as the strengths in the proposal.

Schedule of Topics

Part 1. Inference without models.		
Aug 23.	Causal questions: Observing and intervening	Hernán and Robins (2020) Ch 1 [pdf]
Aug 25.	The target trial	Hernán (2016) [pdf]
Aug 30.	Consistency: Defining potential outcomes	Hernán and Robins (2020) 3.4–3.5 [pdf]
Sep 1.	Sharp bounds and the limits of assumption-free inference	Mullahy et al. (2021) [pdf]
Sep 6.	Exchangeability: Assumptions to block backdoor paths	Greenland et al. (1999) [pdf]
Sep 8.	Population-average causal effects from samples	Westreich et al. (2019) [pdf]
Sep 13.	Positivity: Recognizing the problem of empty cells	Hernán and Robins (2020) 3.3 [pdf]
Part 2. Inference with models.		
Sep 15.	The parametric g-formula: Categorical treatments	Hernán and Robins (2020) Ch 13 [pdf]
Sep 20.	The parametric g-formula: Continuous treatments	Rothenhäusler and Yu (2019) [pdf]
Sep 22.	The generality of the g-formula: Using any estimator	Naimi and Balzer (2018), [pdf]
Sep 27.	The g-formula by matching	Stuart (2010) [pdf]
Sep 29.	The g-formula with propensity scores	Brand and Xie (2010) [pdf]
Oct 4.	Inverse probability weighting	Hernán and Robins (2020) 12.1-12.3 [pdf]
Oct 6.	Marginal structural models	Hernán and Robins (2020) 12.4-12.6 [pdf]
Part 3. Dynamic causal inference.		
Oct 13.	Treatments in many time periods	Hernán and Robins (2020) Ch 19 [pdf]
Oct 18.	Mediation: Controlled direct effects	Acharya et al. (2016) [pdf]
Oct 20.	[No class. Fall break.]	
Oct 25.	Mediation: Natural direct and indirect effects	Imai et al. (2011) [pdf]
Part 4. Complexities that arise in real settings.		
Oct 27.	Defining the estimand is hard	Lundberg et al. (2021) [pdf]
—Deadline. Ideas for the research proposal due Oct 31—		
Nov 1.	Principal stratification: Addressing undefined outcomes	Page et al. (2015) [pdf]
Nov 3.	Principal stratification: Bias in policing	Knox et al. (2020) [pdf]
Nov 8.	Unknown functional forms: Two chances	Glynn and Quinn (2010) [pdf]
Nov 10.	Unknown functional forms: Flexible learners	Díaz (2020), [pdf]
Nov 15.	Measurement error: The problem	Hernán and Cole (2009), [pdf]
Nov 17.	Measurement error: Using proxies	Knox et al. (2022)
—Deadline. Final research proposal due Nov 21—		
Nov 22.	Beyond backdoor adjustment: Regression discontinuity	De la Cuesta and Imai (2016) [pdf]
Nov 24.	[No class. Thanksgiving.]	
Nov 29.	Beyond backdoor adjustment: Instrumental variables	Hernán and Robins (2020) Ch 16 [pdf]
Dec 1.	Course recap: Causal inference in observational settings	
—Deadline. Feedback to two peers due Dec 5—		

References

Acharya, A., Blackwell, M., and Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110(3):512–529.

Brand, J. E. and Xie, Y. (2010). Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review*, 75(2):273–302.

De la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396.

Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48.

Hernán, M. A. (2016). Does water kill? a call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680.

Hernán, M. A. and Cole, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8):959–962.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25.

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565.

Mullahy, J., Venkataramani, A., Millimet, D. L., and Manski, C. F. (2021). Embracing uncertainty: The value of partial identification in public health and clinical research. *American Journal of Preventive Medicine*, 61(2):e103–e108.

Naimi, A. I. and Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. *European Journal of Epidemiology*, 33(5):459–464.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., and Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, 36(4):514–531.

Rothenhäusler, D. and Yu, B. (2019). Incremental causal effects. *arXiv preprint arXiv:1907.13258*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.

Westreich, D., Edwards, J. K., Lesko, C. R., Cole, S. R., and Stuart, E. A. (2019). Target validity and the hierarchy of study designs. *American Journal of Epidemiology*, 188(2):438–443.