Columbia University School of International and Public Affairs
PhD in Sustainable Development

Causal Inference Workshop

Draft Syllabus

Course Title: Causal Inference Workshop

Instructor: Suresh Naidu, sn2430@columbia.edu

Teaching Assistant: Claire Palandri, cp2913@columbia.edu

Semester: Spring 2022

Meeting Date/s Times: Fridays, 9:00-10:00am

Location: IAB 823 or same room as the SDev Colloquium (Fridays, 10-11am)

Credits: 1.5 (Pass/Fail)

Course Description

This 14-week workshop designed for the students in the PhD program in Sustainable Development covers the fundamental theory and techniques of causal inference. Specifically tailored to students trained in econometrics and positioned to conduct interdisciplinary research, it systematically ties back the econometrics approaches covered to the underlying statistical framework, and provides the students with the tools to conduct rigorous empirical analyses and to share and defend their approach in front of both economics and non-economics audiences. Lower-year students learn the fundamental methods for observational studies; upper-year students discuss how they employ them in their own current research. Participants are presented with the most common methods in the field, their limitations and best practices, and introduced to underused statistical methods for causal inference.

Course Overview

The workshop will consist of a weekly class session, led by the teaching assistant. The 14 weeks are organized into four sections:

A. *[The aim of causal inference]* Fundamentals of inferential statistics are reviewed, followed by the theoretical framework of potential outcomes for causal inference, and how it's implemented with observational data through regression modeling.
B. *[How it's done]* The most common identification strategies — special cases of regression adapted to particular forms of natural experiments — are reviewed. For each method, the canonical setup is presented in the first half of the session, in particular: the data generating process assumed; the identifying assumptions; the estimand of interest; the estimator used; best practices; strengths and weaknesses. The second half of the session then puts the theory in practice: two papers are reviewed and discussed: one working paper by a current PhD student and presented by that student, and a published paper on sustainable development.
C. *[How to do it better]* How to improve a causal analysis along the steps of a good inference workflow. Pre- and post-estimation best practices are presented, including how to support the assumptions on which

the inferences rest, and the benefits of matching and prediction for causal inference, highlighting the different positions within the statistics community.

D. Less common and arguably underused statistical methods for causal inference are presented, specifically randomization inference, synthetic control methods, and directed acyclic graphs. The last session will cover a specific topic selected by the students earlier in the semester from a list of options, such as multilevel modeling of varying treatment effects or Bayesian inference.

The workshop does not follow a specific textbook, but the two references in which the participants will find most of the material covered — and that are highly recommended as complements of each other — are Angrist and Pischke (2008) and Gelman et al. (2020).

Grading The course is graded on a Pass/Fail basis. The course grade will be based on attendance and on a home assignment to be turned in on the final week of class. It will consist of the replication of the analysis of a published paper, to supplement with statistical analyses covered during the course (e.g., diagnosis checks of underlying modeling assumptions, model evaluation, matching and graphical modeling).

Course Structure: Week-by-week list of class topics

A. Causal inference fundamentals

1. **Overall presentation + Inferential statistics fundamentals**
   · Modeling assumptions precede identifying assumptions: assumptions of the classical linear regression model, and the estimator properties depending on them.
   · Making statistical inferences is deducing properties of (conditional) probability distributions: regression models as conditional distributions, implications for binary or count data.

2. **The potential outcomes framework and identification**
   · The Neyman-Rubin causal model or potential outcomes framework, expressions of treatment effects.
   · Identification relies on conditional independence; the goal: make us believe in the CIA.
   · Random assignment balances the Xs by treatment. Statistical techniques can't substitute for good design, relevant data, and testing predictions against reality.
   References Rubin (1974), Freedman (1991)

3. **Endogeneity, good and bad controls**
   · Estimating *an* average treatment effect: the relation between observed and potential outcomes can be written as a regression on the treatment.
   · Endogeneity; sources and consequences (imbalance in potential outcomes).
   · Do control for prior relevant variables, don't for irrelevant ones, beware of correlated predictors.

B. Common identification strategies

4. **Instrumental Variables (IV)**
   · Theory: treatment assignment by an instrument; compliance behavior; two-stage least squares.
   · Applications: (i) working paper by a current PhD student; (ii) published paper(s) on an SDev topic.
   A poll will be sent to the students, to decide the topic of the last session of the semester, among a list of options which may include: multilevel modeling of varying treatment effects; Bayesian inference...
   References TBD

5. **Regression Discontinuity Design (RDD)**
   · Theory: deterministic but discontinuous assignment; non-parametric estimation & optimal bandwith.
   · Applications: (i) working paper by a current PhD student; (ii) published paper(s) on an SDev topic.
   References TBD

6. **Difference-in-Differences (DiD) and event-study**

· Theory: pre-trends; justifying a third difference; beware of weighted sums of the average treatment effects with two-way fixed effects.
· Applications: (i) working paper by a current PhD student; (ii) published paper(s) on an SDev topic.
References de Chaisemartin and D'Haultfœuille (2020); Hsiang and Sekar (2019), TBD

C. Improving one's causal analysis along an inference workflow

7. Pre-estimation: EDA; matching

· Exploratory Data Analysis: scatterplot your raw data (and show some summary in your final paper).
· Matching: *in place of* (Angrist and Pischke, 2008) or *on top of* (Ho et al., 2007; Gelman et al., 2020) regression — but never in place of design. Examples: propensity score matching; Mahalanobis distance matching.
References Rosenbaum and Rubin (1983); Almond et al. (2005)

8. Post-estimation: check and support assumptions

· Modeling assumptions; back to inference fundamentals: post-estimation model diagnostics.
· Identifying assumptions; show a balance test table and do falsification tests. Examples of falsification tests for each identifying assumption of common identification strategies (IV, RD, DiD)

9. LATEs and treatment heterogeneity

· Local average treatment effect (LATEs); computing average complier characteristics and getting more out of a LATE.
· Treatment interactions; discussion on LATEs in effect recognizes that treatment effects vary across units (i.e., D_i interacts with pretreatment variables). Given that, model such treatment interactions.
References Angrist and Pischke (2008, eq. 4.4.8); Kowalski (2018); Abadie (2003); Almond and Doyle (2011)

10. Model selection and prediction

· Regularization methods.
· Prediction isn't part of statistical inference, but can help 1. support your assumptions; 2. prove general interest of your results. Measures of performance: information criteria; cross-validation.

D. Other topics in causal inference

11. Which uncertainty matters? Randomization inference

· Design-based vs sampling-based inference. 3 possible motivations: no true sampling variation to speak of; not having to rely on asymptotics; preserving unformalizable clustered data structures.
· Applications: (i) working paper by a current PhD student; (ii) published paper(s) on an SDev topic.
References Athey and Imbens (2017); Cooperman (2017)

12. Synthetic Control Method

· A new counterfactual: the "synthetic unit".
· Applications: (i) working paper by a current PhD student; (ii) published paper(s) on an SDev topic.
References TBD

13. Other approaches to causal modeling

· Graphical causal modeling with Directed Acyclic Graphs (DAGs); Structural Equation Models.
References Pearl (2009); Cunningham (2021, chap. 3)

14. Topic chosen by the students + Wrap-up

References

Abadie, A. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003. ISSN 0304-4076. doi: 10.1016/S0304-4076(02)00201-4.

Almond, D. and Doyle, J. J. After midnight: A regression discontinuity design in length of postpartum hospital stays. *American Economic Journal: Economic Policy*, 3(3):1–34, August 2011. doi: 10.1257/pol.3.3.1.

Almond, D., Chay, K. Y., and Lee, D. S. The Costs of Low Birth Weight. *Q. J. Econ.*, 120(3):1031–1083, 2005. ISSN 0033-5533, 1531-4650.

Angrist, J. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ, Dec. 2008. ISBN 9781400829828. doi: 10.1515/9781400829828.

Athey, S. and Imbens, G. W. The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier, 2017. doi: 10.1016/bs.hefe.2016.10.003.

Cooperman, A. D. Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation. *Polit. Anal.*, 25(3):277–288, July 2017. doi: 10.1017/pan.2017.17.

Cunningham, S. *Causal Inference: The Mixtape*. Yale University Press, Jan. 2021. ISBN 9780300251685, 9780300255881. URL https://mixtape.scunning.com.

de Chaisemartin, C. and D'Haultfœuille, X. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, September 2020. doi: 10.1257/aer.20181169.

Freedman, D. A. Statistical models and shoe leather. *Sociological Methodology*, 21:291–313, 1991. doi: 10.2307/270939. URL http://www.jstor.org/stable/270939.

Gelman, A., Hill, J., and Vehtari, A. *Regression and Other Stories*. Cambridge University Press, July 2020. ISBN 9781107023987.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007. doi: 10.1093/pan/mpl013.

Hsiang, S. and Sekar, N. Does legalization reduce black market activity? evidence from a global ivory experiment and elephant poaching data. Working Paper 22314, National Bureau of Economic Research, April 2019. URL http://www.nber.org/papers/w22314.

Kowalski, A. E. Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform. Working Paper 24647, National Bureau of Economic Research, May 2018. URL http://www.nber.org/papers/w24647.

Pearl, J. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, second edition, Sept. 2009. ISBN 9780521895606.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688–701, Oct. 1974. ISSN 0022-0663, 1939-2176. doi: 10.1037/h0037350.