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

Causal Inference Workshop

Syllabus

(updated February 10, 2022)

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 1101

Credits: 1.5 (Pass/Fail)

Course Description

This 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 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 are presented the fundamental methods for observational studies; upper-year students can discuss how they employ them in their own current research. Participants are presented with the core methods in the field, their limitations and best practices, and less-used statistical methods relevant for causal inference.

Course Overview

The workshop will consist of a weekly class session, led by the teaching assistant. The 13 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. *[Core of 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 part 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 part of the session then puts the theory in practice by discussing the analyzes of working or published papers: work in progress by a current PhD student and/or a published paper on sustainable development.

C. *[How to improve upon it]* How to obtain stronger causal inferences through steps at the analysis stage. 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.
D. Less common topics in causal inference are presented, specifically randomization inference, the synthetic control method, and directed acyclic graphs.

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 mostly on attendance, and also 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...).

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 limited outcome data.
2. **The potential outcomes framework and identification**
   · The Neyman-Rubin causal model or potential outcomes framework. Identification from independence assumptions.
   · Relation between observed and potential outcomes can be written as a regression on the treatment. Simplest case & extensions (limited Y, covariates X, continuous D).
   References Rubin (1974), Freedman (1991)

B. Design stage: Applied identification methods

3. **Instrumental Variables (IV)**
   · Theory: treatment assignment by an instrument; compliance behavior; two-stage least squares.
   · Local average treatment effect (LATEs); computing average complier characteristics and getting more out of a LATE.
   · Application: working paper by a current PhD student and/or published paper related to SDev.
   References Angrist and Pischke (2008, eq. 4.4.8); Kowalski (2018); Abadie (2003); Almond and Doyle (2011)
4. **Regression Discontinuity Designs (RDD)**
   · Theory: deterministic but discontinuous assignment; estimation with flexible functional forms; Sharp RD, Fuzzy RD (imperfect compliance).
   · Application: working paper by a current PhD student and/or published paper related to SDev.
   References TBD
5. **Difference-in-Differences (DiD) and event-studies**
   · Theory: pre-trends; justifying a third difference; beware of weighted sums of the average treatment effects with two-way fixed effects.
   · Application: working paper by a current PhD student and/or published paper related to SDev.
   References de Chaisemartin and D'Haultfœuille (2020); Hsiang and Sekar (2019), TBD

C. Analysis stage: Steps for stronger causal inferences

6. **Limitations of identification strategies; Pre-estimation steps**
   - Limitations of identification strategies; steps can be taken pre-/during-/post-estimation.
   - Exploratory Data Analysis: scatterplot your raw data (and show some summary in your final paper).
   - Restructuring the data by matching to improve overlap: *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)

7. **Estimation steps: Balance & TE heterogeneity**
   - Good/bad controls, adjusting as much as possible for potential imbalance.
   - Allowing for treatment effect heterogeneity. Interacting D_i with pretreatment variables; multilevel modeling of varying TEs.

8. **Post-estimation steps #1: supporting assumptions**
   - Modeling assumptions; back to inference fundamentals: post-estimation model diagnostics, fit the model to simulated data where you know the true parameter values.
   - Identifying assumptions; show a balance test table and do falsification tests. Examples of falsification tests for each identifying assumption of common identification strategies (IV, RDD, DiD).
   - Fake data simulations

9. **Post-estimation steps #2: model selection; external validity**
   - Model selection; 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. Bayesian inference.

D. Other topics in causal inference

10. **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.
    - Application: working paper by a current PhD student and/or published paper related to SDev.

    References Athey and Imbens (2017); Cooperman (2017)

11. **Synthetic Control Method**
    - A new counterfactual: the "synthetic unit".
    - Application: working paper by a current PhD student and/or published paper related to SDev.

    References TBD

12. **Other approaches to causal modeling**
    - Graphical causal modeling with Directed Acyclic Graphs (DAGs)
    - Structural Equation Models

    References Pearl (2009); Cunningham (2021, chap. 3)

E. Wrap-up & *'Replication+'* exercise

13. **Wrap-up + Exercise**
    - "Replication +" exercise: Using data from a published causal inference paper, redo the analysis, conduct post-estimation checks, explore other methods.

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.