Applied Causal Inference & Machine Learning in Political Behaviour

Felix Hartmann
Email: f.hartmann@hu-berlin.de
Chair of Comparative Political Behaviour, Department of Social Science, Humboldt-Universität zu Berlin
Moodle:

Winter Semester 2021/22

General Information

If you decide not to take the course, please sign out from Moodle.

Where/When We meet weekly **Friday, 12.15-15.45**,

Office Hours Monday 2-3 pm:

Overview

Research questions in political behaviour are often causal. Does voter outreach increase turnout? Can political campaigns persuade voters? Does political performance increase incumbent voting? The course will introduce students to tools to answer these questions and apply them to empirical data from the field of political behaviour. We will begin by introducing statistical theory and practice of causal inference. As theoretical frameworks, we will discuss potential outcomes and randomisation. We will also cover various methodological tools including randomized experiments, regression discontinuity designs, regression, instrumental variables, difference-in-differences. Lastly, we apply machine learning techniques for causal inference using regression trees, Lasso, and causal forests. Empirical topics will include political persuasion, the role of the media, and incumbency voting.

Learning Outcomes

Students will understand the potential outcomes framework, and the key assumptions underlying causal inference, and will be able to choose appropriate methods for a variety of research questions posing different identification challenges. They will be able to assess empirical evidence in political behaviour using the tools of causal inference.

Assessment

The assessment consists of problem sets (20%, pass/fail), a student presentation (20%, pass/fail) and a research design essay (60%, 1.0-5.0), in which students will outline how they would address a causal research question of their choice using methods introduced in class. The research design should be structured like a pre-analysis plan (PAP) and include a short literature review, hypotheses, research design, and estimation.The student presentation should cover a substantive research paper from the field of political behaviour.

Prerequisites

Linear Regression, hypothesis testing, probability theory

Involvement Participation includes coming to class; turning in assignments on time; thinking and caring about the material and expressing your thoughts respectfully and succinctly in class. As much as possible, we will be working in groups during the class meetings. This work will require that you have done the assigned reading in advance and that you are an active collaborator.

Problem Sets As part of the seminar students should submit problem sets as a PDF or html document via Moodle on the **Wednesday after each seminar**. You will use Rmarkdown to prepare your problem sets. All students must write up their problem sets individually. However, you may work in groups of up to three (though you are not required to work in groups). Please indicate at the top of your homework the names of the other students you worked with that week. Don't "share" members across groups. Do not copy and paste the answers across group members. All problem sets are pass/fail. A solution set will be provided on the course website and the activity will be discussed in class.

Presentations Student should also prepare a 20-30 min presentation for a book chapter or paper of your choice. Please sign up here:

Final Paper The central assignment for the class is a research design paper in a form of a pre-analysis plan. I'll provide more information on this as we go along. Hand in research design paper by **March 31st, 2022** as PDF via Moodle.)

Computing For some exercises, we will be using R in class. Please install R (http://www.r-project.org) and R-Studio (https://rstudio.com/products/rstudio/download/#download) on your computers before the first class session. As you work on your papers, you will also learn to write about data analysis in R-Markdown. Always show your code and results in your R-Markdown document.

Books

Required Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer

Applied with R Irizarry, R. A. (2019). *Introduction to data science: Data analysis and prediction algorithms with R*. CRC Press

Prerequisites Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press, Princeton

Schedule

Week 1 (Oct. 22): Introduction

Before Class: Make sure you have a satisfactory setup for statistical computing in R.

Topics: Overview of the course, Introduction to R and RStudio.

Reading: Angrist and Pischke, 2014, Introduction, p. xi-xv , Keele (2015)

Week 2 (Oct. 29): Causal Inference

Topics: R markdown, Causality as counterfactuals, Potential outcomes, Identification and estimation, Causal estimands

Reading: Angrist and Pischke, 2014, p. 1-33, Gerber and Green, 2012, Ch. 1-2.6: p. 1-39

Problem Set Go through chapter 1, 2, and 40 of: https://rafalab.github.io/dsbook/getting-started.html

Week 3 (Nov. 5): Randomized Experiments

Topics: Identification of Causal Effects under Randomization; Covariate adjustment

Reading: Gerber and Green, 2012, Ch. 2.7-3: p. 39-86

Problem Set: Potential Outcomes, ATE vs. ATT vs ATC Experimental Design

Week 4 (Nov. 12): Selection on Observables

Topics: Identification under Selection on Observables

Reading: Angrist and Pischke (2014, p. 47-79); Gelman and Hill (2007)

Problem Set: Applied (with R): Olken (2007)

Week 5 (Nov. 19): DAGs and Instrumental Variables

Topics: Directed acyclic graphs (DAGs); Identification and Estimation via Instrumental Variable: Using Exogenous Variation in Treatment Intake Given by Instruments

Reading: Angrist and Pischke (2014, Ch. 3: p. 98-139), Gerber and Green (2012, Ch. 5-6: p. 131-67 & 173-205)

optional: Video Panel Machine Learning and Causal Inference: https://www.youtube.com/watch?v=B4mK8ERsCIM&t=16s

Problem Set: Applied (with R): Acemoglu et al. (2001)

Week 6 (Nov. 26): Regression Discontinuity Design

Topics: Identification, Estimation, Falsification Checks

Reading: Angrist and Pischke (2014, Ch. 4: p. 147-75)

Problem Set: Applied (with R): Go through chapter 4 of: https://rafalab.github.io/dsbook/getting-started.html; replicate Titiunik (2011)

Week 7 (Dec. 3): Difference-in-Differences

Topics: Identification, Estimation, Falsification tests

Reading: Angrist and Pischke, 2014, Ch. 5: p. 178-204; optional: Imai and Kim (2021)

Problem Set: Applied (with R): Dube and Vargas (2013)

Week 8 (Dec 10.): Matching

Topics: Another observational method to control for confounders

Reading: Sekhon, 2009

Week 9 (Dec 17.): Interaction Effects

Topics: How do we interpret interaction effects?

Reading: Hainmueller et al. (2019)

Problem Set: Applied (with R): Boas and Hidalgo (2011)

Week 10 (Jan 7.): Meta Analysis

Topics: How can we aggregate knowledge from experiments?

Paper: Dunning (2016), Incerti (2020)

Week 11 (Jan 14.): General Introduction to Machine Learning

Topics: difference between prediction versus explanation, supervised versus unsupervised Learning, the bias/variance trade-off, cross-validation and test-data and discuss first simple techniques such as regression, nearest neighbour

Reading: Grimmer et al., 2021, James et al., 2013, Ch. 2 (15 - 52), James et al., 2013, Ch.5 (176 - 184)

Week 12 (Jan 21.): Supervised learning 1: Regression & Classification

Topics: multivariate linear regression, classification tasks

Paper: James et al., 2013, Ch. 3 (59 - 104), James et al., 2013, and Ch.4 (127 - 154)

Week 13 (Jan 28.): Supervised Learning 2: Sparse and flexible Regressions

Topics: statistical learning techniques for regression models; include selection and regularization, talking about the subset selection, shrinkage methods and dimension reduction methods.

Reading: James et al., 2013, Ch. 6 (203 - 243)), James et al., 2013, Ch.7 (265 - 287)

Week 14 (Feb 4.): Supervised Learning 3: Tree based methods and Support Vectors Machines

Topics: decision trees and talk about bagging, random forests and boosting, support vectors machines.

Reading: James et al., 2013, Ch. 8 (303 - 323) , James et al., 2013, Ch.9 (337 - 356)

Week 15 (Feb 11.): Research Design and Pre-Analysis Plans

Topics: Reproducibility, Publication bias, analysis of research designs, Defining your estimand

Reading: Stockemer et al. (2018), Blair et al. (2019)

Optional: Lundberg et al. (2021)

Week 16 (Feb 18.): Final Paper

Topics: Brainstorm your own ideas

I References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Blair, G., Cooper, J., Coppock, A., and Humphreys, M. (2019). Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859.

Boas, T. C. and Hidalgo, F. D. (2011). Controlling the airwaves: Incumbency advantage and community radio in brazil. *American Journal of Political Science*, 55(4):869–885.

Dube, O. and Vargas, J. F. (2013). Commodity price shocks and civil conflict: Evidence from colombia. *The review of economic studies*, 80(4):1384–1421.

Dunning, T. (2016). Transparency, replication, and cumulative learning: What experiments alone cannot achieve. *Annual Review of Political Science*, 19:S1–S23.

Gelman, A. and Hill, J. (2007). Causal inference using regression on the treatment variable.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419.

Hainmueller, J., Mummolo, J., and Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*, 27(2):163–192.

Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press, Princeton.

Imai, K. and Kim, I. S. (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415.

Incerti, T. (2020). Corruption information and vote share: A meta-analysis and lessons for experimental design. *American Political Science Review*, 114(3):761–774.

Irizarry, R. A. (2019). *Introduction to data science: Data analysis and prediction algorithms with R*. CRC Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3):313–335.

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.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy*, 115(2):200–249.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12:487–508.

Stockemer, D., Koehler, S., and Lentz, T. (2018). Data access, transparency, and replication: new insights from the political behavior literature. *PS: Political Science & Politics*, 51(4):799–803.

Titiunik, R. (2011). Incumbency advantage in brazil: Evidence from municipal mayor elections (under revision). *University of Michigan*.