# Causal Inference for the Social Sciences

Pablo Geraldo Bastías

2025-01-01

Instructor: Pablo Geraldo Bastías (pablo.geraldo@nuffield.ox.ac.uk)

University of Oxford - Hilary 2025

Location: Seminar room, Department of Sociology

Meetings: Mondays, 3-5pm

## General

The course introduces students to contemporary frameworks of "counterfactual" causal inference, emphasizing the complementarities between potential outcomes (a.k.a., Neyman-Rubin causal model) and the structural causal model (a.k.a., Wright-Pearl graphical approach). The focus will be on identification and estimation of causal effects using observational data as encountered in social science applications.

## Topics

Topics covered in the class include an introduction to the experimental ideal, the use of potential outcomes to formalize the target quantity of an observational study and the assumptions needed for causal identification, the use of graphical models (Directed Acyclic Graphs) to discuss the plausibility of such identifying assumptions, and several "templates" for identification commonly used in empirical research (selection on observables through matching, weighting, regression, and recent machine learning approaches, instrumental variables, regression discontinuity, and difference-in-differences), as well as mediation and sensitivity analysis.

## Learning Outcomes

On successfully completing this course, students should have an understanding of the central role of causality in the social sciences and they should be able to cast a critical eye on the causal claims that social scientists make. Students should also have acquired a thorough knowledge of the potential outcomes and graphical approaches to causal inference, understanding the central role of untestable assumptions in identifying causal effects. Finally, students should be able to decide an appropriate analytical strategy to address their own research questions, and they should be able to estimate a wide range of empirical models in the context of causal inference applications.

## Content and Structure

The course focuses on the identification and estimation of causal effects, including the use of formal languages to define the target quantity ("estimand"), the conditions that should be meet for a causal interpretation to be valid ("identification assumptions"), and the evaluation of different algorithms ("estimators") that can be used to obtain answers from data given the assumptions.

Basic knowledge of probability and statistics, up to generalized linear models, is a prerequisite, but resources will be provided for those who need to refresh their statistics foundations. It is very likely that you will find the answer to your questions in the appropriate sections of the following books (but feel free to ask me if that's not the case):

Aronow, P. M., & Miller, B. T. (2019). Foundations of Agnostic Statistics. Cambridge University Press.

Blackwell, M. (2024). A User's Guide to Statistical Inference and Regression. Free online: https://mattblackwell.github.io/gov2002-book/

Shalizi, C. R. (2024). Advanced Data Analysis from an Elementary Point of View. Free online: https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf

## Weekly plan

We will broadly follow this plan, but there would be some adjustments if needed.

Week 1: Introduction to Potential Outcomes and the experimental ideal Week 2: Introduction to the Structural Causal Model and Directed Acyclic Graphs (DAGs) Week 3: Selection on observables I (matching, regression, weighting) Week 4: As-if-random assignment (instrumental variables and regression discontinuity) Week 5: Parallel trends and beyond (difference-in-differences and synthetic control) Week 6: Time-varying treatment and confounding, mediation analysis Week 7: Disparities, decompositions, and machine learning Week 8: Sensitivity analysis and external validity

## How to prepare for lectures

Each week, we will cover two related topics. While not attempting to be exhaustive, there is still a lot to learn, and attending lectures alone won't be enough. The role of the lecture is to provide a general panorama, and to build intuition to get you started and up to speed. You will encounter more technical material in the readings, and the problem sets will challenge you to apply the concepts and methods covered in class to real examples.

To get the most out of the class (to understand, to the point of being able to build and support causal claims on your own research), you will need to engage. Learning requires not only seeing, but doing. Both the readings (before and after the lecture), and the practical exercises, are essential components of the course. Asking questions (to yourself, to the readings, to your classmates, and to me) is one of the most important parts of active learning. So feel free to ask as many questions as you need, in person and online.

## Course Assessment

Problem sets will be assigned every other week, and you will get two weeks to work on them. Each assignment will include a combination of conceptual and theoretical questions, some practical questions regarding applications and empirical examples, and some coding exercises. All problem sets are equally weighted towards the final grade.

Collaboration is encouraged. Work with your classmates, discuss your answers, and help each other. However, it is required for everyone to produce their own answers, so do not copy verbatim from each other's answers (including code). And make sure to include in your response the list of people you collaborated with.

Problem sets can feel long and relatively challenging. Do not wait until the last minute to start them! The ideal would be to work on them through the two-week period as you attend the lectures and do the readings and tutorials. I recommend answering as much as you can during the time given, and submit whatever you have by the deadline. Try your best, but do not lose sleep over a problem set. If you do too little, I will let you know. But it is generally fine to submit an incomplete problem set.

Problem set 1: handed on January 20th, submit by January 31st Problem set 2: handed on February 3rd, submit by February 14th Problem set 3: handed on February 17th, submit by February 28th Problem set 4: handed on March 3rd, submit by March 14th

For each problem set, you will receive a pdf version and the source code as a .qmd file. You are expected to submit your answers both as a compiled pdf document and include your own .qmd source file (details below).

## Programming and software

During the term, students are expected to conduct data pre-processing steps, model estimation, and interpretation of results. This practical material will be covered through analyses and replications included in the problem sets, to be performed in R. You can find the R installation instructions for your system on the CRAN website. I highly recommend you to install RStudio as your development environment.

You will find many free tutorials and books online, targeted at various levels of experience. If you are not familiar with R, or need to improve your proficiency, I recommend the following resources:

Imai, K. (2018). Quantitative Social Science. An Introduction. Princeton University Press

Llaudet, E., and Imai, K. (2023). Data Analysis for Social Science. A Friendly and Practical Introduction, Princeton University Press (more beginner's friendly than the one above)

Wickham, H., Çentikaya-Rundel, M., and Grolemund, G. (2023). R for Data Science, 2nd Edition, O'Reilly Media, Free online: https://r4ds.hadley.nz/

If you are proficient in Stata and need to migrate to R, you may find some of these resources useful: see here and here.

Finally, Quarto. Quarto is a type of document where you can seamlessly write text, code, run your analyses, and get output (tables and plots) directly incorporated. Even more, from the same source code, you can render a pdf document or an html version (like a local website). It is therefore an amazing tool for research and data science, making your work well documented and reproducible. There is a Quarto chapter in the r4ds book (here), and the official documentation is very well explained. To get started go here: https://quarto.org/docs/get-started/hello/rstudio.html

For the problem sets, we will use Quarto. If you're familiar with Rmarkdown, it is pretty much the same, but better. If you are not, it's ok. You can probably get away by just modifying the template I will provide. Notice that Quarto also supports other programming languages, such as Python. You can use them for the problem sets, but no support will be provided.

## Using ChatGPT and other Generative AI tools

The University has adopted the Russell Group guidelines for the use of Generative AI in education [here]. Additional resources can be found here and here. Up to date policies on Plagiarism can be found here.

In general, using ChatGPT or other Generative AI tools in this class is discouraged but not forbidden, under certain limits. If you need help to clean your answers from grammatical and redaction errors, go ahead. If you need help debugging your own code, go ahead. But make sure to add a note indicating if and when you used the help of a Generative AI model,

including the type of model you used (ChatGPT, Claude, or any other), and the prompt you asked.

However, do not ask an LLM to do the work for you! The first and most important reason is that there is no point in taking this class if you're not doing the work yourself. Another important reason is that, believe it or not, these models do not understand causal inference very well. You may get a lot of ok-sounding answers that are fundamentally flawed (I have checked this many times myself!). The idea of taking this class is precisely being able to separate the wheat from the chaff in the applied causal inference literature, while Generative AI is a really powerful tool to obfuscate their difference. Therefore:

It is important to do the required readings, not just read an automated summary of them. Reading is learning!

It is important to write your own answers to the questions in the problem sets, not just write a prompt. Writing is thinking!

It is important to write your own code, and to check that it correctly implements your desired analysis. It's easier (and safer) to improve what you write yourself, than to fix AI-generated code!

## References

The complete list of readings for this class, organized by topics, can be found here. Here I only list some booksI have found particularly useful to improve my understanding of causal inference.

Angrist, Joshua and Jörn-Steffen Pischke (2009). Mostly Harmless Econometrics. Princeton University Press Chernozhukov, V. & Hansen, C. & Kallus, N. & Spindler, M. & Syrgkanis, V. (2024): Applied Causal Inference Powered by ML and AI; arXiv:2403.02467. Free online: https://CausalML-book.org

Hernan, M., and Robins, J. (2025). Causal Inference. What if. Chapman and Hall/CRC. Free online: https://miguelhernan.org/whatifbook

Huntington-Klein, N. (2022). The Effect: An Introduction to Research Design and Causality (1st edition). Chapman and Hall/CRC. Free online: https://theeffectbook.net/

Morgan, Stephen L. and Christopher Winship (2014). Counterfactuals and Causal Inference: Methods and Principles for Social Research (2nd edition), Cambridge University Press

Pearl, Judea, Madelyn Glymour and Nicholas P. Jewell. 2016. Causal Inference in Statistics. Wiley

Wager, S. (2024). Causal Inference. A Statistical Learning Approach. Free online: https://web.stanford.edu/~swager/causal_inf_book.pdf

## Acknowledgments

I have been influenced by a lot of people in my way of thinking about causal inference in general, and in preparing the materials for this class in particular. I first encountered causal inference more than a decade ago (!) during my master's in the Catholic University of Chile, in an eye-opening class taught by Luis Maldonado. At UCLA, I learn from (in alphabetical order) Onyi Arah Jennie Brand, Erin Hartman, Chad Hazlett, and Rodrigo Pinto. For this class, I am drawing heavily on Chad's class materials. Here at Oxford, both Richard Breen and Sascha Riaz have generously shared their own materials from previous iterations of this course they taught in Sociology and Political Sciences.