HDSR
HARVARD DATA SCIENCE REVIEW

Volume 0 Issue 0
DOI: 00.000
ISSN: 2644-2353

An Undergraduate Course in Causality

Lea Bottmer [†], Guido Imbens [†,*], Jason Weitze [†] Mary Wootters [†]
[†] Stanford University

Abstract. In the Fall quarter of 2024 we (a computer scientist and an economist as the faculty in charge of the course, with two economics graduate students as course assistants) taught an undergraduate course with the title "Causality, Decision Making, and Data Science," cross-listed in the Economics Department, the Data Science Major, the Computer Science Department and the Graduate School of Business undergraduate program. The course was primarily intended for freshmen and sophomores, but because it was the first time we offered it, we also admitted juniors and a few seniors. We restricted enrollment to forty students to make the course interactive. The course was case-based, with minimal statistics requirements. It was successful from our perspective, and student evaluations reflected a similarly positive view. We would like to share here some of what we learned. The materials we put together, including an extensive set of slides, problem sets, and data sets, are available on this website (https://stanford-causal-inference-class.github.io/).

Keywords: causality, decision science, correlation, regression

[*]imbens@stanford.edu

Causality, Decision Making, & Data Science

A Modular Course Template & Executive Summary

Stanford University // ECON 115 // Fall 2025

COURSE PHILOSOPHY

How do we know if a new policy, a business strategy or a medical treatment *actually* works? This course equips students with the foundational toolkit to move beyond correlation and rigorously estimate causal effects from data. We bridge econometrics, statistics, and machine learning to answer real-world questions.

LEARNING OBJECTIVES

- Distinguish causation from correlation.
- Implement and interpret regression models.
- Analyze random and quasi-experiments.
- Critically evaluate data-driven claims.
- Communicate findings effectively.

ASSESSMENT

60% 📘 Problem Sets (4x)
25% 🚀 Final Project
15% 🙋 Participation & Pre-Class Reading Reactions

PREREQUISITES

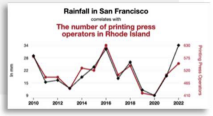
An *Intro to Statistics* course. Experience with programming will be helpful but is not required.

KEY TOOLS & RESOURCES

Python, Gradescope, Canvas. Optional textbook: *Mastering 'Metrics* by Angrist & Pischke

COURSE SCHEDULE & MODULES

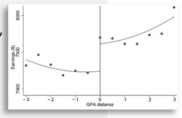
MODULE 0: Introductions and Statistics Refresher *[1 Week]*
What is this class about?
- Topics: Course overview, standard errors and confidence intervals, correlation vs. causality, potential outcomes and ATE
- Milestones: 📘 HW 0 due



MODULE 1: The Gold Standard *[2 Weeks]*
Randomized Experiments
- Topics: Randomized experiments, problems with experiments, adaptive experiments and multi-armed bandits
- Milestone: 📘 HW 1 due

MODULE 2: Universal Basic Income *[2 Weeks]*
What is the effect of cash transfers?
- Topics: Friedman's permanent income hypothesis, linear regression
- Milestones: 🚀 Project proposal due, 📘 HW 2 due

MODULE 3: Returns to Education *[2 Weeks]*
What are you (and everyone else) getting out of your education?
- Topics: Mincer model, twin studies, instrumental variables, local average treatment effects, neural networks and non-parametric models, regression discontinuity
- Milestone: 📘 HW 3 due



MODULE 4: Panel Data *[2 Weeks]*
How do we analyze data with multiple units over time?
- Topics: Difference-in-differences, matrix completion, synthetic controls, placebo methods
- Milestones: 🚀 Project check-in, 📘 HW 4 due

FINALE: Project Showcase *[1 Week]*
Presentations & Final Submissions
- Milestones: 🎤 In-class presentations, 🚀 Final report due

Figure 1. A one-page summary of the course as taught in Fall 2025 (including some updates compared to the Fall 2024 version).

1. The Origins of the Course and our Goals

This course came together out of a confluence of circumstances. One of us, Guido Imbens, had taught graduate courses in causal inference for many years. While there were no courses in this area when he arrived at Stanford in 2012, there are now multiple such graduate courses in different departments, including economics, political science, the graduate school of business, statistics, management science and engineering, and the medical school. He felt these topics should be of importance to undergraduates as well, and would be an improvement over the standard undergraduate econometrics courses. Another of the authors, Mary Wootters had taught a short course like this with Isaac Sorkin from the Stanford economics department while they both visited Istanbul during a sabbatical. In the Spring of 2024, we decided to combine forces and create a new course. It would be cross-listed in the Economics Department, the Data Science Major, the Computer Science Department, and the Graduate School of Business undergraduate program to reflect its interdisciplinary focus and the breadth of the teaching team.

Our main goal for the course was to get the students excited about and comfortable with using modern data science tools for decision making. We wanted to show how to do so in realistic settings, with real data and with ambiguity regarding the models and methods to use. We hoped that the students would see this as exciting and important material, and that some would consider majoring in data science or related fields. We also wanted to stress that problems are not bound by traditional disciplinary boundaries and that working across disciplines can improve insights, as well as be fun.

The combination of having an economist (or really, an econometrician) and a computer scientist coteaching the course made the experience more enjoyable and probably better for the students. We both attended all the classes (with some exceptions because of travel). It generated some discussions that highlighted differences in perspectives from our respective fields. Having a computer scientist and an econometrician as faculty was not important for the material. We both led classes in each of the modules, and for any of the specific lectures both of us could have taught the material. This partly reflected the expertise of both of the faculty, but more generally we are confident that many economists, computer scientists, statisticians, or other quantitative social scientists, could have taught courses very similar to this.

2. Course Design, Logistics, and Class Policies

We decided for the first version of the course to keep it small, with a cap of forty students, and tightly managed enrollment to ensure roughly equal representation from the four majors. As it was, the course was heavily oversubscribed. We selected the students by having them write a paragraph on why they were interested in the course. In the end, we had 1 freshman, 5 sophomores,

We aimed to make a course with minimal requirements so it could serve as an introduction to data science for computer science and social science majors. That was not entirely successful because many juniors and seniors with substantial backgrounds in statistics were very keen on taking the course and persuasive in their arguments to join the course. Our only formal prerequisite was a prior statistics course; although we did not require linear algebra, all students we accepted happened to have taken it beforehand.

Our main design choice was to make the course case-based. Rather than organize the course around the methods, we would focus on questions of substantive interest to policy makers and introduce methods around those applications. Some topics would center on a narrow set of methods, others would allow for the discussion of a range of methods, but they would all be introduced in the service of substantive questions and decisions faced by policymakers.

The substantive questions we focused on rarely had the black and white answers many of the students were accustomed to, causing them some unease. Seeing that given a particular substantive question the choice of statistical analysis was still open to discussion, and never entirely resolved, challenged the students. They were often comfortable with the technical level of the course, but participating in open-ended discussions around the questions and bringing in insights from multiple data analyses was new to them. By the end of the course, the students were getting comfortable differentiating between what they wanted to know and what they thought they were learning from the data. The data were providing an answer to a question, and they were getting more comfortable discussing how that may or may not inform the policy question they were interested in. It was rewarding to see this transition over the course of the quarter. As a plus, this focus made the course much more accessible to a freshman as the key lessons crucially did not require years of technical background. Here part of the motivation for this type of course for Imbens came from a course he took as a freshman at the Erasmus University in Rotterdam (The Netherlands). This course, taught by a lecturer from the Bureau of Economic Analysis in The Netherlands, used the book *Topics in Applied Econometrics* by Kenneth Wallis.

We decided on four modules of two weeks each, preceded by one week of statistics and followed by one week of student presentations. In the first module, students considered how we might evaluate the effects of different active labor market programs. Here, we highlighted the problems potentially plaguing observational comparisons and the benefits of randomized experiments. The second module focused on the effects of universal basic income and introduced regression as a tool for controlling for confounders. Building on these foundations, the third module concerned the returns to education, using this as an opportunity to introduce instrumental variables and regression discontinuity designs. Finally, the fourth module explored the effects of Brexit, while introducing panel data methods such as difference-in-differences, synthetic control, and matrix completion.

Each module consisted of four classes (with the exception of the universal basic income module, which consisted of three classes). The first class in each module typically introduced the substantive problem and why it was important to decision makers. The next three sessions would discuss the methods: This would involve describing them, applying them to the substantive problems (sometimes to different substantive problems), and discussing their shortcomings. Sometimes, there was also time to discuss additional methods related to the module's main methods.

Each module had an empirical problem set consisting of two parts, one for each week. The first part typically asked students to read in the data and present simple summary statistics to ensure they were prepared for the more substantive analyses and discussions in the second part. We did not require the students to have any prior programming experience, so the empirical problem sets contained a substantial amount of starter code.

The empirical problem sets were all group assignments, with the groups changing for every module and assigned by us rather than chosen by the students. This was intentional to create groups with students from different majors, backgrounds, and interests. However, this created some friction, with students preferring to work with their friends and sometimes having difficulty getting in touch with other students. The benefits we saw in the random groups were threefold: (*i*) creating class community, (*ii*) encouraging peer-to-peer teaching since we grouped them based on majors, and (*iii*) removing social barriers. We posted the group schedule for the rest of the quarter right after the students' course shopping period was over and we had a final list of students in the class. Note that we also assigned seating for the class, putting students in the same homework groups next to each other. We hoped this would help students to get to know each other in class before working on their homework together. We also had name cards for the students to make it

easier for us to remember the students' names given that they were sitting in different places every time.

Rather than a final exam, we had students complete a final project to not only assess their understanding but also to provide a final learning experience. Students could choose to work individually or in small teams and on either a theoretical or empirical project based on personal interest. Theoretical projects could probe topics underexplored in class, like multi-armed bandits, while empirical projects asked students to find data and attempt to answer a substantive question of interest. The students had to submit a project proposal to receive early feedback as well as a final presentation and report. Through this process, students, especially those working on empirical projects, confronted the inherent challenges of data-driven decision-making, cementing one of our core messages: causal inference is hard.

The day before each class, we asked the students to answer a question or submit a comment on a short reading. This would force the students to engage with the material prior to the class and helped us see where the students were coming from. The students appreciated that the assignments were not always to read a paper and instead often described a (realistic) scenario, prompting them to think through how they would approach the problem. Typically, the scenario and intended answer would be related to the method we would teach in the following class, though responses were not graded for accuracy. Our hope was for the students to gain intuition and start thinking about the methods we were going to discuss.

As we made clear in the syllabus and communication with students before enrollment, we required attendance in each class. To verify attendance, we used digital exit tickets at the end of each class. While this was an easy way for us to check attendance without physically noting it down, the exit tickets served another purpose as well, namely quick feedback and low pressure assessment. We designed them to be short, typically three questions, with the first one relating to the class content, the second one on students' comfort levels with the material, and the last one typically asking for feedback about classroom culture/lectures/problem sets. The class content questions encouraged students to reflect on the lessons and material learned rather than asking a question with a correct answer. For example, we would ask them to think about how what they learned today could be applied in real life or their own daily life or what surprised them most and why. Based on the feedback we received, we could address common confusions at the start of the next lecture and see what students might be looking for in a next iteration of this class. By requiring everyone to fill out the ticket, we got to hear a more representative sample of feedback. Some comments from the feedback were surprising relative to the questions we got in class. We did not allow the use of laptops in class, other than for students with accommodations. We did allow phones to be brought to class, partly to allow us to use polls and the QR code based exit tickets. For a future iteration of the course, we may switch to a classroom response system (clickers) to minimize students' engagement with their phones.

We had detailed PowerPoint slides for each class. These slides were listed as the most important resources by the students and are available on our websites. While students regularly asked for the slides to be made available prior to class, we decided against that, posting them only after each class. Doing so would have changed the nature of the classes, with the current version of the slides having discussion questions for the students. For students eager to delve deeper into the material, we recommended *Mastering 'Metrics* by Angrist and Pischke as an optional reference, though it was not required for the course. While there are many other textbooks on causal inference, many

of them are more appropriate for graduate courses.[1] We aimed for an interactive classroom setting rather than a simple lecture-style classroom. We made use of group discussion questions as well as PollEverywheres to get students engaged and interact with each other and the material. The students really appreciated the small breaks to digest the material and check their understanding with peers in real time. This is where our decision to have assigned seating next to the students in their problem set groups was important.

At the Graduate School of Business, one of the Stanford departments where the course was listed, faculty are discouraged from prohibiting the use of AI given the lack of enforceability of such policies. In this spirit, we told the students "we encourage you to use AI tools to *enhance* your own thinking, not to think for you." The combination with the stress on class discussions made feasible by the small class size made the challenges from AI use limited.

3. Module 0: Statistics Refresher

The first module was a week-long statistic refresher. This ranged from random variables, (conditional) expectations, and probability distributions to estimators, confidence intervals, and hypothesis testing. This module was designed to lay the foundation for the weeks to come. At the end of the module, we had students submit an individual homework assignment on the material. While we formally only graded the homework for completeness, we also used it as an opportunity to assess the students' background knowledge. An important challenge for the course was accounting for the variation in students' backgrounds, both in disciplines, but also in the level of mathematical sophistication.

While we briefly reviewed a number of basic concepts, we spent a significant amount of time on correlation and independence, as future modules on causality would rely on a firm understanding of these concepts. We began by unpacking the idea of correlation and examining a number of simple examples to help build intuition for what exactly it captures. We also worked through a few small examples by hand to clarify the mechanics of correlation. Afterward, we introduced independence, demonstrating through simple examples how it differed from correlation being equal to zero. At this stage, we refrained from drawing any comparisons with causality as we wanted to wait until we had carefully introduced the concept in the following module.

We also had an extended discussion of resampling methods when discussing uncertainty quantification (*e.g.*, standard errors and confidence intervals), because it was new to many of them. We emphasized bootstrap methods as a practical alternative to traditional analytic techniques, demonstrating how repeated sampling could provide robust, simple estimates of variability. This deeper dive into resampling not only reinforced students' understanding of statistical inference but also showed them how many practitioners in the 21st century quantify uncertainty. Moreover, this discussion laid the foundation for later discussions on randomization/permutation inference, which also leverages resampling.

During the first week, we also offered an optional Python Bootcamp outside the regular class hours to help prepare students for their homework. The intent was to onboard students who were either new to programming or simply new to Python. We started by showing students how to install Python and open a Jupyter Notebook. To make the rest of the session interactive and hands-on, we had the students download a Jupyter Notebook that would introduce them to Python fundamentals. This notebook intentionally had a similar structure to what their homework assignments would look like: a Jupyter Notebook with some text and designated cells for them to fill in answers. The

[1] This includes, among others, Ding (2024), Hernan and Robins (2020), Imbens and Rubin (2015), and Peters et al. (2017).

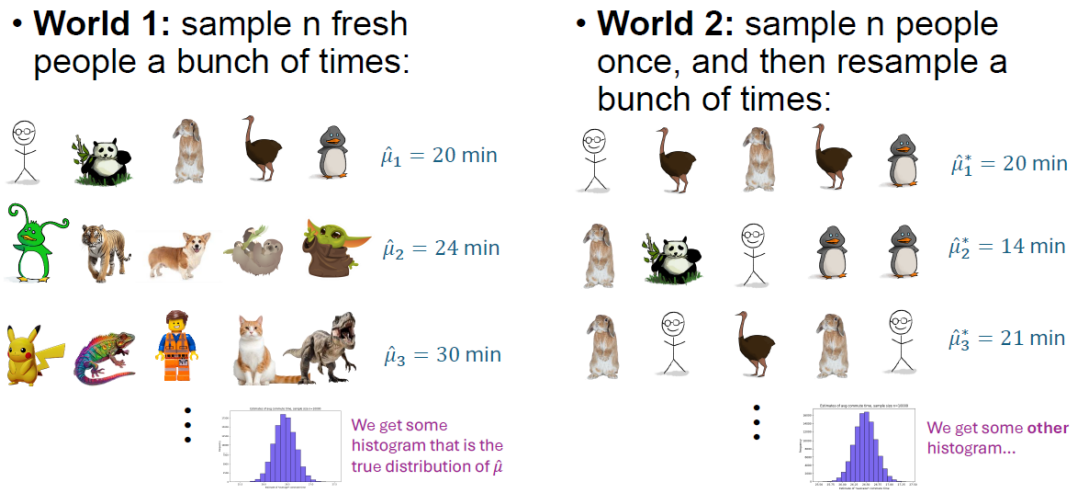


Figure 2. Thought experiment on resampling.

notebook started by introducing the basics, such as variables, data types, and control structures. Afterward, we showed them how to download, import, and use standard packages for data analysis, like NumPy and Pandas, as they would need to use them on most homework assignments. We also posted the instructions for setting up Python as well as the Jupyter Notebook so that students who could not make the Python Bootcamp still had an opportunity to get onboarded.

4. Module 1: An Experimental Evaluation of an Active Labor Market Program – Randomized Experiments

For the first module, the methodological focus was on randomized experiments. The substantive question was about the effect of active labor market programs. In the 1980s there were a number of high profile randomized experiments of such programs in the United States. We used data from one such program, the Greater Avenues to Independence (GAIN) program in California (Freedman et al., 1996; Hotz et al., 2006). Before going too deep into the application's details, we did one lecture introducing causality, taking a potential outcomes perspective (Imbens & Rubin, 2015; Neyman, 1923/1990; Rubin, 1974). Here, we discussed the difference between correlation and causality, stressing the old *adagium* that correlation does not imply causation.

In the second class, we discussed a paper that described an experiment to evaluate the effect of allowing laptops in the classroom (Carter et al., 2017). The setting for this paper was West Point, where during the 2014/2015 academic year for an introductory economics class students in some randomly selected sessions were allowed to bring laptops to class and use them, and students in other sessions were not allowed to use laptops. The results were quite striking, with the students in the sessions that were allowed laptops doing substantially worse on the common exam, by 0.18 of a standard deviation. This was the catalyst for a classroom discussion on external validity, and the relevance of these results from West Point from 2014/2015 for our decision not to allow laptops in the classroom, here at Stanford, in the Fall of 2024, for this specific class. The study was conducted in 2014/2015. Was the technology similar enough to make the study relevant for the current classroom? Is the population of students at West Point sufficiently similar to the population of students in our class that we should view these results as having bearing on our setting? Does the topic of the class matter? Maybe for a coding class having laptops in the class is more beneficial. Another interesting result in the paper was the heterogeneity in the causal effects. There was no evidence of any effect for women and the overall estimated effect was due to a substantial effect for

the men. There was also some evidence that the negative effect of allowing laptops was bigger for students with better prior academic records. This led to further discussion whether we could have a policy that allowed laptops for some students but not for others, and what characteristics should we be allowed to discriminate on. Note that we did allow exemptions to our no-laptop policy for students with special accommodations. One student brought up the issue of agency, suggesting that even if there were a negative effect on learning, we may wish to allow for the students to make their own decisions. None of these questions have clear answers, but in the end we, the faculty teaching the class, had made the decision, partly based on the evidence from the Carter et al., 2017 study, not to allow laptops.

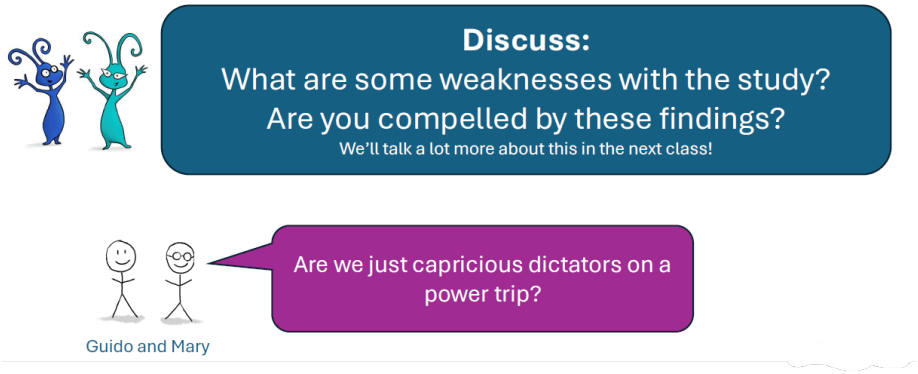


Figure 3. Classroom discussion prompt: "Should we allow laptops in the classroom?"

In the last class for this module, we discussed modern experimental designs in online experimentation. This covered many experimental designs in a single session. In each case, we tried to motivate the designs, and show some evidence of their value, without going into technical detail. This included multi-armed bandits, and experimental designs dealing with interference and spillovers. The students were very interested in these designs and they were the focus of a number of the final projects.

Figure 4. A visual for stratified experiments that we discussed as an extension of randomized control trials.

Homework. The accompanying homework for this first module was centered around analyzing a labor market experiment (Freedman et al., 1996). The first part of the homework asked the students to answer some pre-assignment questions about the context, then asked them to load in the data, and, lastly, provide some descriptive statistics. In the second part of the homework, we asked the students to estimate the average effect of the program using the difference-in-means estimate, and to provide standard errors through bootstrapping.

5. Module 2: The Effect of a Universal Basic Income and Friedman's Permanent Income Hypothesis – Regression

In this module, we investigated Milton Friedman's permanent income hypothesis (Friedman, 2018) and discussed the merits of a universal basic income that Friedman was a strong proponent of. Milton Friedman's Permanent Income Hypothesis states that consumption spending is based on a person's expected long-term typical/average income, the "permanent income" in the hypothesis, rather than their current, possibly transitory income. It implies that temporary income fluctuations have relatively limited effect on consumption. Understanding the effect of changes in permanent and transitory income is important for the design of the social security system, and it affects how we can most effectively help individuals suffering temporary setbacks in the labor market. Here, we focused specifically on the effect on labor supply of a change in permanent income. Friedman himself questions the implication of the irrelevance of transitory income: "... seems highly implausible. Will not a man who receives an unexpected windfall use at least some part of it in 'riotous living,' *i.e.* in consumption expenditures?" (Friedman, 2018, p. 28). The challenge in practice is to separate permanent and transitory components of income.

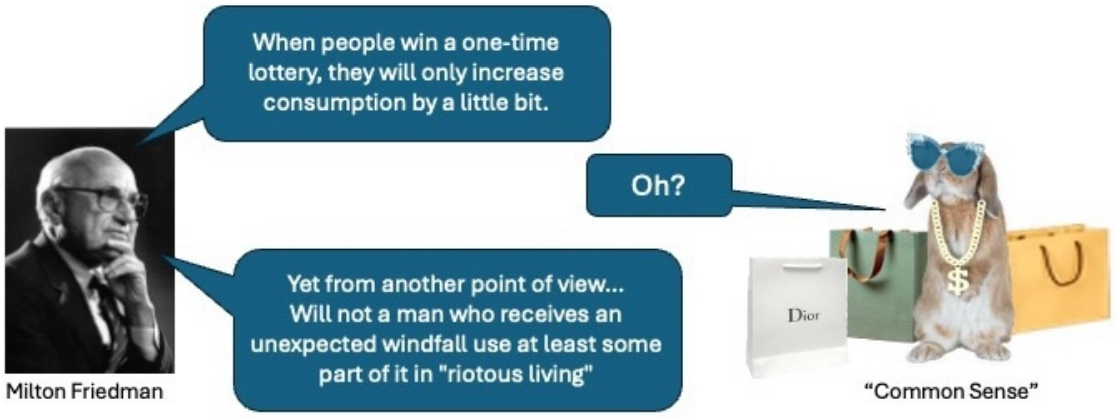


Figure 5. A visualization of Friedman's permanent income hypothesis.

There were two motivations for including this module. One was that the aforementioned class Imbens took as a freshman also included a module on Friedman's permanent income hypothesis. Second, a study Imbens had done surveying lottery winners and their economic behavior (Imbens et al., 2001). The data from this study were the basis for the problem set in this module. Moreover, a number of actual experiments on universal basic income have been conducted (Robins, 1985) and it has become a bigger topic of discussion in many policy circles.

We used the first lecture to introduce the new topic and, specifically, Friedman's permanent income hypothesis. We motivated the hypothesis by introducing students to two relatable hypotheticals – getting a $10,000 one-time payment *versus* getting a $10,000 payment every year for the rest of their lives – and then showing them how Friedman used economic theory regarding consumption and savings decisions to guide his hypothesis about people's spending behaviors for transitory and permanent income. We introduced heterogeneity by asking them to reflect how their parents would respond differently to the one-time and every-year payments of $10,000. We discussed some studies that had compared individuals receiving windfalls that were arguably somewhat close to random (Bodkin, 1959; Klein & Liviatan, 1957; Kreinin, 1961). However, individuals receiving

these windfalls were clearly not actually randomly selected, and there was evidence that there were other differences between those receiving windfalls and those not receiving windfalls.

In the second class, we introduced regression as a new data analysis tool for adjusting for such differences. We discussed the intuition behind linear regression and its interpretation. Moreover, we showed how adding variables to the specification of a regression function can remove some of the biases associated with the presence of confounders. We ended the class with a critical discussion of the regression results from three papers that tried to test Friedman's permanent income hypothesis with different data sets (Bodkin, 1959; Klein & Liviatan, 1957; Kreinin, 1961) that we had summarized and which they had been asked to read before class.

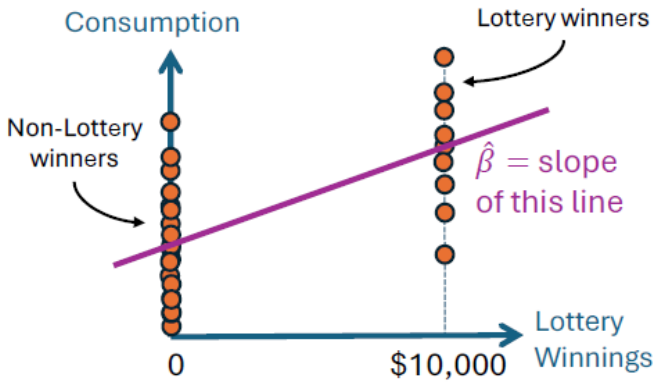


Figure 6. A visualization of a regression for lottery winners vs non-lottery winners.

In the last class, we related our unit back to current political discussions and three more recent papers centered around analyzing universal basic income experiments in the US and Finland (Bartik et al., 2024; Hämäläinen et al., 2019; Jaroszewicz et al., 2022; Vivalt et al., 2024). We analyzed the different studies using the five concepts we taught the students for evaluating the design of a study: (*i*) internal validity, (*ii*) external validity, (*iii*) construct validity, (*iv*) spillover effects, and (*v*) ethical concerns. This was the first unit where we started analyzing observational data and one lesson we wanted to get across in this unit was that being able to make causal claims is really hard.

Homework. The accompanying homework for module two analyzed the lottery data from the Imbens-Rubin-Sacerdote study (Imbens et al., 2001). In the first part of the homework, we asked students to answer some pre-assignment questions about the context as well as loading in the data and providing some descriptive statistics. The second part of this homework asked the students to implement a regression and interpret the results with respect to Friedman's permanent income hypothesis.

6. Module 3: The Returns to Education – Instrumental Variables, Regression Discontinuity

Our third module investigated the returns to education. This is a classic question in labor economics (Card, 1999). It is of course also very relevant to the students, particularly when thinking about why they are in college right now, and whether they want to pursue higher education.

In the first class, we started off again by introducing some economic theory about the returns to education, specifically the Mincer Model. In the Mincer model, individuals choose to acquire more formal education up to the point where the costs and benefits balance each other. What are the costs and benefits? Mincer models the benefits through the increase in wages people would

command with more education. The costs arise from the fact that getting more education requires foregoing working now, and so losing money one would otherwise have earned. The Mincer Model implies under some assumptions that log earnings should have an approximately linear relationship with years of education, thus leading naturally into regressions for the analysis. Moreover, the magnitude of the slope coefficient should be comparable to the interest rate. Here, economic theory is quite powerful, not simply suggesting a model to take to the data, but even suggesting values for the parameters. We discussed that many simple studies found returns to education on the order of 7-8%. We discussed concerns around omitted variable bias in our simple regression. We finished off the lecture by discussing the pros and cons of twin studies, one way to attempt to address omitted variable bias.

The second and third class of this module were reserved for discussions on methodology. We introduced the instrumental variable approach as a way to address the concern of omitted variable bias in settings where not all relevant confounders were observed. In the first lecture, we introduced the idea of Local Average Treatment Effects (LATE) using the Vietnam lottery draft example from J. D. Angrist (1990). We decided to use this clean binary case as an intuitive introduction to LATEs. Lastly, we touched upon how the LATE estimation is useful to answer policy relevant questions. For the third class, we extended our methodological approach to non-binary treatments/instruments, covariates and multiple instruments. We went back to our returns to education context and analyzed how J. D. Angrist and Krueger (1991) used the instrumental variables approach with quarters of birth as instruments to understand the returns to education.

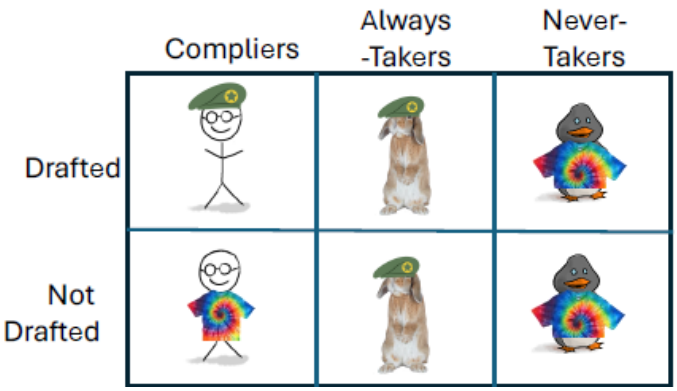


Figure 7. A visualization of different types of individuals in the LATE framework.

Since there are many different ways applied researchers have tried to understand the returns to education, we spent the fourth class in this module analyzing another new methodological approach: regression discontinuity designs. We used the example of SAT score cut-offs for college admission to motivate regression discontinuity designs (RDDs); we discussed its interpretation and limitations. We then introduced fuzzy regression discontinuity designs, motivated by returning to our substantive question: we want to know the effect of actually going to college, not of admission. We discussed two papers that use regression discontinuity designs to understand the returns to education (Hoekstra, 2009; Zimmerman, 2014). Lastly, we briefly introduced regression trees and random forests as a more complex and flexible way to understand returns to education, capturing non-linear relationships and heterogeneous treatment effects at the expense of interpretability.

Homework. In the problem set for the third module, we asked the students to replicate a study by Card (1993) that uses an instrumental variables approach – using geographical proximity to a

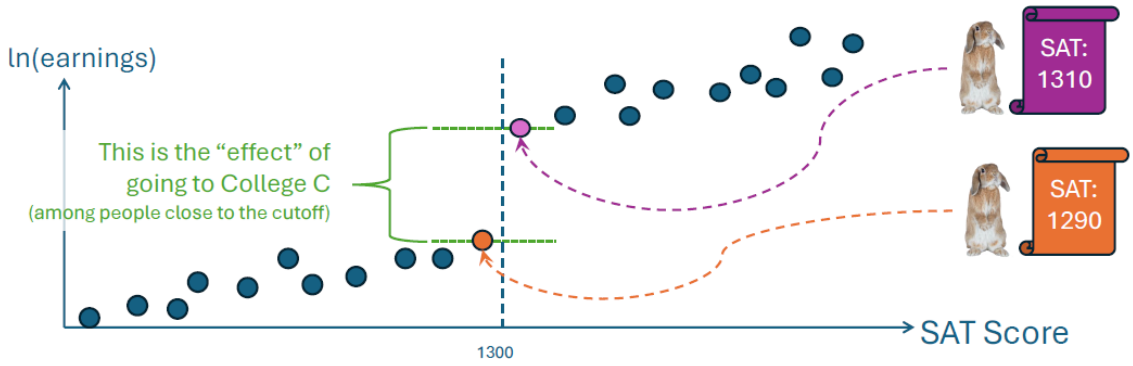


Figure 8. A visualization of RDD using SAT test scores as the running variable.

four-year college growing up as an instrument – to understand the returns to education. In the first part of the homework, students were asked to answer pre-assignment questions demonstrating their understanding of the context and limitations of the analysis. Then they had to load in the data set and report some descriptive statistics. In the second part of the homework, we guided them through a full analysis that started off with running a simple regression (no covariates), interpreting and discussing the results and whether we believe them, then a regression with covariates including interpretation and discussion, and, lastly, implement the IV estimator using indirect least squares (and optionally two-stage least squares). We concluded with asking them to discuss how their findings fit in with all the other estimates we have seen during our class discussions.

7. Module 4: The Effect of Brexit – Panel Data, Difference-in-Differences, Synthetic Control Methods, Matrix Completion

In the fourth module, the substantive question was about the causal effect of the UK leaving the European Union on UK Gross Domestic Product (GDP) per capita. The statistical problem involved panel/longitudinal data with observations on GDP per capita for the UK for a number of years both before and after Brexit, and observations on GDP per capita for other countries for the same years. Although one could bring to bear observations on additional variables, the focus here was primarily on using the $N \times T$ matrix of observations on GDP per capita for the N countries, including the UK, across T time periods. While students extensively studied the Brexit question on their homework, we used this module to motivate a more general paradigm and learn about different quantities we may be interested in estimating in our panel data context as well as several ways to estimate them. In the first class, we discussed the new data type, introduced eight empirical examples where researchers analyzed panel data and introduced the three main estimation strategies for the unit from the econometrics literature: difference-in-differences, matrix completion, and synthetic control.

In the second class, we focused on difference-in-differences (J. Angrist & Pischke, 2008) and introduced it using the famous minimum wage example by Card and Krueger (1994). We introduced the intuition behind the methods first in a simple two group, two period design and the assumptions it relies on. We gave a brief introduction into placebo analyses as a way to test our assumptions. Lastly, we discussed generalizations to the difference-in-differences strategy that include covariates and more time periods.

We used the third class to provide the students with some more fundamentals for introducing machine learning into economics. We started off with motivating feature selection using the LASSO

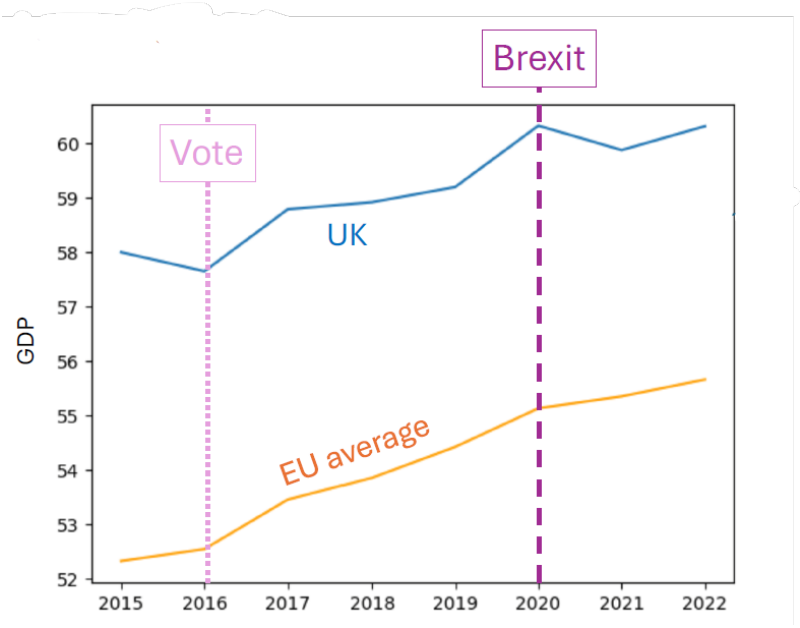


Figure 9. UK's GDP over time compared to the EU average.

before we dove into singular value decomposition (SVD) for our panel data settings as a generalization to feature selection on the outcome data directly. Lastly, we discussed how we can use these prediction tools to answer our causal questions of interest using matrix completion methods.

In our last class, we introduced synthetic control methods. We used a canonical example from Abadie et al. (2010), Proposition 99, a tobacco control program in California, as a running example to explain the intuition behind the synthetic control estimator (again, this is because we asked them to use the Brexit data on their homework). We also talked about placebo analyses in this context. We concluded the module by discussing how the three methods – difference-in-differences, matrix completion, and synthetic control– are related.

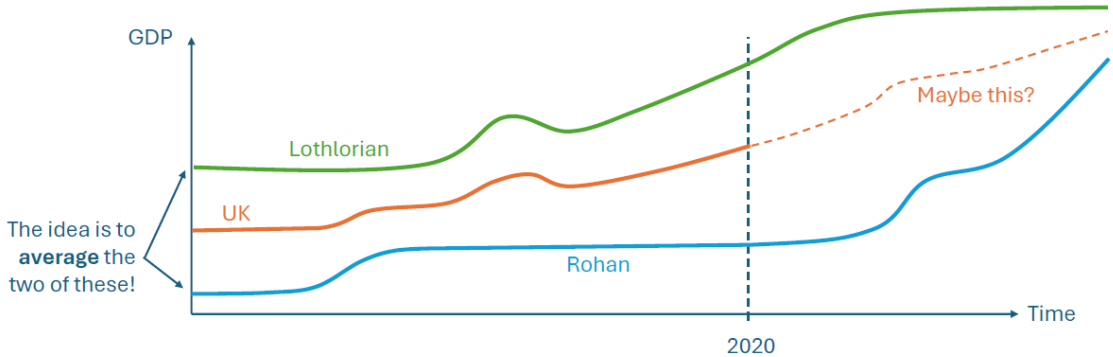


Figure 10. A visualization of the idea behind synthetic control using two imaginary places as control units.

Homework. The accompanying homework was all about analyzing Brexit using the methods we discussed in class. Since the students were supposed to work on their final projects at the same time, we aimed to make the workload for this homework substantially lighter and made some parts optional. As usual, in the first homework we simply asked the students to answer some pre-assignment questions about the setting, load in the data and do a preliminary analysis by computing

some group averages and time averages. In the second part of the homework, we asked students to implement the difference-in-differences estimator, conduct some placebo tests, and finally compute the matrix completion estimator and discuss any differences that arose. The optional part included applying the synthetic control estimator to the data.

8. Office Hours

We held 4 weekly office hours for the students. Since the class was rather small, these served as an effective opportunity for the students to ask a lot of in-depth questions about the class material, the problem sets, and their projects. Often, the questions would start out as problem-set specific questions and turn into bigger questions about the class material. The students highlighted two major advantages of office hours. First, they appreciated having the opportunity to get one-on-one explanations of formulas, concepts, and intuitions in greater depth than was possible in class. Second, they found that having access to someone with yet another perspective and way of explaining things was invaluable. While many students seemed very proficient in the math part of the course, many struggled with the more open-ended interpretation questions. The students had a few prototypical questions. One of the questions they would ask, especially early on before they got the gist of the class, was about how people in practice value different types of evidence when making decisions. At first a number of them were looking for black and white answers, right and wrong, and for many of them it was in office hours that they started to see it is all different shades of gray. Another common question was, 'why this way,' highlighting their curiosity. We also noted that office hours became a lot more attended the closer it got to the final project deadline. Some students would stick around for 30-45 min taking advantage of no one else in office hours to fully dive into their idea, the data they collected and the different identification strategies they were considering.

In the future, we may require every group to sign up for a 15-minute slot to get project feedback a few weeks before the due date. A few weeks out, most groups had gotten a good start on their project and had enough specific questions for this to be productive. Moreover, they would have had time to read through and start addressing the feedback on the proposals that we provided. Making this engagement with the course staff mandatory may have improved the project quality of the marginal group that did not choose to come to office hours of their own volition.

Many students suggested in the course evaluations to do the problem sets individually. This may partly reflect the unease of the students in being assigned to groups, rather than having groups that they could choose themselves.

9. Final Projects

In addition to the problem sets we had the students do final projects. Here we gave the students more leeway than with the problem sets. They could choose their own groups of up to three students, but some students chose to work on their own. In the end, we had seventeen projects. We also gave them a lot of freedom in the topics and the parameters of the project. For instance, their projects could be theoretical or empirical. The students first had to submit a proposal, which we read and approved, adding comments and suggestions as needed. During the last week of class, the students had to give an eight-minute presentation on their projects before submitting a final report a week later. We evaluated the presentations mainly on the students' ability to teach their peers something new, while the write-ups were primarily evaluated based on their content, quality, and clarity.

For a future iteration of the course, there are a few things we might change about the presentations. Through this experience many students learned how short 8 minutes really are. Part of the exercise was for students to practice distilling the core lessons/takeaways from their project and concisely communicating it, but a number of the groups tried unsuccessfully to communicate their entire project in eight minutes. In the future, we plan to give more guidance in the form of a recommended number of slides and emphasizing that sometimes less is more. Moreover,

in the spirit of the course, we randomized the order of the groups for the presentation schedule. However, we realized afterward that it may have been smarter to group similar topics together.

There was a wide range of topics, some of which we had covered in class, and some of which we had not. There were four presentations on adaptive experiments. There were seven empirical presentations using differences-in-differences and synthetic control methods. There were two projects on causal discovery, which we had not covered in class. We had shown some causal graphs, in particular for unconfoundedness and instrumental variables settings, but had not touched at all on causal discovery. There was one presentation focusing largely on regression adjustments. There was one project using regression discontinuity. There was also one project using causal forests to explore heterogeneous treatment effects.

The empirical projects explored a wide range of substantive topics, illustrating students' diverse interests. The topics included the effect of daylight savings time on ozone concentrations, the effect of black lives matter protests on crime, the effects of tax reforms in Colombia, the effects of inflation on political violence, the effects of having admissions be partly based on standardized tests on the diversity of the student body, and adaptive experimentation in the context of an online food delivery service.

A key insight students gained from doing empirical projects was the challenging nature of empirical work; they discovered that locating data is difficult and its analysis is complex. Many of them were concerned before presenting that they did not have statistically significant results, but while listening to other presentations, they realized that that was the norm rather than the exception. All groups had great discussions of their limitations and the shortcomings of their projects, emphasizing a main take-away we wanted to get across to the students: causal inference is hard.

The students doing empirical projects seemed to get a lot more out of the final project, and if doing it again, we would probably require the projects to be empirical. An important lesson from the class was that this is hard and you sometimes have to make decisions with imperfect data, and confronting that difficulty seems to have been valuable for the groups that tried empirical projects. Crucially, we were not grading them on finding the most credible causal effect, but rather if they could go through the motions, reason about the results, their credibility, and what they would do in the future if they had more data. In contrast, those doing the theoretical projects did not need to reckon with the inherent difficulty of empirical questions beyond what we guided them through in class and homework. Students may have picked the theory project because they appeared easier, and more familiar; you can open a textbook, learn a new concept or algorithm, and teach it to the next person. It is a good exercise, but there are also many courses out there teaching this kind of skill, and at least for the Stanford students we taught, it seemed too familiar. In contrast, the empirical projects really pushed students out of their comfort zone, to think critically and view things from a different light. If we chose to offer both project options again, we might make the empirical project easier (*e.g.,* by providing more data sets for them to use) to encourage more students to choose it.

10. Other Possible Modules

Student feedback on the course modules was generally positive, though there were several comments indicating that some lectures covered foundational concepts that many felt they had already mastered. Here, we discuss some additional modules that might fit in well with a longer or more advanced version of this course.

10.1. **Multi-armed Bandits.** Many final projects were about multi-armed bandits and their applications, demonstrating the students' high interest in this topic. In the course we had devoted one lecture to modern experimentation going beyond the treatment/control group set up. That lecture included a discussion on multi-armed bandits, but also discussed other experimental designs such as cross-over designs. Given the students' interest in this area, and the widespread use of modern experimental designs in online experiments, we would consider covering bandits in more depth in the next version of the class.

10.2. **Graphical Models.** In this course, we took as potential outcome perspective to causality (Imbens & Rubin, 2015). This was effective for the case studies and methodologies we wished to cover. However, there are other approaches to analyzing causality, with a prominent one based on graphical models (Pearl, 1995). One student presentation focused on graphical models, and it would be an interesting set of issues to discuss in the class.

10.3. **Random Forests.** In general, we would like to include more modern machine learning techniques into the class. One of these methods that was touched upon in the current version is random forests. These would naturally fit into the discussion on regression in the second module. We could also include some discussion in the first module to allow for more discussion on heterogeneous treatment effects. A more advanced module on estimating treatment effects under unconfoundedness would also have been a natural place to introduce methods that use the propensity score for estimating causal effects, and perhaps matching methods or even doubly robust methods. This would be a natural choice if the audience had a more advanced statistics background. Given our goal of a course that serves as an entry to data science, computer science, and statistics majors, we decided to avoid more advanced statistical techniques, and instead focus on conceptual insights related to causality.

10.4. **Regression Discontinuity Designs.** The discussion of Regression Discontinuity Designs was limited to a single class. These designs are very widely used in social science these days, and it would be easy to extend this to a full module with its own application.

10.5. **Sensitivity.** The arc of the course, with the initial module on experimentation, emphasized our view that experimental evidence has a special place in causal inference. In the other modules, we would again and again come back to concerns why the observational evidence might not be compelling, and how one should attempt to convey in a decision making process the uncertainty about the evidence beyond the statistical uncertainty captured in standard errors. Another way of doing so would be to have a module on sensitivity analyses and robustness. Some of this comes up in the placebo analyses, in particular in the regression discontinuity and synthetic control discussions, but sensitivity analyses have gone much beyond that.

11. An Evaluation and Thoughts for the Future

After the class, we had some discussions about the experience. What went well, what would we do differently, and did we achieve our learning objectives?

There are of course many ways to assess and measure this. In terms of the university run course evaluations the course was highly rated, substantially better than the average course in the school. Individual student comments also reflected their general satisfaction with the course, with one student writing "One of the best (if not the best) class I've taken at Stanford." Two students subsequently enrolled in a graduate class taught by one of us. There was clear appreciation for the connection between the technical material and the real-world case studies that students felt was missing in some other more methodological courses.

From our side the experience was similarly positive, and we plan to teach the class again in the Fall of 2025. We view the decision to not allow laptops in the classroom as a good one, and we will continue doing that. We may go further and also ask students to keep their phones in their bags, and use other methods for giving in-class feedback. The in-class discussions worked well, and we will put more structure on them to make them even more effective.

The range of backgrounds, with sophomores, juniors, and seniors, created some tension that was evident in the course evaluations. Some students felt the material and problem sets were not challenging enough and wanted to see more mathematical detail. However, part of this seems to be coming from students looking for a familiar form of difficulty, mathematics, rather than recognizing the difficulty inherent to the subjective part of causal inference. Even students well-versed in specific methodologies deepened their understanding by applying these approaches to ambiguous real-world scenarios, reinforcing that methodological knowledge alone is insufficient. To better align students' expectations with our learning goals, in future iterations, we plan to further emphasize this as one of our key learning goals at the start of the quarter.

Similarly, computer science majors expressed a desire for more challenging computational exercises. While coding proficiency was not a prerequisite, most enrolled students demonstrated coding competence—again likely reflecting the upper-division demographic. To address this in future offerings, we could provide tiered coding assignments with varying levels of scaffolding, allowing students to self-select based on their comfort level.

While AI use was not a major problem in the class in its initial form, it may become an issue in the future. Moreover, we did not attempt to harness the benefits of AI use in any systematic way. This is an interesting challenge for future versions of the course.

Disclosure Statement. The authors have no conflicts of interest to declare.

Acknowledgments. We thank Xiao-Li Meng for encouraging us to write this article, and two anonymous reviewers for comments. We particularly would like to thank the students in our class for making it a joy to teach.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, *105*(490), 493–505.

Angrist, J., & Pischke, S. (2008). *Mostly harmless econometrics: An empiricists' companion*. Princeton University Press.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, 313–336.

Angrist, J. D., & Krueger, A. (1991). Does compulsory schooling affect schooling and earnings. *Quarterly Journal of Economics*, *106*(4), 979–1014.

Bartik, A. W., Rhodes, E., Broockman, D. E., Krause, P. K., Miller, S., & Vivalt, E. (2024). *The impact of unconditional cash transfers on consumption and household balance sheets: Experimental evidence from two us states* (tech. rep.). National Bureau of Economic Research.

Bodkin, R. (1959). Windfall income and consumption. *American Economic Review*, *49*(4), 602–614.

Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (tech. rep.). National Bureau of Economic Research.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, *3*, 1801–1863.

Card, D., & Krueger, A. (1994). Minimum wages and employment: Case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, *84*(4), 772–793.

Carter, S. P., Greenberg, K., & Walker, M. S. (2017). The impact of computer usage on academic performance: Evidence from a randomized trial at the united states military academy. *Economics of Education Review*, *56*, 118–132.

Ding, P. (2024). *A first course in causal inference*. Chapman; Hall/CRC.

Freedman, S., Friedlander, D., Lin, W., & Schweder, A. (1996). The gain evaluation: Five-year impacts on employment, earnings, and afdc receipt. *New York: MDRC*.

Friedman, M. (2018). Theory of the consumption function.

Hämäläinen, K., Kanninen, O., & Verho, J. (2019). Evaluation of the finnish basic income experiment: First year results.

Hernan, M. A., & Robins, J. M. (2020). *Causal inference*. CRC Boca Raton, FL:

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The review of economics and statistics*, *91*(4), 717–724.

Hotz, V. J., Imbens, G. W., & Klerman, J. A. (2006). Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the california gain program. *Journal of Labor Economics*, *24*(3), 521–566.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W., Rubin, D. B., & Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 778–794.

Jaroszewicz, A., Jachimowicz, J., Hauser, O., & Jamison, J. (2022). *How effective is (more) money? randomizing unconditional cash transfer amounts in the us*. SSRN.

Klein, L. R., & Liviatan, N. (1957). The significance of income variability on savings behaviour. *Bulletin of the Oxford University Institute of Economics & Statistics*, *19*(2), 151–160.

Kreinin, M. E. (1961). Windfall income and consumption: Additional evidence. *The American Economic Review*, *51*(3), 388–390.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465–472.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT press.

Robins, P. K. (1985). A comparison of the labor supply findings from the four negative income tax experiments. *Journal of human Resources*, 567–582.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Vivalt, E., Rhodes, E., Bartik, A. W., Broockman, D. E., & Miller, S. (2024). *The employment effects of a guaranteed income: Experimental evidence from two us states* (tech. rep.). National Bureau of Economic Research.

Zimmerman, S. D. (2014). The returns to college admission for academically marginal students. *Journal of Labor Economics*, *32*(4), 711–754.