*Ivan A. Canay*

# *Econometrics for MMSS II*

| ***Econ 386-2 :: Ver. May 19, 2026***

# Contents

# *Foreword*

These lecture notes were prepared for Math 386-2 at Northwestern University, the third quarter of the econometrics sequence in the MMSS program. The class assumes that students have completed Math 385 and 386-1, which cover the fundamentals of probability, statistics, and predictive methods such as regression. The focus of this course is not on technical details but rather on developing a solid understanding of when we can assign causal interpretations to commonly used parameters that applied researchers seek to estimate.

There are many resources available online and in print that cover these topics in greater detail. Each chapter of these notes highlights the key references that influenced the ideas presented. So why write these notes? The primary goal is to provide a coherent framework for the concepts I plan to cover in the class, using consistent notation. The secondary goal is to strike a balance suitable for MMSS students—advanced beyond most undergraduate courses, yet not quite at the level of graduate classes. Personally, I do not believe these notes would be particularly useful outside the context of Math 386-2, nor are they intended to be a comprehensive treatment of any of the topics. If you are not an MMSS student, you are welcome to use them at your discretion, but I strongly encourage readers to explore the additional resources referenced in each chapter for a deeper understanding.

I would like to express my gratitude to several friends and colleagues who made it easier for me to write these notes. First and foremost, I want to thank Stefan Wager and Peng Ding for their excellent notes on many of the topics covered here, which they have generously made publicly available. I also want to extend my gratitude to those who were kind enough to share their own notes with me, even when such notes are not publicly available. These individuals include Alex Torgovitsky for his outstanding notes on IV, Clément de Chaisemartin for sharing his work in progress on Difference in Differences, and Matias Cattaneo for his teaching material on RDD. Special thanks go out to my friend and long-time collaborator, Azeem Shaikh, whose influence on my way of thinking about econometrics makes it challenging for me to distinguish how much of what is written in these notes is novel and how much is derived from our conversations, scribbles, and the notes he has shared with me over our 15+ years of collaboration. Finally, I want to express my sincere appreciation to Sebastian Poblete-Coddou, whose contributions as my research assistant played a key role in developing and revising these notes.

# Part I

# Causality and Conditional Independence

# 1

# *Causality and Potential Outcomes*

This class has two objectives. One is to build intuition for you to assess questions that are *causal* in nature. The other is to introduce a simple formal framework—potential outcomes—that lets us state causal questions and causal assumptions precisely. Causal in the context of our class means *cause–effect* relationships that may be quantified using statistics and intuitions from other disciplines. We want to measure effects of an event $A$ in an outcome $Y$. Examples of these kind of questions are: What is the effect of public expenditure in hospitals on patient health? What is the effect of minimum wage on employment? What is the effect of climate change on migration? Below, we will see why we require a whole class (and why there is a whole literature and sub-field of study) about addressing causal questions. Our goal is to make sure that you can build your causal intuition, and along with the mathematical notation introduced here, help you to understand *when* a certain parameter or statistic may be interpreted causally or not. This notation, which will be used throughout the class, is necessary to (i) precisely define causal concepts, and (ii) to build the statistical tools necessary to answer causal questions.

## 1.1 Causality, Counterfactuals, and What-if

The "fundamental problem" in causal inference is that we *don't observe* counterfactual states: If we want to perfectly measure the effect of a drug in a patient health outcome, ideally we would like to have two clones, equal in every way, and give the drug to one clone, and none (or a placebo) to the other clone. Then, we would measure the difference between the clones of the health outcome. We would call this quantity the *causal effect* of the drug, since there is no other factor that could have changed the health outcome other than the drug. Besides the bioethical implications of cloning, you can see that this ideal scientific and statistical scenario is rather scarce or impossible. Therefore, we need a conceptual and statistical framework to address this fundamental problem, and actually be able to measure causal effects of interest.

In this class we will use the so-called potential outcomes framework (Neyman, 1923; Rubin, 1974) to mathematically represent counterfactual states of

the world. In this framework, an experiment, or at least a thought experiment, has an intervention, a manipulation, or a treatment, and we are interested in its effect on an outcome or multiple outcomes. Let's start with some examples of real counterfactual questions that appear in different areas of study:

- What is the effect of minimum wage in employment? [Labor]
- What would happen to prices/welfare if two firms merged? [IO]
- What effect would this medication have on heart disease? [Biostatistics]
- What will happen to global temps if emissions decrease? [Climatology]

In 2013, American Airlines and US Airways decided to merge. The effect of mergers in prices can be ambiguous: They can *go up* if the merger increases industry concentration in a way that makes a market less competitive, or can *bring it down* if the efficiency gains of the merger are such that it actually makes the market more competitive. In 2020, a year after the merger, the price of the product sold by the new merged company (American Airlines Group) went down [Das, 2019]. Imagine that we can somehow know — perhaps by divine revelation — that had these two firms stayed as separate firms, prices a year later would have remained unchanged. Equipped with this information most would agree that it was the merger that caused the decrease in prices, probably because of an efficiency gain.

Consider two other firms, C and D, that also decide to merge in 2019. A year after the merger, the price of the product sold by the new company goes down relative to the previous year. Imagine that we can somehow know that, in the absence of the merger, the prices charged by firms C and D would have gone down by exactly the same amount a year later. Hence the merger did not have a causal effect on prices in the following year; it probably was just one firm getting more efficient.

These two vignettes illustrate how humans reason about causal effects: we compare the outcome when an action $A$ is taken versus the outcome when the action $A$ is withheld. If the two outcomes differ, we say that the action $A$ has a causal effect on the outcome. Otherwise, we say that the action $A$ has no causal effect on the outcome. Epidemiologists, statisticians, economists, and other social scientists refer to the action $A$ as an intervention, an exposure, a policy, or **a treatment**.

## 1.2 Potential Outcomes

Now let's introduce the mathematical notation of potential outcomes, which is probably the easiest way to think about causal relationships in this class. As a simple illustration, consider again our drug example but in a more realistic

setting: think about a randomized controlled experiment where individuals are randomly assigned to a treatment (a drug) that is intended to improve their health status. Let $Y$ denote the observed health outcome and $A \in \{0, 1\}$ denote whether the individual takes the drug or not. The causal relationship between $A$ and $Y$ can be described using the so-called *potential outcomes*:

| | |
|---|---|
| $Y(0)$ | potential outcome in the absence of treatment |
| $Y(1)$ | potential outcome in the presence of treatment |

In other words, we imagine two potential health status variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 1. Recalling our "cloning" example, $Y(0)$ would be the health status or outcome for the clone without the drug, and $Y(1)$ the outcome of the clone with the drug. As mentioned earlier, in reality, we *only observe one* of the two potential outcomes, the other being the *counterfactual* outcome.

The difference

$$\Delta := Y(1) - Y(0) \tag{1.1}$$

is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*.

**Remark 1.1** For now, we will focus on population parameters, so we will not discuss the availability of a random sample or the indexing of random variables by subjects, units, or individuals (i.e., by $i$). Later in the course, we will address these topics and introduce the notation $Y_i(0)$ and $Y_i(1)$ for the potential outcomes of the $i$-th individual. At that point, the treatment effect in (1.1) will be indexed by $\Delta_i$, reflecting the fact that the treatment effect is unit-specific and may be heterogeneous across individuals. ■

Neyman [1923] first used potential outcomes notation and, as simple and intuitive as it appears to be, it has some hidden assumptions. Rubin (1980) clarified the implicit assumptions and called them no-interference and consistency.

The first assumption, *no-interference*, states that the potential outcomes of a unit or individual do not depend on other units' treatments. This is often a reasonable assumption in medicine (i.e., that the treatment prescribed to patient 1 doesn't affect patient 2), but may be less appropriate in social sciences where *network effects* may arise. For now, however, we will ignore network effects and assume **no-interference**.

The second assumption, *consistency*, states that there are no alternative versions of the treatment; that is, the treatment must be well defined and unambiguous for the outcome of interest. For example, in studying the effect of cigarette smoking on lung cancer, differences in cigarette types might matter, and in examining the effect of college education on income, variations in institution type or major may be relevant. Consistency effectively rules out

these variations. While one could redefine each version of the treatment as a separate intervention, this is not always feasible or desirable — especially if data on the specific treatment version is unavailable.

The two assumptions combined imply that the observed outcome $Y$ is equal to the potential outcome $Y(A)$ for the treatment level $A \in \{0, 1\}$ that was *actually* observed,

$$Y = Y(A) = Y(1)A + Y(0)(1 - A) \ .$$

We can re-write this equation as

$$Y = Y(0) + \Delta A \ ,$$

which shows that the observed outcome is just the baseline potential outcome $Y(0)$ plus the treatment effect times the treatment level. Importantly, all of these objects, $Y(0)$, $A$ and $\Delta$, are **random**. In particular, since $\Delta$ is random, it has a distribution, and we can talk about its mean, variance, and other properties. In other words, the treatment effects can be summarized in many ways and in turn lead to many possible *parameters of interest*.

**Example 1.1 (Program Evaluation)** Suppose that $A \in \{0, 1\}$ indicates participation in a job training program and that $Y$ is a scalar labor market outcome such as earnings. If $A = 1$ we observe $Y(1)$ - but not $Y(0)$ - and if $A = 0$ we observe $Y(0)$. There are many possible questions one could ask (and therefore, *parameters of interest* to analyze), such as:

- What would be average earnings if everyone were trained, i.e. $E[Y(1)]$?
- What is the average effect of the program, i.e. $E[Y(1) - Y(0)]$?
- What about only for those who are trained, i.e. $E[Y(1) - Y(0)|A = 1]$?

■

### 1.2.1 About Identification of the Average Treatment Effect

Our goal in this class will be to credibly identify and estimate features of the distribution of the treatment effect $\Delta$. We will devote particular attention to the **average treatment effect** (ATE), defined as

$$\theta := E[\Delta] = E[Y(1) - Y(0)] \ ,$$

due to its prevalence in empirical work. The main barrier to credibly identify features of the distribution of the treatment effect $\Delta$, such as the ATE, is that only one treatment can be assigned to a given individual, and so only one of $Y(0)$ and $Y(1)$ can ever be observed. In other words, the treatment effect $\Delta$ is *never* observed and so we need to find ways to deal with this missing data problem. What we observe is the outcome $Y$ for a given treatment level $A$, and so the **problem of identification** is to be able to characterize the true $\theta$ as a function of the observed data: observed outcomes and treatment assignment $(Y, A)$.

## 1.3 General Setting

The potential outcomes framework is a way to formalize the notion of counterfactuals. While for the vast majority of our exposition we will focus on the case where the treatment variable $A$ is binary, it is useful to think about the case where $A$ can take on more than two values. For instance, in the case of the effect of education on income, $A$ could be the number of years of education.

To keep notation simple in this chapter, we will focus on the case where $A$ is discrete, with a finite or countable support $\mathcal{A}$ (e.g., $\mathcal{A} = \{0, 1\}$ or $\mathcal{A} = \{0, 1, \ldots, 25\}$). For each $a \in \mathcal{A}$, define the potential outcome

$$Y(a) = \text{ the outcome that would have been observed if } A = a \ .$$

The treatment effect could be defined relative to any baseline $a_0 \in \mathcal{A}$ as $\Delta(a) = Y(a) - Y(a_0)$, or between two different values of $a$ as $\Delta(a, a') = Y(a) - Y(a')$. In this setting, the observed outcome $Y$ is equal to

$$Y = \sum_{a \in \mathcal{A}} Y(a) I\{A = a\} = Y(A) \ ,$$

where $I\{A = a\}$ is an indicator function that equals 1 if $A = a$ and 0 otherwise. Note that $Y = Y(A)$ is observed, but the potential outcomes $Y(a)$ for $a \neq A$ are always **unobserved**. This problem is illustrated in Figure 1.1. In an ideal scenario, we would observe the potential outcomes $Y(a)$ for different values of $a$ for the same individual. However, in practice, we can observe only one outcome, $Y(A)$ , for each person.

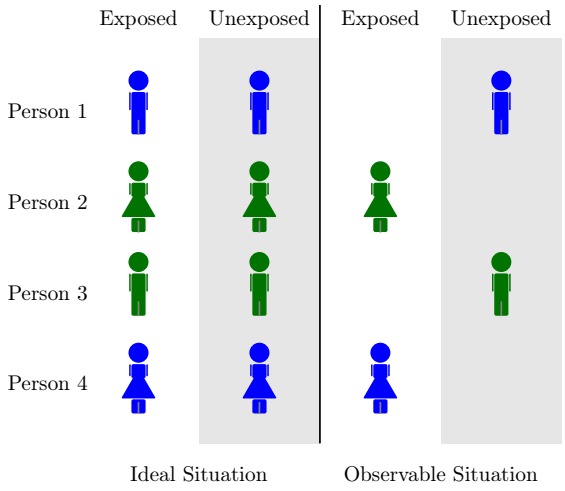



FIGURE 1.1: Illustration of the counterfactual problem

Before we move on to discuss identification of the ATE in perhaps the simplest possible setting, we should mention that potential outcomes are not the only way to formalize the notion of counterfactuals. There are at least two other ways that receive a lot of attention across disciplines: structural models (or latent variable models) and directed acyclic graphs (DAGs). Latent variable models generally refer to models where the outcome is a function of an unobserved variable and some other variables. For instance, in the case of the effect of education $A$ on income $Y$, the latent variable could be IQ (denoted by $U$), and be related to the outcome by the relationship

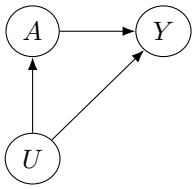
$$Y = g(A, U) \ ,$$

for some well defined function $g(\cdot)$. A **causal** interpretation of this model is implicitly saying:

$$Y(a) = g(a, U) \text{ for every } a \in \mathcal{A}$$

and it could impose assumptions depending on what $g$ and $U$ are (and, more importantly, how $U$ is assumed to be related to $A$).

DAGs are a graphical representation of a set of assumptions about the causal structure of the data generating process. For example, in the case of the effect of education $A$ on income $Y$, we could have the following DAG




where an arrow indicates a direct causal effect. Here $U$ is an (unobserved) factor such as "ability" that can affect both education $A$ and income $Y$. The arrow from $A$ to $Y$ represents the causal effect we care about, while the arrows out of $U$ remind us that differences in outcomes across people with different $A$ may reflect differences in $U$ as well. No arrow means no direct causal effect. The variables can still be related indirectly, or look related because of other variables.

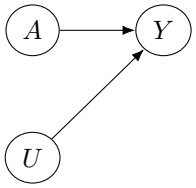
Some are dogmatic about the use of, say, potential outcomes versus latent variable models; this often follows field-specific conventions (e.g., labor versus IO). We will not discuss these alternative approaches in detail in this class, but you should be aware of them and keep in mind that this is just *notation* - and that you can use the above to translate.

## 1.4 About Experiments

Imagine we are comparing incomes between individuals who grew up in high-income versus low-income neighborhoods. While we might expect higher salaries for those that grew up in the high-income neighborhood, there could be many factors at play. The neighborhood itself may not be the sole explanation for the observed differences. To start, wealthier families could shape the expectations of their young residents. These individuals may perceive dropping out of school as forfeiting a professional future — something that might be seen as unattainable in a lower-income neighborhood. Additionally, families in wealthier neighborhoods are more likely to afford better schooling and higher education, which could contribute to the observed income disparities.

Note that many of the factors mentioned above are hard to measure, and are likely *correlated* with neighborhood choices: families do not pick neighborhoods at random. Using the notation we just reviewed, the neighborhood would represent the "treatment" $A$ , while other background characteristics (family resources, expectations, etc.) would be denoted by $U$.

One way to address this issue is through **experiments**: we can randomly assign the treatment $A$ (e.g., assign some families access to a higher-income neighborhood and others not). In that case, $A$ is unrelated to $U$. The corresponding DAG looks like:




In words: because $A$ is randomly assigned, it is not related to $U$. This is why experiments help: differences in outcomes across groups reflect the causal effect of $A$ rather than pre-existing differences in $U$. Of course, for many treatments we cannot literally control the choice (e.g., where a family lives), but we can sometimes still *influence* $A$ in a random way, as illustrated in the next example (MTO).

**Example 1.2 (Moving to Opportunity experiment)** Between 1994 and 1998, the US Department of Housing and Urban Development (HUD) enrolled 4,604 low-income families living in five US cities (Baltimore, Boston, Chicago, Los Angeles, and New York) in the Moving To Opportunity (MTO) project. Families were randomly assigned into three groups: (i) the experimental group, which received housing vouchers that subsidized private market rents and could initially (for the first year) only be used in census tracts with 1990 poverty rates below 10 percent; (ii) the Section 8 group, which received regular

housing vouchers without any MTO-specific relocation constraint; and (iii) a control group, which received no assistance through MTO. Chetty et al. [2016] evaluated the impact of the program on children, using the random voucher assignment to generate differences in neighborhood access.

The authors found an average effect of $3,477 for those who received the experimental voucher, a 31 percent increase relative to the control group mean of $11,270. This income boost led to several positive outcomes for those impacted by the intervention. The key point is that, because assignment to the voucher groups was random, the groups should be similar in background characteristics (like prior wealth or expectations) on average. Therefore, differences in average outcomes across groups can be interpreted causally. ■

## 1.5 Key Concepts

- **Potential Outcomes:** For each unit and each treatment level $a \in \mathcal{A}$, $Y(a)$ denotes the outcome that would be realized for that unit if (possibly contrary to fact) the treatment were set to $a$.
- **Counterfactual:** Given the realized treatment $A$, the counterfactual outcomes are $\{Y(a) : a \in \mathcal{A}, a \neq A\}$; these are not observed for that unit.
- **Identification of the ATE:** Identification of the ATE means showing that $\theta = E[Y(1) - Y(0)]$ is uniquely determined by the joint distribution of observables (e.g., $(Y, A)$), given maintained assumptions.

## 1.6 Concluding Remarks

The contents of this chapter are based on the notes by Peng Ding [2023], the notes by Stefan Wager [2020], the book by Hernán and Robins [2025], and notes shared by Alex Torgovitsky. The book by Angrist and Pischke [2008] is another reference that includes discussion on many of the topics we covered.

## 1.7 Problems

**Problem 1.1** *In the context of $A/B$ testing, suppose a company is testing two versions of a webpage: version $A$ with a red "Buy" button and version $B$ with a green "Buy" button. The outcome $Y$ is whether or not a user clicks the "Buy" button, with $A = 1$ representing the red button and $A = 0$ representing the green button.*

1. *What does $E[Y(1)]$ represent in this $A/B$ testing scenario?*
2. *Why is $E[Y(1)]$ important for evaluating the effectiveness of the red "Buy" button in comparison to the green one?*
3. *What type of counterfactual "policy" would the ATE capture in this scenario?*

**Problem 1.2** *Consider a field experiment studying the effect of offering cash incentives on voter turnout. In this experiment, some individuals are randomly selected to receive cash incentives ($A = 1$) and others do not ($A = 0$). The outcome of interest is $Y$, which is voter turnout (1 for turnout, 0 for non-turnout).*

1. *What does $E[Y(1) - Y(0)]$ measure in this setting?*
2. *Why is it difficult to directly observe $Y(1) - Y(0)$ for everyone in the sample?*
3. *What does $E[Y(1) - Y(0) \mid A = 0]$ measure in this setting?*

**Problem 1.3** *Consider a clinical trial testing the effectiveness of a new drug. Let $A \in \{0, 1\}$ indicate whether a patient receives the drug ($A = 1$) or a placebo ($A = 0$), and let $Y$ be a binary outcome representing whether the patient experiences a negative side effect (1 for a side effect, 0 for no side effect).*

1. *What does $E[Y(0)]$ represent in this clinical trial context?*
2. *Suppose that the placebo drug has no side effect, what would $E[Y(1)|A = 1]$ measure in this case ?*
3. *What would be the probability that individuals would experience a side effect if everyone were treated?*

**Problem 1.4** *Suppose that $A \perp\!\!\!\perp Y(a)$ for $a \in \{0, 1\}$. Show that*

$$E[Y \mid A = a] = E[Y(a)] \ .$$

**Problem 1.5** *Suppose a firm introduces a job training program ($A_i \in \{0,1\}$) for some employees in a given workplace. Employees work in teams, and workers who are trained often share the tips they learned in the training with their teammates.*

1. *Using the concepts from the chapter, explain in words which assumption(s) would this situation violate (no-interference, consistency, or both).*
2. *Formalize the violation by writing (using the notation introduced in the chapter) how a worker $i$'s potential outcomes could depend on other workers' treatments.*

**Problem 1.6** *Let $A \in \mathcal{A} = \{0,1,2\}$ denote "no training", "short training", and "long training", and let $Y(a)$ be earnings under treatment level $a$.*

1. *Give two examples of causal contrasts that could be of interest in this setting, and define them using the potential outcomes $\{Y(0), Y(1), Y(2)\}$. Interpret each contrast in words.*
2. *Write the observed outcome $Y$ as a function of $\{Y(0), Y(1), Y(2)\}$ and indicator functions $I\{A = a\}$.*
3. *Explain why, for any given individual, at most one element of $\{Y(0), Y(1), Y(2)\}$ is observed.*

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

R. Chetty, N. Hendren, and L. F. Katz. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902, April 2016. doi: 10.1257/aer.20150572. URL https://www.aeaweb.org/articles?id=10.1257/aer.20150572.

S. Das. Effect of merger on market price and product quality: American and us airways. *Review of Industrial Organization*, 55(3):339–374, 2019.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2025.

J. S. Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Annals of Agricultural Sciences*, 5:1–51, 1923. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/2245382.

S. Wager. Causal inference. Standford University, 2020.

# 2

# *Causality and Random Assignment*

At the end of the previous chapter, we introduced the potential outcomes framework and discussed the counterfactual problem. We also examined an example demonstrating how an experiment can help us interpret an average comparison between a "treated" and a "control" group as causal.

In this chapter, we will formalize these intuitions. First, we will focus on population quantities. We will prove that under certain assumptions, random assignment of an action $A$ (in the MTO Example 1.2, $A$ corresponds to the experimental group or voucher allocation) allows for the identification of the potential outcomes distribution $Y(a)$. This result highlights the power of experiments in addressing causal questions.

Next, we will explore how to implement these intuitions and estimate treatment effects using actual data from an experiment. To illustrate, consider the following example, which we will later analyze using the tools developed in this chapter:

**Example 2.1** Bertrand and Mullainathan [2004] conducted a randomized experiment on resumes to study the effect of perceived race on callbacks for interviews. They randomly assigned African-American - or White - sounding names on fictitious resumes to help-wanted ads in Boston and Chicago newspapers. The following two-by-two table summarizes perceived race and callback:

| | callback | no callback |
|---|---|---|
| African-American | 157 | 2278 |
| White | 235 | 2200 |

We can compare the probabilities of being called back among African-American - and White - sounding names:

$$\frac{157}{2435} - \frac{235}{2435} = 6.45\% - 9.65\% = -3.20\% < 0 \ .$$

That is, white names received more callbacks. In this case we could do a Fisher's exact test (later on in class) to learn that this difference is statistically significant with a p-value smaller than 0.001. In Bertrand and Mullainathan [2004]'s experiment, the treatment is the "perceived race" (as opposed to race) which can be manipulated by experimenters and so tackles a well-defined

causal question. In the rest of the chapter, we are going to see *why* we can interpret this probability difference (or mean differences in general) as the *causal* effect of perceived race on callbacks, and also how to implement this estimator. ■

## 2.1 Identification via Random Assignment

The main element in our search for identification is the assumption we make on the treatment (or intervention) variable $A$. Random assignment is the assumption that the treatment variable $A$ is independent of the potential outcomes $Y(a)$ for all $a \in \mathcal{A}$. In other words, random assignment is the assumption that

$$\{Y(a) : a \in \mathcal{A}\} \perp\!\!\!\perp A \ . \tag{2.1}$$

Under random assignment, the distribution of $Y(a)$ is identified,

$$F_a(y) := P\{Y(a) \leq y\} = P\{Y(a) \leq y \mid A = a\} = P\{Y \leq y \mid A = a\} \ , \tag{2.2}$$

where the second equality is due to random assignment, and the last equality follows from recalling that under our assumptions $Y = Y(A)$. The intuition of this step is that under random assignment, conditioning on treatment does not change the distribution of potential outcomes. In other words, there is nothing systematically different about the treatment and control groups. Identification follows from the last equality since the distribution of $Y$ given $A = a$ is identified from the data. It follows from this result that any parameter that is a function of $\{F_a : a \in \mathcal{A}\}$ is also identified. Some common parameters of interest when $A$ is binary are :

- Average treatment effect (ATE): $E[Y(1) - Y(0)]$
- Average treatment on the treated (ATT): $E[Y(1) - Y(0) \mid A = 1]$
- Average treatment on the untreated (ATU): $E[Y(1) - Y(0) \mid A = 0]$

Under random assignment, all of these parameters are identified and $ATE = ATT = ATU$ (we will prove this formally in Problem 2.1).

To see more directly how the ATE is identified by the distribution of the observed data $(Y, A)$ under random assignment, note that if we take the differences in population means for the "treated" and "control" groups we get:

$$\begin{aligned} E[Y \mid A = 1] - E[Y \mid A = 0] &= E[Y(1) \mid A = 1] - E[Y(0) \mid A = 0] \\ &= E[Y(1)] - E[Y(0)] \\ &= E[Y(1) - Y(0)] \\ &= \theta \ , \end{aligned} \tag{2.3}$$

where the first equality follows from the treatment assignment (the outcome we observe is the potential outcome of the actual assignment). The second equality follows from random assignment: $A$ is independent of the potential outcomes $Y(0), Y(1)$ –see (2.1) and the Hint 2.1 below–, and the third equality follows from $Y = Y(1)A + (1 - A)Y(0)$.

**Hint 2.1** *In the proof for (2.3) above, we used the fact that when two random variables $W, Z$ are independent, then $E[W|Z] = E[W]$.*

Note that even under the assumption in (2.1), the joint distribution of the potential outcomes $Y(0)$ and $Y(1)$ is not identified. That is,

$$P\{Y(1) \leq y_1, Y(0) \leq y_0\} \ ,$$

is not identified. This is because we never observe both potential outcomes for the same unit, which is another way of stating the *fundamental problem of causal inference*. Importantly, most features of $\Delta = Y(1) - Y(0)$ are not identified. For instance, we cannot identify the proportion of individuals who are hurt by the treatment, $P\{Y(1) < Y(0)\}$.

## 2.2 Estimation via Difference in Means

Recalling Example 2.1, people were randomly assigned names commonly associated with a certain racial group. This generates data on two groups, one with names commonly associated with African–Americans, and one group with names commonly associated with Caucasian-Americans. Per the results in Section 2.1, an intuitive path to obtain a causal effect would be to compare the average probability of a callback of the treatment group, with the average probability of callback in the control group. Since the assignment of the name is random, the groups should be *comparable*, given that nothing should be different between the groups on average other than having a different name. This is the approach that the authors use in their study. Let's prove that this approach actually estimates the ATE.

Consider the setting from a randomized controlled experiment (or trial, i.e., RCT) where we have a binary treatment variable $A$ (where $A = 1$ when the unit is treated and $A = 0$ when the unit is part of the control group) and an outcome variable $Y$. We assume that we have a random sample of size $n$ from the distribution of $(Y, A)$, denote the distribution by $P$. Assume that $A$ is exogenous in the sense of (2.1), i.e., the treatment is assigned randomly. In the previous section we characterized the ATE parameter $\theta$ as a function of $(Y, A)$ in (2.3). Call $n_1$ the number of individuals $i$ in the treatment group, and $n_0$ the number of individuals in the control group. Given this representation, the natural estimator of the ATE is the difference in means estimator (which

we will denote by $\hat{\theta}_n$) given by

$$\hat{\theta}_n := \frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Y_i - \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} Y_i \ , \qquad (2.4)$$

where we used the notation

$$\mathcal{I}_a := \{i \in \{1,\dots,n\} : A_i = a\}$$

to denote the set of units with treatment status $a$ and $n_a := |\mathcal{I}_a|$ to denote the number of units with treatment status $a$. That is, the estimator in (2.4) is the difference in sample means of the outcome variable between the treatment and control groups.

Under the random assignment assumption in (2.1), the estimator in (2.4) is unbiased for $\theta$; that is

$$E[\hat{\theta}_n] = \theta \ .$$

To see this, first note that

$$\begin{aligned} E[\hat{\theta}_n] &\overset{(1)}{=} E\Big[E[\hat{\theta}_n \mid A_1, ..., A_n]\Big] \\ &\overset{(2)}{=} E\left[\ E\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i \mid A_1,\dots,A_n\right] - E\left[\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) \mid A_1,\dots,A_n\right]\ \right] , \end{aligned}$$

where $\overset{(1)}{=}$ follows from the law of iterated expectations (LIE, see Hint 2.2), and $\overset{(2)}{=}$ follows from the linearity of the expectation. Then, taking the conditional expectation of the treatment group average (i.e. the first term above) we get:

$$\begin{aligned} E\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i \mid A_1,\dots,A_n\right] &\overset{(3)}{=} \frac{1}{n_1}\sum_{i=1}^{n} E[A_i Y_i \mid A_1,\dots,A_n] \\ &\overset{(4)}{=} \frac{1}{n_1}\sum_{i=1}^{n} E[A_i Y_i(1) \mid A_1,\dots,A_n] \\ &\overset{(5)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1) \mid A_1,\dots,A_n] \\ &\overset{(6)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1) \mid A_i] \\ &\overset{(7)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1)] \\ &\overset{(8)}{=} E[Y(1)] \ , \end{aligned}$$

where $\overset{(3)}{=}$ follows from the linearity of the expectation; $\overset{(4)}{=}$ follows from $Y_i A_i =$

$Y_i(1)A_i$ is non-random conditional on $(A_1, \ldots, A_n)$; $\overset{(5)}{=}$ uses that $A_i$ is non-random conditional on $(A_1, \ldots, A_n)$; $\overset{(6)}{=}$ follows from i.i.d. sampling across units; $\overset{(7)}{=}$ uses random assignment $(Y_i(1) \perp\!\!\!\perp A_i)$; and $\overset{(8)}{=}$ uses that $\sum_{i=1}^{n} A_i = n_1$. Similar arguments show that

$$E\left[\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - A_i) \mid A_1, \ldots, A_n\right] = E[Y(0)] \ ,$$

which implies that $E[\hat{\theta}_n] = E[Y(1)] - E[Y(0)] = \theta$ by the law of iterated expectations (LIE, see Hint 2.2).

**Hint 2.2** *In this class we will use the law of iterated expectations (LIE) all the time. Recall that the LIE states that for any random variables $X$ and $Y$, we have:*

$$E[Y] = E[E[Y \mid X]] = \int E[Y \mid X = x] f(x) dx \ .$$

*This means that the expectation of $Y$ can be computed by first conditioning on another variable $X$, and then taking the expectation of the conditional expectation of $Y$. The LIE is particularly useful for breaking down complex expectations into simpler components, especially when dealing with conditional expectations.*

Later in this class we will show that this estimator is also consistent and asymptotically normal under these assumptions. For now, assume that it is and let's see how to implement this simple estimator and perform a t-test (to test whether the estimator is statistically different from zero) in R, using the data on Bertrand and Mullainathan [2004]'s study:

```r
# Load necessary libraries
library(dplyr)
library(haven)

# Load the dataset, which is originally in STATA format
# (but R can easily read it)
data <- read_dta("lakisha_aer.dta")

# View the structure and details of the dataset since we imported
    from STATA
str(data)
summary(data)

# Shows a Table of the proportion of callbacks for each perceived
    race
means_comparison <- data %>%
  group_by(race) %>%
  summarise(mean_callback = mean(call, na.rm = TRUE),
            count = n())

```

```r
# Print table results
print(means_comparison)

# If you want to perform a t-test for statistical significance
t_test_result <- t.test(data$call[data$race == "b"], # Selects callbacks for A-A
                        data$call[data$race == "w"], # Selects callbacks for C-A
                        paired = TRUE)

# Print the t-test results
print(t_test_result)
```

Code Snippet 2.1: Means Difference in an Experimental Setting

Using this code, we should get the result of a mean difference of -3.20% of callbacks for people assigned African–American sounding names with respect to the control group assigned Caucasian–American names. At the end of the snippet, a t-test is performed and we obtain a p-value smaller than 0.001, which indicates statistical significance of the estimate.

## 2.3 Scope of Random Assignment

One may wonder when is random assignment a good assumption. The answer is that this is a reasonable assumption only in contexts where the experimenter has control over $A$ (as in Examples 2.1 and 1.2) or, in other words, settings where agents have no control over $A$. It is much less likely to hold in settings where units choose $A$, since they typically choose $A$ using information about $\{Y(a) : a \in \mathcal{A}\}$. For example, thinking about Example 1.2, the MTO program, in general people choose themselves into neighborhoods depending on income. In those cases, we expect *selection* into treatment.

Formally, we say there is selection into the treatment state $A$ if

$$Y(a)|A = a \text{ is distributed } \textbf{differently} \text{ from } Y(a)|A = a' \text{ for } a \neq a' \ .$$

This is expected to occur if agents choose $A$ with knowledge of $\{Y(a) : a \in \mathcal{A}\}$. For example, agents who choose to join a job training program might do so because of a low value of $Y(0)$. In the neighborhood case, we could think of $A = 1$ as a high rent neighborhood and $A = 0$ as a lower rent neighborhood. Then, families that can afford it will choose $A = 1$, their children obtaining $Y(1)$ as future income, and vice versa, which would make observable data of $Y(1), Y(0)$ correlate with family income and neighborhood selection. In either case, we expect $Y(0)$ to be distributed differently across the two groups and this in turn leads to the so-called *selection bias*.

The difference in means estimator in (2.4) converges to the ATE only if there is no selection into the treatment state. More generally, under our

random sampling assumption, it converges to the observed mean difference,

$$E[Y \mid A = 1] - E[Y \mid A = 0] \ ,$$

which equals the ATE only under random assignment (no selection into the treatment state). We can then decompose the contrast into a causal effect and selection bias:

$$\begin{aligned} E[Y \mid A = 1] - E[Y \mid A = 0] = & \underbrace{E[Y(1) \mid A = 1] - E[Y(0) \mid A = 1]}_{\text{ATT}} \\ & + \underbrace{E[Y(0) \mid A = 1] - E[Y(0) \mid A = 0]}_{\text{Selection Bias}} \ . \end{aligned}$$

The first difference on the right-hand side is the causal effect for those who were treated (the ATT), while the second difference is the selection bias term that captures how the treated would have been different anyway. These effects could cancel out if the ATT is $(+)$ while the selection bias is $(-)$, but they could also go in the same direction and exacerbate each other.

Recall the study of the effect of neighborhoods on future income of Example 1.2. Now, imagine we just compare the future income of children coming from high-income neighborhoods ($A = 1$) with the future income of people that grew up in low–income neighborhoods ($A = 0$), without an experiment. Then, we compute the sample version of $E[Y \mid A = 1] - E[Y \mid A = 0]$, where $Y$ is the future income. If high income families select themselves into high income neighborhoods, then the selection bias in our estimator would be positive: We could be summing the effect of the high-income neighborhood (the first term above, the ATT) with the effect on future income that coming from a wealthy family has (the selection bias). Then, we would be *overestimating* the effect of neighborhoods.

## 2.4 Key Concepts

- **Average Treatment Effect (ATE):** The average treatment effect is the population mean of the treatment effect. In the binary case, $\theta := E[Y(1) - Y(0)]$.
- **Difference in Means Estimator:** The difference in means estimator compares the sample mean outcome in the treated group to the sample mean outcome in the control group: $\hat{\theta}_n = \bar{Y}_1 - \bar{Y}_0$. Under random assignment, it identifies the ATE.
- **Average Treatment Effect on the Treated (ATT):** The av-

erage treatment effect for the units that are treated, $ATT := E[Y(1) - Y(0) \mid A = 1]$.

- **Selection Bias:** Selection bias is the component of

$$E[Y \mid A = 1] - E[Y \mid A = 0]$$

due to systematic differences between treated and control units, rather than the causal effect of $A$.

## 2.5 Concluding Remarks

The contents of this chapter are based on the notes by Peng Ding [2023], the notes by Stefan Wager [2020], and notes shared by Alex Torgovistky. The book by Angrist and Pischke [2008] is another reference that includes discussion on many of the topics we covered.

## 2.6 Problems

**Problem 2.1** *Show that under random assignment we obtain*

$$ATE = ATT = ATU \ .$$

**Problem 2.2** *In this problem, we look into more details of Bertrand and Mullainathan (2004)'s study on the effect of perceived race on callbacks for interviews. They first generated a pool of resumes for the fictitious job applicants. The resumes are classified into two categories: high and low quality. They also generated a pool of names for the fictitious job applicants. The names are classified into four categories: African-American male, African-American female, white male, and white female.*

1. *Suppose for each resume randomly drawn from the pool of resumes, a name is randomly drawn from the pool of names to generate a fictitious job applicant. In this case, is the random assignment (of perceived race) assumption valid for each of the following subgroups? Justify your answer.*

   *(a) high quality resume;*

   *(b) low quality resume;*

   *(c) female;*

*(d) male.*

2. *For the four subgroups in part 1, conduct the same analysis as in Section 1.5. That is, for each subgroup, calculate the difference in means estimator for the effect of perceived race on the probability of callback. Report the estimates in a table and explain your findings.*

**Problem 2.3** *Imagine a study evaluating the effect of regular exercise on cardiovascular health. The analyst decided to survey patients at a large hospital and collect information about their daily routine. Then, the analyst decided to use a mean contrast,* $E[Y|A=1] - E[Y|A=0]$ *with the intention to identify the average treatment effect.*

1. *Explain whether you think the mean contrast would identify or not the ATE.*

2. *If the answer to the previous question is yes, show that* $A \perp\!\!\!\perp Y(a)$ *for* $a \in \mathcal{A}$.

3. *If the answer to part (1) is no, then explain what your concerns are and argue whether you would expect* $E[Y|A=1] - E[Y|A=0] > ATE$ *or* $E[Y|A=1] - E[Y|A=0] < ATE$*, both mathematically and in words.*

**Problem 2.4** *Recall the parameter* $ATT := E[Y(1) - Y(0) \mid A = 1]$*. Given what we saw earlier in the chapter, and under the assumptions we discussed (random sampling, random assignment (2.1), and* $Y_i = Y_i(A_i)$*), propose an unbiased estimator for ATT based on the observed data* $\{(Y_i, A_i) : 1 \le i \le n\}$.

**Problem 2.5** *Consider a randomized experiment with a binary treatment* $A \in \{0, 1\}$ *and outcome* $Y$*. Let* $\pi := P\{A = 1\}$ *denote the (known) assignment probability. Define the Horvitz–Thompson (inverse probability weighted) estimator*

$$\hat{\theta}_{HT,n} := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{A_i Y_i}{\pi} - \frac{(1 - A_i) Y_i}{1 - \pi} \right) .$$

*Assume we have a random sample from the joint distribution of* $(Y, A)$*, random assignment (2.1), and recall that under our assumptions* $Y_i = Y_i(A_i)$*. Show that* $\hat{\theta}_{HT,n}$ *is unbiased for the average treatment effect* $\theta = E[Y(1) - Y(0)]$.

**Problem 2.6** *Fix* $a \in \mathcal{A}$*. We proved that under random assignment and the rest of our maintained assumptions, the distribution of* $Y(a)$ *is identified by the observed data, i.e.*

$$P\{Y(a) \le y\} = P\{Y \le y \mid A = a\} \qquad \text{for all } y \in \mathbf{R} .$$

*Let* $h : \mathbf{R} \to \mathbf{R}$ *be any function such that* $E[|h(Y(a))|] < \infty$*. Show that*

$$E[h(Y(a))] = E[h(Y) \mid A = a] .$$

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

S. Wager. Causal inference. Standford University, 2020.

# 3

# *Linear Regression*

In the previous lecture, we introduced the concepts of counterfactuals, potential outcomes, and causal parameters such as the average treatment effect (ATE). We also discussed how the ATE can be identified under random assignment, leading to the difference-in-means estimator. In this lecture, we will introduce the linear regression model and explore how it can be used to estimate causal parameters. We will also examine the assumptions necessary to interpret the parameters in the linear regression model causally, and discuss how the ATE can be estimated using linear regression under random assignment.

First, we will review how to estimate a linear regression model using ordinary least squares (OLS). Next, we will provide an overview of the interpretations of the linear regression model. Finally, we will address the question of when, if at all, linear regression results can be interpreted as causal parameters.

**Note about notation:** We will distinguish between population random variables and sample observations. When referring to population quantities, we will use notation *without* a subscript $i$. When referring to sample data, we will use subscripts such as $Y_i$ and $X_i$ to denote the $i$th observation in a sample of size $n$.

## 3.1 Linear Regression with Exogenous Regressors

### 3.1.1 Solving for $\beta$

Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ (they are scalars) and $X$ takes values in $\mathbf{R}^{k+1}$ (i.e., $X$ is a vector of dimension $k+1$). Assume that the first component of $X$ is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Consider the linear model

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$ and that $E[XX']$ exists and is invertible (i.e., there is no perfect collinearity in $X$). Under these assumptions, we can solve for the population regression coefficient $\beta$.

The assumption $E[XU] = 0$ will play different roles depending on how we interpret the linear model (we return to this later in the lecture). The "no perfect collinearity" condition simply rules out cases where one component of $X$ is an exact linear combination of the others, ensuring that $E[XX']$ is invertible. This is summarized in the following lemma, which you can prove in Problem 3.1.

**Lemma 3.1** *Let $X$ be a random vector such that $E[XX'] < \infty$. Then $E[XX']$ is invertible if and only if there is no perfect collinearity in $X$.*

Now, let's solve for $\beta$. Starting with the linear model and isolating $U$ , we get:

$$\begin{aligned} U &= Y - X\prime\beta \\ XU &\overset{(1)}{=} X(Y - X'\beta) \\ E[XU] &= E[X(Y - X'\beta)] \\ E[XU] &\overset{(2)}{=} E[XY] - E[XX']\beta \\ 0 &\overset{(3)}{=} E[XY] - E[XX']\beta \\ E[XX']\beta &= E[XY] \end{aligned}$$

Here, (1) follows from pre-multiplying the equation by $X$, (2) uses the linearity of expectation, and (3) comes from the first assumption $E[XU] = 0$. Since $E[XX']$ is invertible, we can solve for $\beta$ by multiplying both sides of the equation by $E[XX']^{-1}$ :

$$\beta = E[XX']^{-1}E[XY] \ . \tag{3.1}$$

If $E[XX']$ is not invertible, i.e., there is perfect collinearity in $X$, then there will be more than one solution to this system of equations. Importantly, any two solutions $\beta$ and $\tilde{\beta}$ will necessarily satisfy $P\{X'\beta = X'\tilde{\beta}\} = 1$. Depending on the interpretation we give to the linear model, this may be an important distinction or not. For instance, in the second interpretation, each such solution corresponds to the same "best" linear predictor of $Y$ given $X$, whereas in the third interpretation different values of $\beta$ could have wildly different implications for how $X$ affects $Y$ holding $U$ constant.

## 3.2 Estimating $\beta$

Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ and $X$ takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of $X$ is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, $E[XX'] < \infty$ and that there is no perfect collinearity in $X$. Above we determined that under these assumptions, we can solve for $\beta$. We now discuss estimation of $\beta$ when we observe an i.i.d. sample of outcomes $Y$ and covariates $X$.

### 3.2.1 Ordinary Least Squares

Let $(Y, X, U)$ be distributed as described above and denote by $P$ the marginal distribution of $(Y, X)$. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sequence of random vectors with distribution $P$. By analogy with the expression we derived for $\beta$ in (3.1) under these assumptions, the natural estimator of $\beta$ is simply

$$\hat{\beta}_n = \left( \frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} X_i Y_i \right) .$$

This estimator is called the *ordinary least squares* (OLS) estimator of $\beta$ because it can also be derived as the solution to the following minimization problem, which corresponds to the sum of squared errors $U_i = Y_i - X_i'\beta$:

$$\hat{\beta}_n = \underset{b \in \mathbf{R}^{k+1}}{\operatorname{argmin}} \frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i'b)^2 .$$

To see this, note that $\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i'b)^2$ is convex (as a function of $b$) and so

$$\frac{\partial \frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i'b)^2}{\partial b} = \frac{1}{n} \sum_{1 \leq i \leq n} -2X_i(Y_i - X_i'b) ,$$

where we used the fact that $\frac{\partial A'z}{\partial z} = A$. Hence the solution, $\hat{\beta}_n$, must satisfy

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i(Y_i - X_i'\hat{\beta}_n) = 0 ,$$

or, re-arranging terms,

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i Y_i = \frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \hat{\beta}_n .$$

The matrix

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i'$$

may not be invertible, but, since $E[XX']$ is invertible, it will be invertible with probability approaching one. Then, if we premultiply our last equality by $\left( \frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1}$ we obtain the OLS estimator $\hat{\beta}_n$ above.

The $i$th *fitted value* is denoted by $\hat{Y}_i = X_i'\hat{\beta}_n$. The $i$th *residual* is denoted by $\hat{U}_i = Y_i - \hat{Y}_i$. By definition, we therefore have that

$$\frac{1}{n}\sum_{1\le i\le n} X_i\hat{U}_i = 0\ .$$

**Example 3.1 (Bertrand and Mullainathan (2004), cont.)** In this example, we apply the OLS estimator to the work by Bertrand and Mullainathan [2004] on perceived race and job callbacks. For details, please refer to Example 2.1. Below is the code that loads the study's data, performs the mean comparison, and then conducts an OLS regression with the following specification:

$$Y_i = X_i'\beta + U_i, \quad X_i' = (1\ A_i), \quad \beta = (\beta_0\ \beta_1)'\ ,$$

where $A = 1$ if the perceived race is African-American and 0 otherwise (this variable is created in line 19 of the code below).

```
1  # Load necessary libraries
2  library(dplyr)
3  library(haven)
4
5  # Load the dataset, which is originally in STATA format
6  # (but R can easily read it)
7  data <- read_dta("lakisha_aer.dta")
8
9  # Shows a Table of the number of callbacks for each perceived
       race
10 means_comparison <- data %>%
11   group_by(race) %>%
12   summarise(mean_callback = mean(call, na.rm = TRUE),
13             count = n())
14
15 # Print means comparison results
16 print(means_comparison)
17
18 # Create an indicator/dummy variable that is 1 if the perceived
       race is A
19 data$race_dummy <- as.integer(data$race == "b")
20
21 # Fit a linear regression model that compares the number of
       callbacks
22 # between perceived race:
23 linear_model <- lm(call ~ 1 + race_dummy, data = data)
24
25 # Print the linear regression model summary
26 print(summary(linear_model))
```

Code Snippet 3.1: Regression in an Experimental Setting

The OLS estimate (or LS estimate for short) is $\hat{\beta}_1 = -3.2\%$, consistent with the difference-in-means estimator. Why is this the case? It turns out that in certain settings the LS estimate is exactly the difference-in-means estimator and so, under random assignment, it admits a causal interpretation. But is

it true that we can interpret $\hat{\beta}_1$ causally in general? In other words, can we interpret $\beta$ in (3.1) causally? To address this, we will first explore general interpretations of the linear model and discuss when we can interpret $\beta$ as a causal parameter. ■

## 3.3 Interpretations of the Linear Regression Model

Let's return to population values to interpret the linear model. Let $(Y, X, U)$ be a random vector where $Y$ and $U$ are scalars and $X \in \mathbf{R}^{k+1}$ has first component equal to 1 (an intercept), and consider

$$Y = X'\beta + U \ , \qquad \beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}.$$

The parameter $\beta_0$ is the *intercept* and the remaining $\beta_j$'s are *slopes*. We will discuss three common interpretations of $\beta$; first we introduce the conditional expectation, which describes the relationship between $Y$ and $X$ more generally.

### 3.3.1 About the Conditional Expectation Function

In the previous class, we learned that under random assignment, the average treatment effect $\theta$ can be identified with the mean contrast $E[Y|A = 1] - E[Y|A = 0]$. Later in the course, we will extend our analysis by including additional variables in the conditional expectations. It will become clear that many causal parameters are functions of the conditional expectation of the outcome $Y$, conditional on a set of variables (including the treatment $A$ and others), which we will denote by $X$, i.e., $E[Y|X]$. However, it is crucial to understand that this does not give conditional expectations a *causal* interpretation. Conditional expectations provide a natural way to summarize the joint distribution of $Y$ and $X$ , and can be seen as summarizing the relationship between them. In fact, conditional expectations are often used for *prediction* problems where one is interested in finding the value of $Y$ given a value of $X$. But this does not imply that $E[Y|X]$ can be interpreted causally as "if we change X from $x$ to $x'$ , then Y would be expected to change by $E[Y|X = x'] - E[Y|X = x]$." We will revisit this point in more detail later in class.

From a purely mechanical point of view, we can always decompose $Y$ into the conditional expectation with respect to another random variable $X$, plus an additive mean-zero error. Specifically, if we call this error $V$, we have:

$$Y = E[Y|X] + V \quad \text{with} \quad E[V|X] = 0 \ .$$

This follows by defining $V := Y - E[Y|X]$ and taking conditional expectations, $E[V|X] = E[Y|X] - E[Y|X] = 0$. Notice that this decomposition is always

true. Two important takeaways emerge from this. First, writing a model where $Y = E[Y|X] + V$ with $E[V|X] = 0$ is a tautology, and does not lead to causal statements; this holds for *any* variable $X$. Second, in order for this decomposition to be useful, we need to either focus on prediction questions, where the interest lies in $E[Y|X]$ itself, or we need to impose a model on $E[Y|X]$, such as a linear one. We will discuss this further in the next section.

### 3.3.2 Interpretation 1: Linear Conditional Expectation

Suppose that the true conditional expectation is linear, $E[Y|X] = X'\beta$, and define $U = Y - X'\beta$. This implies that $E[U|X] = 0$ and therefore that $E[U] = 0$, by applying the law of iterated expectations; see Hint 2.2. Moreover, $E[XU] = 0$, so $\text{Cov}[X, U] = 0$. It immediately follows that

$$E[XU] = E[X(Y - X'\beta)] = E[XY] - E[XX']\beta = 0 \ ,$$

which leads to the same characterization of $\beta$ we previously got (which only relies on the assumption $E[XU] = 0$ that here is a consequence of assuming a linear conditional expectation).

It is tempting to interpret the coefficient $\beta_j$ for $1 \leq j \leq k$ as the *ceteris paribus* (i.e., holding $X_{-j}$ and $U$ constant) effect of a one unit change in $X_j$ on $Y$, but this is incorrect. Indeed, more generally, it is not appropriate to think of differences in (or derivatives of) conditional expectations causally. After all, $Y$ could be an indicator for rain and $X$ could be an indicator for carrying an umbrella. In this case, it may be the case that $E[Y|X]$ is increasing in $X$, but one would not want to think of carrying an umbrella as causing rain. What is missing is a model of how $Y$ is determined as a function of $X$ (and possibly other unobserved variables).

### 3.3.3 Interpretation 2: "Best" Linear Approximation to the Conditional Expectation or "Best" Linear Predictor

In general, one would not expect the conditional expectation to be linear. Suppose $E[Y^2] < \infty$ and $E[XX'] < \infty$ (equivalently, $E[X_j^2] < \infty$ for $1 \leq j \leq k$). Under these assumptions, we can attempt to find the "best" linear approximation to the conditional expectation. A linear approximation is a function of the form $X'b$ for some choice of $b \in \mathbf{R}^{k+1}$ that is "closest" to the true, possibly non-linear, conditional expectation $E[Y|X]$. To this end, consider the minimization problem

$$\min_{b \in \mathbf{R}^{k+1}} E[(E[Y|X] - X'b)^2].$$

Denote by $\beta$ a solution to this minimization problem. When the solution is unique (after, for example, assuming there is no-perfect collinearity in $X$), the solution to this problem is once again $\beta = E[XX']^{-1}E[XY]$.

To see this last claim more clearly, let $V = E[Y|X] - Y$ so $E[XV] = 0$. Note that

$$\begin{aligned} E[(E[Y|X] - X'b)^2] &= E[(E[Y|X] - Y + Y - X'b)^2] \\ &= E[(V + Y - X'b)^2] \\ &= E[V^2 + 2V(Y - X'b) + (Y - X'b)^2] \\ &= E[V^2] + 2E[VY] - 2E[VX']b + E[(Y - X'b)^2] \\ &= \text{constant} + E[(Y - X'b)^2] \ . \end{aligned}$$

Thus, $\beta$ also solves

$$\min_{b \in \mathbf{R}^{k+1}} E[(Y - X'b)^2] \ .$$

Note that $E[(Y - X'b)^2]$ is convex (as a function of $b$) and thus the first order condition of this problem is

$$\frac{\partial E[(Y - X'b)^2]}{\partial b} = E[-2X(Y - X'b)] \ .$$

Hence, $\beta$ must satisfy

$$E[X(Y - X'\beta)] = 0 \ .$$

If we further assume that $E[XX']$ is invertible and solve for $\beta$, we obtain $\beta = E[XX']^{-1}E[XY]$ as before. Importantly, note that if we define $U = Y - X'\beta$, then we may rewrite this equation as

$$E[XU] = 0 \ .$$

Under this second interpretation, $\beta$ is simply a convenient way of summarizing another feature of the joint distribution of $Y$ and $X$, namely, the "best" linear approximation to the conditional expectation. For the same reasons as before, it is not correct to interpret the coefficient $\beta_j$ for $1 \leq j \leq k$ causally, i.e., as the *ceteris paribus* effect of a one unit change in $X_j$ on $Y$.

**Example 3.2 (Correlation does not imply causation)** As is suggested by the two interpretations above, even if we manage to establish correlation between variables $Y$ and another $X$, it does not follow that there is a causal relationship between them. Note that in Figure 3.1, conditional on a number of babies born Johnny, we can predict the burglaries in New Hampshire in the corresponding year. The relationship over time even could be approximated linearly. But it's hard to think that one **causes** the other!

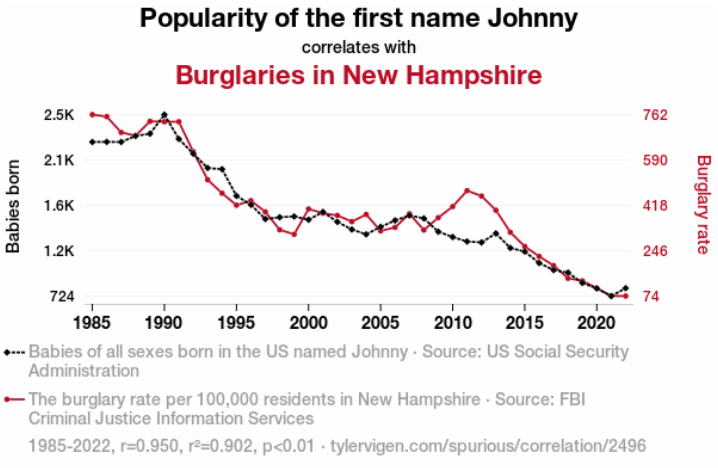



FIGURE 3.1: Do Johnnys cause burglaries or burglaries cause Johnnys? Neither!

■

### 3.3.4 Interpretation 3: Causal Model

Suppose $Y = g(X, U)$, where $X$ are the observed determinants of $Y$ and $U$ are the unobserved determinants of $Y$. Such a relationship is a model of how $Y$ is determined and may come from physics, economics, etc. The effect of $X_j$ on $Y$ holding $X_{-j}$ and $U$ constant (i.e., *ceteris paribus*) is determined by $g$. If $g$ is differentiable, then it is given by $D_{X_j} g(X, U)$. If we assume further that

$$g(X, U) = X'\beta + U \ ,$$

then the *ceteris paribus* effect of $X_j$ on $Y$ is simply $\beta_j$. We may normalize $U$ so that $E[U] = 0$ (by replacing $U$ with $U - E[U]$ and $\beta_0$ with $\beta_0 + E[U]$ if this is not the case). On the other hand, $E[U|X]$, $E[U|X_j]$ and $E[UX_j]$ for $1 \leq j \leq k$ may or may not equal zero. These are now statements about the relationship between the observed and unobserved determinants of $Y$.

It is probably fair to say that the causal interpretation of $\beta$ as described above is in disuse. The main reason not only lies in the difficulty of justifying in applications that $g(\cdot)$ is linear in $X$, but also in the fact that the model implicitly assumes that the effect of $X$ on $Y$ is *homogenous* across individuals - in other words, the model assumes that every single agent responds to changes in $X$ in the same way. However, it is also fair to say that least squares is prevalent in applied work and often interpreted as capturing *some type* of causal effect. We will explore this interpretation in the next section, after we have a characterization of $\beta$ as a function of the distribution of $(Y, X)$, and show that $\beta$ can sometimes be expresses as a *weighted average* of *heterogeneous*

causal effects. Interpreting $\beta$ in this way is delicate and heavily depends on the specifics of the application.

## 3.4 Key Concepts

- **Estimand and Estimator of the Linear Regression:** The estimand is the true population parameter, $\beta$, while the estimator is the value computed from sample data, $\hat{\beta}_n$.
- **Linear Conditional Expectation:** This interpretation assumes that the conditional expectation of the outcome $Y$ given $X$, $E[Y|X]$, is linear in the covariates, represented as $Y = X'\beta + U$.
- **Best Linear Predictor:** In this view, $X'\beta$ represents the best linear approximation to the conditional expectation $E[Y|X]$, minimizing the mean squared error between $X'b$ and $E[Y|X]$.
- **Causal Linear Model:** In this interpretation $\beta$ can be interpreted as the causal effect of $X$ on $Y$, but this interpretation requires a model for how $Y$ is a function of observed and unobserved variables.

## 3.5 Concluding Remarks

These notes are adapted from materials I have used in previous classes and are significantly influenced by Azeem Shaikh, whose notes and our ongoing discussions about teaching have been invaluable. Additional related concepts can be found in the books by Bruce Hansen [Hansen, 2022], Jeff Wooldridge [Wooldridge, 2025], and the one by Angrist and Pischke [Angrist and Pischke, 2008].

## 3.6 Problems

**Problem 3.1** *Prove Lemma 3.1.*

**Problem 3.2** *Using the law of iterated expectations, show that if* $E[U|X] = 0$ *it follows that* $E[U]$, $E[XU]$, *and* $\text{Cov}[X, U]$ *are all equal to zero.*

**Problem 3.3** *Answer each of the following TRUE or FALSE*

*(a) If* $E[XU] = 0$ *then it must be true that* $E[U|X] = 0$

*(b) Suppose that* $X \in \{0, 1\}$*. Then if* $E[XU] = 0$ *then it must be true that* $E[U|X] = 0$

**Problem 3.4** *Let* $(Y, X)$ *be a random vector taking values in* $\mathbf{R}^2$ *with finite first and second moments.*

*(a) Show that, without loss of generality, we can write*

$$Y = h(X) + U \ ,$$

*where* $U$ *is a scalar random variable satisfying* $E[U|X] = 0$ *and* $h(X)$ *is a function of* $X$*.*

*(b) Give an interpretation to* $h(X)$*.*

**Problem 3.5** *Let* $X$ *and* $U$ *denote random variables. Suppose that* $E[U \mid X] = 1$ *and that* $E[X] = 2$*.*

*(a) What is* $E[U]$*?*

*(b) What is* $E[XU]$*?*

*(c) What is* $\mathrm{Cov}[X, U]$*?*

*(d) Is* $U$ *mean independent of* $X$*?*

*(e) Is* $U$ *uncorrelated with* $X$*?*

*(f) Do you have enough information to determine whether* $X$ *is independent of* $U$*?*

**Problem 3.6** *Let* $(Y, X_1)$ *be a random vector taking values in* $\mathbf{R} \times \{0, 1\}$*, this is, the random variable* $X_1$ *is binary. Assume that* $X_1$ *is randomly assigned, and let* $(Y(0), Y(1))$ *denote potential outcomes as defined in class. Assume we observe* $Y(1)$ *when* $X_1 = 1$ *and* $Y(0)$ *when* $X_1 = 0$*, so that*

$$Y = Y(0)(1 - X_1) + Y(1)X_1 \ . \tag{3.2}$$

*(a) Show that we can write*

$$Y = \beta_0 + \beta_1 X_1 + U \ , \tag{3.3}$$

*for a random variable* $U$ *satisfying* $E[X_1 U] = 0$ *and constant coefficient* $(\beta_0, \beta_1)$*. Find expressions for* $\beta_0$*,* $\beta_1$*, and* $U$ *in terms of* $(Y(0), Y(1), X_1)$*.*

*(b) Let* $\theta$ *denote the average treatment effect. Show that*

$$\mathrm{Cov}[Y, X_1] = \theta \, \mathrm{Var}[X_1] \ .$$

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach 8th ed.* Cengage learning, 2025.

# 4

# *Properties of LS*

Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ (they are scalars) and $X$ takes values in $\mathbf{R}^{k+1}$ (i.e., $X$ is a vector of dimension $k+1$). Assume that the first component of $X$ is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Consider the linear model

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$ and that $E[XX']$ exists and is invertible (i.e., there is no perfect collinearity in $X$). Under these assumptions, we can solve for the population regression coefficient $\beta$ and also obtain the OLS estimator of $\beta$, which we denote by $\hat{\beta}_n$, and is given by

$$\hat{\beta}_n = \left( \frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \le i \le n} X_i Y_i \right) \ . \tag{4.1}$$

In this chapter, we will review the properties of the OLS estimator, namely unbiasedness, consistency, and asymptotic normality. Unbiasedness we can prove with the tools we already know: the properties of expectations and the law of iterated expectations. Consistency and asymptotic normality, on the other hand, require three main tools that are routinely used in econometrics: the Law of Large Numbers, the Continuous Mapping Theorem, and the Central Limit Theorem. Before we study those properties, we therefore need to review these three tools.

## 4.1 (Un)Bias of LS

Suppose that, in addition to the assumptions of the linear model, we have that

$$E[U|X] = 0 \ .$$

Equivalently, assume that

$$E[Y|X] = X'\beta \ .$$

Recall that the least squares estimator can be written as

$$\hat{\beta}_n = \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Y_i \ .$$

Using the model $Y_i = X_i'\beta + U_i$, we obtain

$$\hat{\beta}_n = \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i (X_i'\beta + U_i) = \beta + \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i U_i \ .$$

We first show that $\hat{\beta}_n$ is unbiased conditional on the regressors. Conditioning on $(X_1, \ldots, X_n)$,

$$E[\hat{\beta}_n \mid X_1, \ldots, X_n] = \beta + \left( \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i \, E[U_i \mid X_1, \ldots, X_n] \ .$$

Since the observations are independent across $i$, $(X_j, U_j)$ is independent of $(X_i, U_i)$ for $j \neq i$. Hence,

$$E[U_i \mid X_1, \ldots, X_n] = E[U_i \mid X_i] \ .$$

By the assumption $E[U \mid X] = 0$, it follows that $E[U_i \mid X_i] = 0$ for all $i$, and therefore

$$E[\hat{\beta}_n \mid X_1, \ldots, X_n] = \beta \ .$$

Unconditional unbiasedness then follows by the law of iterated expectations:

$$E[\hat{\beta}_n] = E\big[E[\hat{\beta}_n \mid X_1, \ldots, X_n]\big] = E[\beta] = \beta \ .$$

Thus, under $E[U|X] = 0$, the least squares estimator is unbiased both conditionally on the sample of regressors and unconditionally. In particular, for any sample size $n$, the sampling distribution of $\hat{\beta}_n$ is centered at the true parameter vector $\beta$.

## 4.2 Consistency of LS

To show that the LS estimator is consistent, we need to show that $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To do this, we will use the law of large numbers and the continuous mapping theorem.

### 4.2.1 A Quick Review of the Law of Large Numbers

Consistency proofs usually rely on two workhorses:(a) the (weak) law of large numbers (LLN), and (b) the continuous mapping theorem (CMT).

Let $(X_1, \ldots, X_n)$ be an i.i.d. sequence of random variables with distribution $P$ such that the mean $E[X]$ exists; i.e., $E[|X|] < \infty$. The LLN states that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X] \quad \text{as } n \to \infty \ . \tag{4.2}$$

The result is not restricted to scalar random variables; it applies verbatim to vectors. If $(X_1, \ldots, X_n)$ is an i.i.d. sequence of random vectors taking values in $\mathbf{R}^k$ with mean $E[X]$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X] \quad \text{as } n \to \infty \ . \tag{4.3}$$

Note that (4.2) and (4.3) look identical; the only difference is the dimension of $X$ and $E[X]$.

To build some intuition for the law of large numbers, it is useful to think of sample averages as approximating population expectations when the sample size is large. Informally, the WLLN (or just LLN) states that the average of "something" converges in probability to the expected value of that "something", provided the "something" has a finite expectation (a feature that is typically simply assumed). This is:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{something}) \xrightarrow{P} E[\text{something}] \quad \text{as } n \to \infty \ . \tag{4.4}$$

There are two formal statements that capture the intuition in (4.4). First, let $h(\cdot)$ be any function such that $E[h(X)]$ exists. Then, under the same i.i.d. assumption, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{P} E[h(X)] \quad \text{as } n \to \infty \ . \tag{4.5}$$

An immediate application of this result would be

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} E[X^2] \quad \text{as } n \to \infty \ ,$$

or

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{P} E[XX'] \quad \text{as } n \to \infty \ ,$$

when $X \in \mathbf{R}^k$. Second, when the observed data is a sequence of multiple

random variables, like an i.i.d. sample $(Y_1, X_1), \ldots, (Y_n, X_n)$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i) \xrightarrow{P} E[h(X,Y)] \quad \text{as } n \to \infty \,, \tag{4.6}$$

for any $h(\cdot)$ such that $E[h(X,Y)]$ exists. An immediately application of this result would be the case where $h(\cdot)$ is the product of these variables, and so

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{P} E[XY] \quad \text{as } n \to \infty \,.$$

The Law of Large Numbers (LLN) is routinely used to show that estimators—such as least squares, two-stage least squares, machine learning estimators, and others—converge in probability to the parameter of interest. However, in most cases, the LLN alone is not sufficient to establish such results. The reason is that these estimators are typically not simple averages, but rather functions of several sample averages. To handle this, we rely on an additional tool known as the **Continuous Mapping Theorem** (CMT). We now state a formal version of the CMT.

**Theorem 4.1** *Let $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ be a sequence of random vectors satisfying*

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{P} E[h(X)] \quad \textit{as } n \to \infty \,.$$

*Assume the function $g(\cdot)$ is continuous at the point $E[h(X)]$. Then*

$$g\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right) \xrightarrow{P} g\left(E[h(X)]\right) \quad \textit{as } n \to \infty \,. \tag{4.7}$$

Note that while the function $h$ in the CMT can be *any* function, the function $g$ must be *continuous* at the limit point. The most common applications of the CMT arise when we deal with products or ratios of averages. For instance, the CMT implies that

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \xrightarrow{P} E[X]E[Y] \quad \text{as } n \to \infty \,,$$

and

$$\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \xrightarrow{P} \frac{E[Y]}{E[X^2]} \quad \text{as } n \to \infty \,,$$

where in the second case we require the assumption $E[X^2] \neq 0$ to ensure that the function is continuous at the limit point. You may be unsure about what it means for a function to be continuous "at the limit point," but this example helps clarify that requirement. A ratio is a continuous function provided the

denominator is nonzero. The CMT does not require the denominator to be nonzero for every $n$ along the sequence, but only that the limit of the denominator is nonzero. This is why we assume $E[X^2] \neq 0$, rather than requiring $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \neq 0$ for all $n$.

A few important remarks:

- A common mistake when invoking the LLN is to forget the factor $\frac{1}{n}$. Always make sure you have $\frac{1}{n}\sum_{i=1}^{n}$(something) as opposed to $\sum_{i=1}^{n}$(something).
- Sometimes the expression within the sum can be messy. For example, in this class it is not unusual to find terms like

$$\frac{1}{n}\sum_{i=1}^{n} Y_i I\{A_i = a, W_i = w\} \ .$$

  A bullet-proof approach to dealing with cases that are complicated consists in re-naming the expression inside the sum as $X_i$, apply the LLN, and then replace $X$ back with the original expression like this :

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{Y_i I\{A_i = a, W_i = w\}}_{X_i} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X] = E[Y I\{A = a, W = w\}] \ .$$

- In this class, it is also common to encounter cases where the factor multiplying the sum is *not* $\frac{1}{n}$. For example, consider the case

$$\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i \ .$$

  The best way to handle such cases is to rewrite the $n_1$ by its expression, make sure you have the required $\frac{1}{n}$ in front of all sums (perhaps by dividing and multiplying by $n$), and then apply the CMT. That is,

$$\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i A_i}{\frac{1}{n}\sum_{i=1}^{n} A_i} \xrightarrow{P} \frac{E[YA]}{E[A]} \ .$$

### 4.2.2 Proving Consistency of LS

We are now ready to prove consistency of the LS estimator in (4.1). To do this, we will use the LLN and the CMT. It turns out that we can prove consistency of the LS estimator without adding additional assumptions. Note that $E[XY] < \infty$ since $XY = XX'\beta + XU$, and both $E[XX']$ and $E[XU]$ exist. Under this assumption, the OLS estimator, $\hat{\beta}_n$ is consistent for $\beta$, i.e.,

$\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To see this, simply note that by the LLN

$$\frac{1}{n}\sum_{1\le i\le n} X_i X_i' \xrightarrow{P} E[XX']$$
$$\frac{1}{n}\sum_{1\le i\le n} X_i Y_i \xrightarrow{P} E[XY]$$

as $n \to \infty$. The desired result therefore follows from the CMT since

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{1\le i\le n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\le i\le n} X_i Y_i\right) \xrightarrow{P} E[XX']^{-1}E[XY] = \beta \ .$$

## 4.3 Asymptotic Normality of LS

So far, we have used the LLN and CMT to establish consistency of $\hat{\beta}_n$. However, when we want to go beyond consistency and study the *asymptotic distribution* of an estimator — for example, to construct confidence intervals or conduct hypothesis tests — the LLN is no longer enough. We need a different tool: the Central Limit Theorem (CLT).

### 4.3.1 A Quick Review of the Central Limit Theorem

It is useful to recall the basic statement and interpretation of the CLT. Suppose $(X_1, \ldots, X_n)$ is an i.i.d. sequence with $E[X] = \mu$ and $\text{Var}[X] < \infty$. Then, as $n \to \infty$,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N\left(0, \text{Var}[X]\right) \ . \tag{4.8}$$

This result tells us that, after appropriate centering (by $\mu$) and scaling (by $\sqrt{n}$), the sample average converges in distribution to a normal random variable. The CLT is central to modern statistical inference because it allows us to approximate the (usually unknown) finite-sample distribution of an estimator with a known limiting distribution.

Note that when the mean of $X$ is zero, i.e., $\mu = 0$, the expression on the LHS of (4.8) simplies as follows :

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \ .$$

Just like with the LLN, most estimators are not simple averages but rather functions of sample averages. Therefore, to move from the convergence in distribution of a sample average (as given by the CLT) to the convergence

in distribution of a function of sample averages, we need a version of the Continuous Mapping Theorem that applies to convergence in distribution.

**Theorem 4.2 (Continuous Mapping Theorem)** *Let $Z_n$ be a random sequence satisfying $Z_n \xrightarrow{d} Z$ and let $g(\cdot)$ be a function that is continuous at all points in the support of $Z$. Then,*

$$g(Z_n) \xrightarrow{d} g(Z) \ .$$

To build some intuition, consider the following setup. Suppose $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are i.i.d. sequences with $E[X] = \mu_X \neq 0$, $E[Y] = 0$, and finite second moments. Then, by the LLN and CLT:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu_X \ ,$$
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \xrightarrow{d} N(0, \sigma_Y^2) \ .$$

- **Product of averages.** Consider

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right) \ .$$

  Since the first factor converges in probability and the second converges in distribution, the CMT implies

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right) \xrightarrow{d} \mu_X \cdot N(0, \sigma_Y^2) = N(0, \mu_X^2 \sigma_Y^2) \ .$$

- **Ratio of averages.** Now consider

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} \ .$$

  Again, the numerator converges in distribution and the denominator converges in probability to a nonzero constant. Applying the CMT yields

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} \xrightarrow{d} \frac{N(0, \sigma_Y^2)}{\mu_X} = N\left(0, \frac{\sigma_Y^2}{\mu_X^2}\right) \ .$$

These examples highlight the usefulness of combining the CLT with the CMT to handle expressions that are more complex than sample means — especially products and ratios, which arise frequently in econometric estimators.

### 4.3.2 Deriving the Limiting Distribution of LS

A usual assumption in the linear regression is to impose the distribution in the $U$ term, in particular a normal distribution. That implies a normal distribution on $Y$ and therefore on $\beta$ and its estimate $\hat{\beta}_n$ since it comes from data on $(Y, X)$. In what follows, we will see that using asymptotics, we don't need to assume normality of $U$: With "large enough" samples, normality will come and we will be able to do inference, with some assumptions.

Suppose $E[XX'] < \infty$ and that $\text{Var}[XU] = E[XX'U^2] < \infty$. Then,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \quad \text{as} \quad n \to \infty \ , \tag{4.9}$$

where

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \ . \tag{4.10}$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left( \frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} X_i U_i \right) \ .$$

The WLLN implies that

$$\frac{1}{n} \sum_{1 \le i \le n} X_i X_i' \xrightarrow{P} E[XX'] \tag{4.11}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} X_i U_i \xrightarrow{d} N(0, \text{Var}[XU]) \tag{4.12}$$

as $n \to \infty$. Thus, the desired result follows from the CMT. Now we have a (limiting) distribution for $\hat{\beta}_n$ to work on and perform statistical difference.

The result in (4.9) is often referred to as the **asymptotic normality** of the OLS estimator. It is important to note that the asymptotic normality of the OLS estimator is a consequence of the LLN and the CLT, and not an assumption on the distribution of $U$. Mechanically, however, the results is often interpreted as if $\hat{\beta}_n$ is normally distributed for large $n$, i.e.,

$$\hat{\beta}_n \approx N\left( \beta, \frac{1}{n} \mathbb{V} \right) \ . \tag{4.13}$$

For the purposes of this class, we will interpret (4.13) as saying that $\hat{\beta}_n$ is approximately normal for large $n$. Later in class, this will make it easy to compute standard errors and perform hypothesis tests.

## 4.4 Key Concepts

- The OLS estimator is unbiased under $E[U|X] = 0$ .
- The OLS estimator is consistent, i.e., $\hat{\beta}_n \xrightarrow{P} \beta$ .
- The Law of Large Numbers (LLN) ensures sample averages converge in probability to population expectations.
- The Continuous Mapping Theorem (CMT) allows us to extend LLN results to functions of sample averages, such as ratios or products.
- The Central Limit Theorem (CLT) gives the asymptotic distribution of scaled sample averages and justifies inference based on normal approximations.
- The OLS estimator is asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \quad \text{as} \quad n \to \infty ,$$

  where $\mathbb{V} = E[XX']^{-1}E[XX'U^2]E[XX']^{-1}$.

## 4.5 Concluding Remarks

These notes are adapted from materials I have used in previous classes and part of the material that was covered in Math 385. Additional related concepts can be found in the books by Bruce Hansen [Hansen, 2022], Jeff Wooldridge [Wooldridge, 2025], and the one by Angrist and Pischke [Angrist and Pischke, 2008].

## 4.6 Problems

**Problem 4.1** *Let* $\{Z_i\}_{i=1}^n$ *be an i.i.d. sequence of random variables with* $E[Z] = 1$. *Define*

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \ .$$

*(a) What is the probability limit of $\bar{Z}_n$?*

*(b) Suppose that instead you were interested in finding the probability limit of $\frac{1}{n}\sum_{i=1}^{n} Z_i^2$. What additional assumption would you need to make in order to find this probability limit?*

**Problem 4.2** *Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be i.i.d. sequences with finite first moments. Define*

$$A_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{and} \quad B_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \ .$$

*(a) TRUE or FALSE: $A_n B_n \xrightarrow{P} E[X]E[Y]$.*

*(b) TRUE or FALSE: $\frac{B_n}{A_n} \xrightarrow{P} \frac{E[Y]}{E[X]}$.*

*(c) TRUE or FALSE: $\frac{B_n}{A_n^2} \xrightarrow{P} \frac{E[Y]}{E[X^2]}$.*

**Problem 4.3** *Consider the linear model $Y_i = X_i\beta + U_i$, where $\{(Y_i, X_i, U_i)\}_{i=1}^n$ are i.i.d., $E[XU] = 0$, and $E[X^2] \neq 0$. This model does not contain an intercept term and $X$ is scalar. The OLS estimator is*

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \ .$$

*(a) Find the limiting distribution of $\hat{\beta}_n$.*

*(b) How does this distribution simplify if we further assume that $E[U \mid X] = 0$ and $E[U^2 \mid X] = \sigma^2$?*

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach 8th ed.* Cengage learning, 2025.

# 5

# *More on Linear Regression*

Previously in class, we learned that the least squares (LS) coefficient has multiple interpretations. When discussing "causal" statements, it's not surprising that we are particularly interested in the third interpretation of LS: its causal interpretation. Generally, it is difficult to claim that the so-called "homogeneous" linear model is causal. After all, very few models in the social sciences lead to a linear relationship between an outcome of interest, $Y$, and some covariates, $X$.

However, we also learned that in one of our running examples, the LS estimate coincides with the difference-in-means estimator, which admits a causal interpretation in that case due to random assignment. A natural question, then, is: when can we draw a connection between LS and the average treatment effect (ATE)? In this chapter, we explore this question further.

## 5.1 Linear Regression with Binary Covariates

Recall that this is the case of Example 2.1 where $A \in \{0, 1\}$ is the perceived race, which is assigned randomly, and $Y$ is the job callback rate. As we discussed in previous lectures, perhaps the easiest way to think about causal relationships is in terms of potential outcomes. The causal relationship between an action $A$ and an outcome $Y$ can be described using the so-called *potential outcomes*:

| | |
|---|---|
| $Y(0)$ | potential outcome in the absence of treatment |
| $Y(1)$ | potential outcome in the presence of treatment |

In other words, we imagine two potential outcome variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 1.

The difference $Y(1) - Y(0)$ is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*. Using

this notation, we may rewrite the observed outcome as

$$\begin{aligned} Y &= AY(1) + (1-A)Y(0) \\ &= E[Y(0)] + (Y(1) - Y(0))A + (Y(0) - E[Y(0)]) \\ &= \gamma_0 + \gamma_1 A + U \ , \end{aligned}$$

where

$$\gamma_0 = E[Y(0)], \quad \gamma_1 = Y(1) - Y(0), \quad \text{and} \quad U = Y(0) - E[Y(0)] \ .$$

If we let $X = (1, A)'$ and $\gamma = (\gamma_0, \gamma_1)'$, we obtain

$$Y = X'\gamma + U \ .$$

This resembles the linear regression model we discussed previously. However, appearances can be misleading because there are important differences. The most significant difference is that the potential outcomes are random variables, meaning that

$$\gamma_1 = Y(1) - Y(0) \ ,$$

the treatment effect, is also **random**. This contrasts with the linear regression model, where the slope $\beta$ is a constant, unknown parameter.

The fact that $\gamma_1$ is random captures the concept of *heterogeneous treatment response*, which suggests that the treatment effect may vary across individuals. For this model to be equivalent to the linear regression model we discussed earlier, we would need to make the additional assumption that the treatment effect is constant across individuals. Specifically, we would require:

$$P\{Y(1) - Y(0) = c\} = 1$$

or, using a subindex $i$ to denote units in the sample,

$$Y_i(1) - Y_i(0) = c \quad \text{for all } i \leq n \ .$$

Under this, arguably strong, additional assumption, we can then write the linear causal model as $Y = X'\beta + U$ with $\beta$ being a constant and $U \perp\!\!\!\perp X$ - since under random assignment the unobserved variable $U = Y(0) - E[Y(0)]$ is independent of the treatment assignment $A$. Notice that, in order to have a linear causal model, a randomized controlled experiment is not enough; we also need *constant treatment effects*.

Understanding the correct interpretation of the regression coefficients in a linear regression model is crucial due to its widespread use in applied research. As discussed above, even in a simple setting with a single binary covariate $A$, the assumption of *homogeneous* treatment effects is necessary for obtaining a causal interpretation. Generally, there is no easy way around this requirement.

However, an alternative approach, which could be described as "reverse engineering," exists. This approach involves estimating the least squares (LS)

coefficient without assuming causality—following, for example, the second interpretation we discussed in the previous class. From there, we attempt to see if this parameter can be expressed as a function of *heterogeneous* treatment effects. The idea is simple: let's continue using our preferred method (regression), but explore whether we can interpret our results in a broader sense without explicitly assuming that the true model for $Y$ is linear and homogeneous.

In general, this second approach has serious limitations, as the LS estimator is not easily interpretable as a function of treatment effects. However, in the special case where $X = (1, A)'$ and $A$ is binary and randomly assigned, the answer turns out to be positive. This will help explain the results we observe when running LS in Example 3.1. Stay tuned.

## 5.2 When is LS equal to the ATE?

Recall that we have a representation of $\beta$ as a function of the distribution of $(Y, X)$, i.e. the population regression result $\beta = E[XX']^{-1}E[XY]$. Again, consider the case where $X = (1, A)'$ and $A$ is a binary treatment satisfying $A \perp\!\!\!\perp (Y(1), Y(0))$. In this case, we may wonder what would the right interpretation for $\beta$ be if we believe that the treatment effect is heterogeneous. In other words, we would like to characterize the least squares estimand $\beta = E[XX']^{-1}E[XY]$ as a function of the treatment effect $Y(1) - Y(0)$. To this end, let $\pi := E[A]$ and note that

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = E[XX']^{-1}E[XY] = E\left[\begin{bmatrix}1\\ A\end{bmatrix}\begin{bmatrix}1 & A\end{bmatrix}\right]^{-1} E\left[\begin{bmatrix}1\\ A\end{bmatrix} Y\ ,\right]$$

which then leads to:

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = E\begin{bmatrix}1 & A\\ A & A\end{bmatrix}^{-1} E\begin{bmatrix}Y\\ AY\end{bmatrix} = \frac{1}{\pi(1-\pi)}\begin{bmatrix}\pi & -\pi\\ -\pi & 1\end{bmatrix}\begin{bmatrix}E[Y]\\ E[AY]\end{bmatrix}\ .$$

The result follows from simple matrix algebra. Below in following chapters we will learn tricks that will allow us to derive the same expression without inverting matrices at all, but for the moment we rely on the formula to invert a 2x2 matrix as a way to find the expression for $\beta_1$. In particular, we have that focusing on $\beta_1$ in the system above leads to:

$$\beta_1 = \frac{E[AY] - \pi E[Y]}{\pi(1-\pi)}\ .$$

This expression, however, is not particularly insightful and our goal is to be able to express $\beta_1$ as a function of the treatment effect $Y(1) - Y(0)$. This expression is a function of $Y$, $A$, and $\pi$.

**Hint 5.1** *Consider a binary random variable $B$. Then $P\{B = 1\} = E[B]$. This is because by definition of the expectation,*

$$E[B] = 1 \cdot P\{B = 1\} + 0 \cdot P\{B = 0\} = P\{B = 1\} \ .$$

*Furthermore, for any function $g(\cdot)$ of $B$,*

$$E[g(B)] = g(1)P\{B = 1\} + g(0)P\{B = 0\} \ .$$

Note that

$$E[AY] = E[E[AY \mid A]] = P\{A = 1\}E[1 \times Y \mid A = 1] = \pi E[Y|A = 1] \ ,$$

where we first use the Law of Iterated Expectations (LIE), followed by the definition of expectation, and finally apply Hint 5.1. Using this expression, we can write $\beta_1$ as follows:

$$\begin{aligned}
\beta_1 &= \frac{\pi E[Y \mid A = 1] - \pi E[Y]}{\pi(1 - \pi)} \\
&= \frac{E[Y \mid A = 1] - E[Y]}{(1 - \pi)} \\
&= \frac{(1 - \pi)E[Y \mid A = 1] - (1 - \pi)E[Y \mid A = 0]}{(1 - \pi)} \\
&= E[Y \mid A = 1] - E[Y \mid A = 0] \ . \qquad (5.1)
\end{aligned}$$

This last expression is still a function of $Y$ and $A$, but it deserves special attention. It states that $\beta_1$, the least squares (LS) slope in a regression of $Y$ on a constant and a dummy variable $A$, equals what we previously referred to as the mean contrast — the difference in the conditional expectations of $Y$ given $A = 1$ and $A = 0$. However, as we discussed in Chapter 1, $E[Y \mid A = a] \neq E[Y(a)]$ in general. Therefore, $\beta_1$ in such a regression is always equal to the mean contrast, but whether it can be expressed as a function of the treatment effect $Y(1) - Y(0)$ depends on additional assumptions.

We have already covered the one assumption that leads to $E[Y \mid A = a] = E[Y(a)]$, which is *random assignment*. Therefore, if we assume that $A \perp\!\!\!\perp Y(a)$ for $a \in \{0, 1\}$, we have:

$$\begin{aligned}
\beta_1 &= E[Y \mid A = 1] - E[Y \mid A = 0] \\
&= E[Y(1) \mid A = 1] - E[Y(0) \mid A = 0] \\
&= E[Y(1)] - E[Y(0)] \ , \qquad (5.2)
\end{aligned}$$

where the last equality follows from random assignment. We conclude that the slope coefficient in a least squares regression of $Y$ on $(1, A)$ identifies $E[Y \mid A = 1] - E[Y \mid A = 0]$, which in turn equals the average treatment effect under random assignment. In other words, we can interpret the coefficient $\beta_1$

causally in the context of a linear regression when the treatment is randomly assigned via a randomized control experiment.

It is important to recognize that we have arrived at a causal interpretation of $\beta_1$ without assuming that the true relationship between $Y$ and $A$ follows a linear, homogeneous model. Instead, we used regression as the best linear predictor and demonstrated that, by coincidence, this slope coefficient equals the ATE. This was not our initial goal, but rather a result that emerged coincidentally from the regression analysis.

This result breaks down quickly (particularly when $A$ takes more than two values), but it is significant enough to highlight. For instance, we can confidently interpret the regression in Example 2.1 on job callbacks and perceived race as identifying the ATE, given that $A$ is binary and randomly assigned, which makes the result derived here applicable.

A few considerations are important about this special case. The linear model in this case is not viewed as a linear causal model (as, for example, in interpretation 3 in the last class). Instead, the linear model is viewed as a convenient way to summarize the joint distribution of $(Y, X)$ (as in interpretation 2 last class, where we proved that the linear model is the best *linear approximation* to $E[Y|X]$), and then it happens to be that this summary statistic (the slope coefficient) is equal to the ATE when the treatment variable $A$ is randomly assigned to units, therefore independent of potential outcomes. Broadly speaking, the linear model with homogenous partial effects is rarely seen as the "true" causal model, but most often the focus is on trying to understand when this estimand, i.e., $\beta$ as in (3.1), can be written as a function (often a weighted average) of well-defined causal effects.

## 5.3 Inference in Linear Regression Models

Let us return to the setting where we observe an i.i.d. sample of size $n$ from the distribution $P$ under the same assumptions used in the previous chapter, and where the OLS estimator $\hat{\beta}_n$ is given by (4.1). In that chapter we proved that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \quad \text{as} \quad n \to \infty \,, \tag{5.3}$$

where

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \,. \tag{5.4}$$

It turns out we can exploit this asymptotic approximation to test hypotheses about $\beta$. Since $\beta$ is a vector, we may want to test whether all components are zero, whether a linear combination equals a particular value, or whether a single component equals a particular value. In this section we focus on the last case. We start by proposing an estimator of $\mathbb{V}$, the asymptotic variance of the LS estimator.

### 5.3.1 Estimation of $\mathbb{V}$

In order to test hypotheses and construct confidence intervals for components of the vector $\beta$, we will require a consistent estimator of $\mathbb{V}$. Note that $\mathbb{V}$ has the so-called sandwich form, with the bread given by $E[XX']^{-1}$ and the meat by $E[XX'U^2]$.

The main principle behind the construction of a consistent estimator of $\mathbb{V}$ is to notice that the expression for $\mathbb{V}$ is a product of three expectations, so if we can consistently estimate each of these expectations, we can consistently estimate $\mathbb{V}$. In turn, a natural way to consistently estimate a given expectation is to use the sample mean of the expression inside the expectation. For example, we can consistently estimate $E[XX']$ by the sample mean of $X_iX_i'$, i.e.,

$$\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i' \xrightarrow{P} E[XX'] \ . \tag{5.5}$$

The case of the "meat", $E[XX'U^2]$ is, however, more complicated as $U_i$ is not directly observed. There are two ways to deal with this: (i) make assumptions to simplify the expression for $\mathbb{V}$ and (ii) replace $U_i$ by $\hat{U}_i$ and use more sophisticated asymptotic arguments (beyond the scope of this class) to show that the resulting estimator is consistent.

Focusing our attention to the meat, we first consider the case where $E[U|X] = 0$ and $\text{Var}[U|X] = \tau^2$ (i.e., under homoskedasticity). Under these conditions,

$$\text{Var}[XU] = E[XX'U^2] = E[XX']\tau^2 \ .$$

Hence,

$$\mathbb{V} = E[XX']^{-1}\tau^2 \ .$$

A natural choice of estimator is therefore

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1} \hat{\tau}_n^2 \ ,$$

where $\hat{\tau}_n^2$ is a consistent estimator of $\tau^2$. This estimator is quite simple, but it unfortunately relies on the assumption that $\text{Var}[U|X] = \tau^2$, which is often violated in applications.

When we do not assume $\text{Var}[U|X] = \tau^2$, a natural choice of estimator is

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1} \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\hat{U}_i^2\right) \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1} \ . \tag{5.6}$$

This estimator is consistent without additional assumptions, i.e.,

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \text{ as } n\to\infty \ , \tag{5.7}$$

regardless of the functional form of $\text{Var}[U|X]$. It is important to note that, by default, a lot of statistics software (such as `Stata` and R) report homoskedastic-only standard errors.

### 5.3.2 Basic testing problem

For a pre-specified scalar $c$, consider testing

$$H_0 : \beta_j = c \text{ versus } H_1 : \beta_j \neq c \ , \tag{5.8}$$

at level $\alpha$. For good or bad reasons, the most popular case of this hypothesis test is when $c = 0$; that is, testing whether a particular component of $\beta$ is equal to zero.

Last class we showed that the LS estimator of $\beta$ is consistent and asymptotically normal. Since this applies to each component of $\beta$, we have that

$$\sqrt{n}(\hat{\beta}_{n,j} - \beta_j) \xrightarrow{d} N(0, \sigma_j^2) \tag{5.9}$$

as $n \to \infty$. Here $\sigma_j^2$ is the (asymptotic) variance of the $j$th component of $\hat{\beta}_{n,j}$, which is none other than the $j+1$th diagonal element of $\mathbb{V}$; i.e.,

$$\sigma_j^2 := \mathbb{V}_{[j+1,j+1]} \ .$$

Dividing by $\sigma_j$ standardizes the left-hand side so that the limiting distribution no longer depends on unknown parameters; this makes the result useful for testing. The CMT then implies that

$$\frac{\sqrt{n}(\hat{\beta}_{n,j} - \beta_j)}{\sigma_j} \xrightarrow{d} N(0,1) \ , \tag{5.10}$$

as $n \to \infty$.

The main intuition behind testing hypotheses goes as follows. Choose a test statistic $T_n$ that satisfies two properties: (i) it is asymptotically $N(0,1)$ (or $|N(0,1)|$) under the null hypothesis, and (ii) it is such that "large values of $T_n$" provide evidence against the null hypothesis. We can then rely on decision rules of the form "reject $H_0$ if $T_n$ is greater than a certain threshold". This threshold value is usually called *critical value*. This rule is used everywhere in statistics, as it leads to tests with good properties provided the critical value is chosen correctly. In this class, we will focus on the case where the test statistic is the absolute value of the t-statistic, which is called the *t-test*.

### 5.3.3 The t-test

The result in (5.10) motivates the use of the following test statistic

$$T_n := |t_{\text{stat}}| \ , \tag{5.11}$$

where $t_{\text{stat}}$ is known as the t-statistic. It is defined as follows:

$$t_{\text{stat}} := \frac{\hat{\beta}_{n,j} - c}{\hat{\sigma}_{n,j}/\sqrt{n}} \ , \tag{5.12}$$

where $\hat{\sigma}_{n,j}$ is a consistent estimator of $\sigma_j$, i.e., $\hat{\sigma}_{n,j} \xrightarrow{P} \sigma_j$ as $n \to \infty$. Given the result in (5.7), a natural candidate for $\hat{\sigma}_{n,j}$ is simply the square root of the diagonal element of $\hat{\mathbb{V}}_n$, i.e.,

$$\hat{\sigma}_{n,j} = \sqrt{\hat{\mathbb{V}}_{n,[j+1,j+1]}} \ . \tag{5.13}$$

The ratio $\hat{\sigma}_{n,j}/\sqrt{n}$ is often called the "standard error" of the $j$th component of $\beta$. It is important to note that this standard error is scaled by $\sqrt{n}$, a feature that is sometimes overlooked by students and may lead to confusion. Pay close attention to this scaling factor when computing standard errors and test statistics.

Note that $T_n$ satisfies both properties mentioned above. First, it is asymptotically $|N(0,1)|$ under the null hypothesis, since $t_{\text{stat}}$ is asymptotically $N(0,1)$ under the null hypothesis. Second, it is such that "large values of $T_n$" provide evidence against the null hypothesis, since large values of $|t_{\text{stat}}|$ happen when $\hat{\beta}_{n,j}$ is "far" from $c$.

Given our choice of test statistic, we are left with the problem of choosing the critical value. Since the test statistic is asymptotically $|N(0,1)|$ under the null hypothesis, it turns out that a convenient choice is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. The rejection rule (or test) is then

$$\text{Reject } H_0 \text{ if } T_n > z_{1-\frac{\alpha}{2}} \ . \tag{5.14}$$

Since $t_{\text{stat}}$ is asymptotically $N(0,1)$ under the null hypothesis, we have that

$$\begin{aligned}
P\{\text{Reject } H_0\} &= P\{T_n > z_{1-\frac{\alpha}{2}}\} \\
&= P\{t_{\text{stat}} > z_{1-\frac{\alpha}{2}}\} + P\{t_{\text{stat}} < -z_{1-\frac{\alpha}{2}}\} \\
&\to 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(-z_{1-\frac{\alpha}{2}}) \\
&= 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(z_{\frac{\alpha}{2}}) \\
&= 1 - (1 - \frac{\alpha}{2}) + \frac{\alpha}{2} \\
&= \alpha \ .
\end{aligned}$$

There is a useful connection between tests and confidence intervals: a $(1-\alpha)$-level confidence interval for $\beta_j$ is the set of all values $c$ that would *not* be rejected by the test above at level $\alpha$. This feature, known as the duality between hypothesis testing and confidence intervals, gives us a confidence interval for each component $\beta_j$ of $\beta$:

$$\begin{aligned}
C_n &= \left\{ c \in \mathbf{R} : \left| \frac{\hat{\beta}_{n,j} - c}{\hat{\sigma}_{n,j}/\sqrt{n}} \right| \leq z_{1-\frac{\alpha}{2}} \right\} \\
&= \left[ \hat{\beta}_{n,j} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{n,j}}{\sqrt{n}} \ , \ \hat{\beta}_{n,j} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_{n,j}}{\sqrt{n}} \right] \ .
\end{aligned}$$

By the previous derivation and Problem 5.5, this confidence region satisfies

$$P\{\beta_j \in C_n\} \to 1 - \alpha \text{ as } n \to \infty \ . \tag{5.15}$$

### 5.3.4 Empirical Example in R

In Example 3.1, we implemented the linear regression estimator of the ATE and found that a typical African-American name on a CV causes a 3.2% difference in the probability of receiving a callback, compared to a typical Caucasian name. By default, R reports the homoskedastic-only estimator when running least squares.

```
    Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.096509   0.005505  17.532  < 2e-16 ***
race_dummy  -0.032033   0.007785  -4.115 3.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2716 on 4868 degrees of freedom
Multiple R-squared:  0.003466,  Adjusted R-squared:  0.003261
F-statistic: 16.93 on 1 and 4868 DF,  p-value: 3.941e-05
```

The table above shows the coefficient estimates, their standard errors (derived from $\hat{\mathbb{V}}_n$), the corresponding t-statistics for testing the null hypothesis $\beta_j = 0$, and the associated p-values.

```
    t test of coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.0965092  0.0059841 16.1277 < 2.2e-16 ***
race_dummy  -0.0320329  0.0077834 -4.1156 3.926e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R can also implement heteroskedastic-robust (HC) variance estimators. To do this, we first need to load the "lmtest" and "sandwich" packages, then use the `coeftest` function. This function takes as arguments the linear model and the covariance matrix, where we specify the degrees of freedom (usually set to "`Inf`" for the normal asymptotic approximation) and the desired type of variance-covariance matrix. See the Code Snippet and output below, where the standard errors differ slightly from the homoskedastic case:

```
# Load necessary libraries
library(dplyr)
library(haven)

# Load libraries for HC robust standard errors:
library(lmtest)
library(sandwich)

# Load the dataset, which is originally in STATA format
# (but R can easily read it)
```

```r
data <- read_dta("lakisha_aer.dta")

# Create an indicator/dummy variable that is 1 if the perceived
    race is A
data$race_dummy <- as.integer(data$race == "b")

# Fit a linear regression model that compares the number of
    callbacks
# between perceived race:
linear_model <- lm(call ~ 1 + race_dummy, data = data)

# Now let's load the results of the linear model fit into
    coeftest() function:
coeftest(linear_model, vcov = vcovHC(linear_model, df = Inf, type
     = "HC0"))
```

Code Snippet 5.1: Implementing HC Robust Variance Estimators

## 5.4 Key Concepts

- **LS slope equals the mean contrast:** In a regression of $Y$ on a constant and a binary variable $A$, the slope $\beta_1$ always equals the mean contrast $E[Y \mid A = 1] - E[Y \mid A = 0]$.
- **LS slope and the ATE:** When $A$ is binary and randomly assigned, $\beta_1 = E[Y(1)] - E[Y(0)]$, the average treatment effect. This causal interpretation does not require assuming a linear or homogeneous causal model.
- **Robust variance estimator:** The EHW estimator $\hat{\mathbb{V}}_n$ in (5.6) is robust to heteroskedasticity and consistently estimates $\mathbb{V}$ without assuming $\text{Var}[U|X]$ is constant. The simpler homoskedastic estimator is valid only when $\text{Var}[U|X] = \tau^2$.
- **The t-test:** To test $H_0 : \beta_j = c$, we use $T_n = |t_{\text{stat}}|$ where $t_{\text{stat}} = (\hat{\beta}_{n,j} - c)/(\hat{\sigma}_{n,j}/\sqrt{n})$. We reject $H_0$ at level $\alpha$ when $T_n > z_{1-\alpha/2}$.
- **Confidence intervals:** A $(1-\alpha)$-level confidence interval for $\beta_j$ collects all values of $c$ not rejected by the t-test: $C_n = [\hat{\beta}_{n,j} \pm z_{1-\alpha/2}\,\hat{\sigma}_{n,j}/\sqrt{n}]$.

## 5.5 Concluding Remarks

These notes are adapted from materials I have used in previous classes and are significantly influenced by Azeem Shaikh, whose notes and our ongoing discussions about teaching have been invaluable. Additional related concepts can be found in the books by Bruce Hansen [Hansen, 2022], Jeff Wooldridge [Wooldridge, 2025], and the one by Angrist and Pischke [Angrist and Pischke, 2008]. The results in Angrist [1998] are related to recent results on treatment effects with delayed outcomes, Bugni et al. [2023], and contamination bias when $A$ is not binary, Goldsmith-Pinkham et al. [2022].

## 5.6 Problems

**Problem 5.1** *We showed that, without assumptions on $A$, the LS coefficient from a regression of $Y$ on a constant and $A$, $\beta_1$, equals $E[Y(1) \mid A = 1] - E[Y(0) \mid A = 0]$. Explain, intuitively, why this quantity may not be considered causal.*

**Problem 5.2** *Consider the setting in Section 5.2. Show that $\beta_0$ equals $E[Y \mid A = a]$ for some $a \in \{0, 1\}$. Under what condition would this coefficient be equal to $E[Y(a)]$?*

**Problem 5.3** *Consider the setting where the binary treatment $A$ is randomly assigned. Suppose you run a regression of $Y$ on $A$ without including a constant term, and denote the slope coefficient $\tilde{\beta}_1$. Is $\tilde{\beta}_1$ equal to the ATE ?*

**Problem 5.4** *Suppose that $A \perp\!\!\!\perp Y(a)$ for $a \in \{0, 1\}$; that is, $A$ is randomly assigned and binary. Let $\hat{\beta}_{n,1}$ be the LS estimate of the slope coefficient of a regression of $Y$ on $X = (1, A)'$.*

*(a) Show that $\hat{\beta}_{n,1}$ equals the difference-in-means estimator $\hat{\theta}_n$ in (2.4).*

*(b) Show that $\hat{\beta}_{1,n}$ is unbiased for the average treatment effect.*

**Problem 5.5** *Prove (5.15).*

**Problem 5.6** *Consider the same setting as in the hypothesis test (5.8), but now with a one-sided alternative:*

$$H_0 : \beta_j = c \quad \textit{versus} \quad H_1 : \beta_j > c \ .$$

*(a) Explain why the test statistic $T_n = |t_{\text{stat}}|$ used for the two-sided test is not well suited for this problem.*

*(b) Propose a test statistic and critical value so that the test has asymptotic level $\alpha$. Verify that the rejection probability converges to $\alpha$ under the null.*

*(c) Using the duality between tests and confidence intervals, derive the one-sided $(1-\alpha)$-level confidence interval for $\beta_j$ implied by your test.*

## Bibliography

J. D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66 (2):249–288, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

F. A. Bugni, I. A. Canay, and S. McBride. Decomposition and interpretation of treatment effects in settings with delayed outcomes. *arXiv preprint arXiv:2302.11505*, 2023.

P. Goldsmith-Pinkham, P. Hull, and M. Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach 8th ed.* Cengage learning, 2025.

# 6

# *Covariate Adjustment in Experiments*

In this lecture, we build on the concepts learned in previous classes about estimating the Average Treatment Effect (ATE) in randomized controlled experiments. We introduce the idea of covariate adjustment, which involves using additional information from covariates (such as age, gender, etc.) to improve the precision of our ATE estimates. While the difference-in-means estimator is simple and effective, incorporating covariates can potentially yield more accurate results, especially when these covariates are correlated with the outcome of interest. We will explore both linear and non-linear models for adjusting for covariates and compare the efficiency of these estimators.

## 6.1 Setup and Problem Formulation

Consider a randomized controlled experiment, where the treatment variable $A$ is independent of the potential outcomes $Y(a)$ for all $a \in \mathcal{A}$. We focus on the case where $A$ is binary, and the interest lies in estimating the Average Treatment Effect (ATE):

$$\theta = E[Y(1) - Y(0)] \ .$$

In the previous class, we learned that the ATE can be estimated using the difference-in-means estimator:

$$\hat{\theta}_n := \frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Y_i - \frac{1}{n_0}\sum_{i\in\mathcal{I}_0} Y_i \ . \tag{6.1}$$

We also learned (see Problem 5.1) that this estimator can be equivalently obtained by running a regression of the observed outcome $Y$ on a constant term and the treatment variable $A$, i.e., $Y = \beta_0 + \beta_1 A + U$, where $\beta = (\beta_0, \beta_1)$ are the best linear projection coefficients. That is,

$$\hat{\beta}_{n,1} = \frac{\sum_{i=1}^{n}(A_i - \bar{A}_n)Y_i}{\sum_{i=1}^{n}(A_i - \bar{A}_n)^2} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) \ . \tag{6.2}$$

Now, suppose that the experiment also has information on other variables.

Let's call these variables "covariates" and denote them by the vector $W$, which takes values in $\mathbf{R}^{d_w}$. The observed variables are then $(Y, A, W)$. A question we may ask ourselves is whether we can obtain a better estimator of the ATE by using the information contained in $W$. To answer this question, we introduce the following additional notation:

$$\mu_a(w) := E[Y(a) \mid W = w] \tag{6.3}$$

to denote the conditional expectations of potential outcomes given $W$.

We then address the question in the context of two separate cases: one where we assume that the true conditional expectations $\mu_a(w)$ are linear in $w$, and another where we do not impose such a restriction.

### 6.1.1 The linear case

Let's start by making the assumption that $\mu_a(w)$ is a linear function of $w$ for each $a \in \{0, 1\}$;

$$\mu_a(w) = \delta_a + W'\gamma_a \ , \tag{6.4}$$

where $\delta_a$ is a scalar and $\gamma_a$ is a $d_w$-dimensional vector of parameters. It is important to understand that $\mu_a(w)$ may not be linear, so (6.4) is indeed an assumption. This assumption is equivalent to assuming a linear model for potential outcomes; i.e.,

$$Y(0) = \delta_0 + W'\gamma_0 + \epsilon_0 \tag{6.5}$$

$$Y(1) = \delta_1 + W'\gamma_1 + \epsilon_1 \tag{6.6}$$

where $E[\epsilon_a \mid W] = 0$ and $E[\epsilon_a] = 0$. This model implies that treatment effect is given by

$$\Delta = Y(1) - Y(0) = \delta_1 - \delta_0 + W'(\gamma_1 - \gamma_0) + \epsilon_1 - \epsilon_0 \ ,$$

and so it is heterogeneous due to the presence of $W'(\gamma_1 - \gamma_0)$ and $\epsilon_1 - \epsilon_0$; both of which are random. Often times researchers distinguish these terms by saying that the term $W'(\gamma_1 - \gamma_0)$ captures "observed" heterogeneity while the term $\epsilon_1 - \epsilon_0$ captures "unobserved" heterogeneity.

It follows from this representation that the ATE can be written as

$$\theta = \delta_1 - \delta_0 + E[W]'(\gamma_1 - \gamma_0) \ . \tag{6.7}$$

This characterization of $\theta$ depends on $W$ in two ways. First, we can see how it depends on the mean of $W$, $E[W]$. But, in addition, it depends on $(\gamma_0, \gamma_1)$, which are the coefficients associated to $W$ in each conditional mean function $\mu_a(w)$.

The expression in (6.7) suggests that we could then estimate $\theta$ by running **two separate** least squares regression of $Y$ on $(1, W)$ for each group:

- **regression 1**: LS of $Y$ on $(1, W)$ for units with $A_i = 1$

- **regression 2**: LS of $Y$ on $(1, W)$ for units with $A_i = 0$

This involves running the regressions suggested by equations (6.6) and (6.5). If we denote the least squares estimators of $(\delta_a, \gamma_a)$ by $(\hat{\delta}_{a,n}, \hat{\gamma}_{a,n})$, then our covariate-adjusted estimator becomes:

$$\hat{\theta}_{n,\text{adj}} = \hat{\delta}_{1,n} - \hat{\delta}_{0,n} + \bar{W}_n'(\hat{\gamma}_{1,n} - \hat{\gamma}_{0,n}) \ , \tag{6.8}$$

where $\bar{W}_n := \frac{1}{n}\sum_{i=1}^n W_i$. It can be shown that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V_{\text{dm}})$$
$$\sqrt{n}(\hat{\theta}_{n,\text{adj}} - \theta) \xrightarrow{d} N(0, V_{\text{adj}})$$

where

$$V_{\text{adj}} \leq V_{\text{dm}} \ . \tag{6.9}$$

Moreover, whenever $\gamma_a \neq 0$, the inequality is strict. This implies that adjusting for covariates leads to a more precise estimator of $\theta$ (at least asymptotically). In other words, using the additional information contained in the covariates reduces the asymptotic error in the linear model. Recall that $\hat{\theta}_n$ is also a regression estimator, but it does not utilize information on $W$.

If $\gamma_0 = \gamma_1 = 0$, then it can be shown that $V_{\text{adj}} = V_{\text{dm}}$, so there are no improvements in using covariates. This should not be surprising, as these values imply that the potential outcomes are mean independent of the covariates.

While we do not formally prove (6.9) in this class, it is important to understand the implications of the result: in a model where the conditional means $\mu_a(w)$ are linear functions of $w$, the estimator that uses information on covariates, $\hat{\theta}_{n,\text{adj}}$, is more precise (asymptotically) than the vanilla difference-in-means estimator, $\hat{\theta}_n$.

### 6.1.2 The non-linear case

The previous result is perhaps not too surprising given that we assumed from the get-go that $\mu_a(w)$ were linear functions of $w$ for each $a \in \{0, 1\}$. That is, if we assume a linear model, then using an estimator that leverages linearity ought to help, right? However, it is possible to prove a much stronger result for OLS in randomized trials: the OLS estimator that introduces covariate-adjustment is never worse (in terms of asymptotic variance) than the difference-in-means estimator (which is OLS of $Y$ on $A$ without covariates). Moreover, most often this estimator strictly improves upon $\hat{\theta}_n$.

We do not provide a proof of this result in this class, but let's try to provide a rigorous statement and intuition. Recall that $\mu_a(w)$ is the conditional expectation of $Y(a)$ given $W = w$, assumed to have a flexible unknown

shape; possibly non-linear. We can still think of this as a "model" of potential outcomes given by

$$Y(0) = \mu_0(w) + \epsilon_0 \tag{6.10}$$
$$Y(1) = \mu_1(w) + \epsilon_1 \tag{6.11}$$

where $E[\epsilon_a \mid W] = 0$ and $E[\epsilon_a] = 0$ by construction (recall Problem 3.4). This model implies that treatment effect is given by

$$\Delta = \mu_1(W) - \mu_0(W) + \epsilon_1 - \epsilon_0 \ ,$$

and so it is heterogeneous due to the presence of $\mu_1(W) - \mu_0(W)$ and $\epsilon_1 - \epsilon_0$; both of which are random. It follows from this representation that the ATE can be written as

$$\theta = E[\mu_1(W) - \mu_0(W)] \ . \tag{6.12}$$

This characterization of $\theta$ depends on $W$ via the two unknown conditional expectations.

If we knew these two conditional expectations, we could estimate $\theta$ using the following natural estimator,

$$\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \left(\mu_1(W_i) - \mu_0(W_i)\right) \ . \tag{6.13}$$

This estimator is, of course, infeasible. The idea then would be to provide the best linear approximation to these conditional means and use them in place of the true, unknown, conditional means. To accomplish this, we still run **two separate** least squares regression of $Y$ on $(1, W)$ for each group:

- **regression 1**: LS of $Y$ on $(1, W)$ for units with $A_i = 1$
- **regression 2**: LS of $Y$ on $(1, W)$ for units with $A_i = 0$

If we denote by $(\delta_1^*, \gamma_1^*)$ the coefficients of the first regression, note that these coefficients would be interpreted as

$$\begin{aligned}(\delta_1^*, \gamma_1^*) &= \min_{(\delta,\gamma)\in\mathbf{R}^{k+1}} E[(\mu_1(W) - \delta - W'\gamma)^2] \\ &= \min_{(\delta,\gamma)\in\mathbf{R}^{k+1}} E[(Y(1) - \delta - W'\gamma)^2] \ ,\end{aligned}$$

since when $A = 1$, we know that $Y = Y(1)$. The same applies to $(\delta_0^*, \gamma_0^*)$ since $Y = Y(0)$ when $A = 0$.

Denote the least squares estimators of $(\delta_a^*, \gamma_a^*)$ by $(\hat{\delta}_{a,n}^*, \hat{\gamma}_{a,n}^*)$, and the best linear predictor of $Y(a)$ given $W$ by

$$BLP_a(W_i) = \hat{\delta}_{a,n}^* + W_i'\hat{\gamma}_{a,n}^* \ .$$

The covariate-adjusted estimator would then replace the unknown conditional means with these best linear predictions, which leads to

$$\begin{aligned}
\hat{\theta}^*_{n,\text{adj}} &= \frac{1}{n}\sum_{i=1}^{n}\left(BLP_1(W_i) - BLP_0(W_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((\hat{\delta}^*_{1,n} + W_i'\hat{\gamma}^*_{1,n}) - (\hat{\delta}^*_{0,n} + W_i'\hat{\gamma}^*_{0,n})\right) \\
&= \hat{\delta}^*_{1,n} - \hat{\delta}^*_{0,n} + \left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)'(\hat{\gamma}^*_{1,n} - \hat{\gamma}^*_{0,n}) \\
&= \hat{\theta}_{n,\text{adj}} \ .
\end{aligned}$$

It follows that the covariate-adjusted estimator in the general case is *identical* to $\hat{\theta}_{n,\text{adj}}$. The reason we use different notation to denote these two estimators, despite them being the same, is to emphasize that they both admit drastically different interpretation. While $\hat{\theta}_{n,\text{adj}}$ has properties that are determined by the linearity assumption, $\hat{\theta}^*_{n,\text{adj}}$ does not rely on such assumption but has different asymptotic properties as a result. However, we still get the following result:

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta) &\xrightarrow{d} N(0, V_{\text{dm}}) \\
\sqrt{n}(\hat{\theta}^*_{n,\text{adj}} - \theta) &\xrightarrow{d} N(0, V^*_{\text{adj}})
\end{aligned}$$

where

$$V^*_{\text{adj}} \leq V_{\text{dm}} \ . \tag{6.14}$$

Moreover, whenever $\gamma^*_a \neq 0$, the inequality is strict. On the other hand, if $\gamma^*_0 = \gamma^*_1 = 0$, then it can be shown that $V^*_{\text{adj}} = V_{\text{dm}}$, so there are no improvements in using covariates.

In other words, whether or not the true conditional expectation function $\mu_a(w)$ is linear, estimating the ATE using OLS with covariates always reduces the asymptotic variance relative to the difference-in-means estimator. Moreover, the amount of variance reduction scales by how well OLS approximates the conditional expectations.

To recap, the individual treatment effect $\Delta = Y(1) - Y(0)$ is a central object of interest in causal inference. These effects $\Delta$ themselves are fundamentally unknowable; however, a randomized controlled trial lets us consistently recover the average treatment effect $\theta = E[\Delta]$. Moreover, even without assuming linearity, OLS covariate-adjustments generally improve on the performance of the simple difference-in-means estimator.

## 6.2 How to Not Adjust for Covariates

Throughout our analysis, we used two separate regressions—one for each treatment group—rather than a single pooled regression of $Y$ on $(1, A, W)$ with a common slope on $W$:

$$Y = \beta_0^\star + \beta_1^\star A + W'\beta_2^\star + V \ . \tag{6.15}$$

It is natural to ask why. Under random assignment, the coefficient $\beta_1^\star$ in (6.15) does in fact equal the ATE $\theta$, so the pooled regression correctly identifies the treatment effect. The issue is *efficiency*: the variance reductions in (6.9) and (6.14) rely on allowing $W$ to have different slopes in each treatment group, which is what the arm-specific regressions achieve. By forcing a common slope on $W$, the pooled regression (6.15) generally cannot achieve the optimal variance reduction. Even though it gets the right answer on average, it produces a less precise estimator than necessary.

At the same time, the arm-specific adjustment has a convenient one-step implementation via an *interacted* regression. Consider the population linear projection of $Y$ on $(1, A, W, AW)$:

$$Y = \beta_0 + A\beta_1 + W'\beta_2 + AW'\beta_3 + U \ , \tag{6.16}$$

where the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ are best linear prediction (BLP) coefficients. Under random assignment and the normalization $E[W] = 0$, one can show that

$$\beta_2 = \gamma_0^* \ , \qquad \beta_3 = \gamma_1^* - \gamma_0^* \ , \qquad \beta_1 = \delta_1^* - \delta_0^* = \theta \ .$$

More importantly, we can show that the LS estimator of $\beta_1$ in (6.16) is identical to the covariate-adjusted estimator $\hat{\theta}_{n,\text{adj}}$. We will not provide a proo here.

A key practical considertation concerns the covariates, $W$.The centering $E[W] = 0$ is essential for the above manipulations. In practice, the interacted regression is implemented by first centering the covariates by subtracting the sample mean; that is, we replace $W$ with $W - \bar{W}_n$ before running the interacted regression. We illustrate this in the next section.

Finally, note that $\hat{\theta}_{n,\text{adj}}$ (or $\hat{\theta}^*_{n,\text{adj}}$, as they are the same) can be written as

$$\hat{\theta}_{n,\text{adj}} = \frac{1}{n}\sum_{i=1}^{n} \left( \underbrace{(\hat{\delta}_{1,n} + W_i'\hat{\gamma}_{1,n})}_{\hat{\mu}_{1,n}(W)} - \underbrace{(\hat{\delta}_{0,n} + W_i'\hat{\gamma}_{0,n})}_{\hat{\mu}_{0,n}(W)} \right) \tag{6.17}$$

where $\hat{\mu}_{a,n}(w)$ denotes OLS predictions at $w$. Could we use other methods to estimate $\hat{\mu}_{a,n}(w)$ rather than OLS (e.g., deep nets, random forests, Lasso, or other machine learning methods)? The answer is yes and we will discuss some of these later in class. But the main mechanical ideas on how to correctly

implement these ideas can all be illustrated with OLS and do not fundamentally change when we use more sophisticated ways (i.e., machine learning) of estimating conditional expectations.

## 6.3 Empirical Illustration

A common strategy among companies is to offer a low-cost product that isn't necessarily profitable, but serves as a gateway to attract new customers. Once the company has acquired these customers, it can then cross-sell more profitable products. Suppose you work for a coffee delivery company. The main product is a low-cost monthly subscription, allowing customers to receive high-quality, curated coffee delivered weekly. Beyond this basic and affordable subscription, your company offers a premium option, which includes brewing perks and the world's finest coffee, such as that from local producers in Brazil. This premium service is by far your most profitable, and your goal is to increase its sales among customers who have already subscribed to the low-cost, entry-level product.

To achieve this, your company has a marketing team that randomly selects customers and assigns them to one of the following three treatments:

- No email
- A long, beautifully designed email about the premium subscription
- A short and direct email about the premium subscription

To illustrate the ideas in the previous section, we focus on the causal effect of a short email relative to no-email, and ignore the information contained in sending a long email (something you have studied in your problem sets). To do this, let $A_l$ equal 1 when the customer receives a long email and 0 otherwise, and $A_s$ equal 1 when the customer receives a short email and 0 otherwise. Our entire analysis then restricts attention to the sample with $A_l = 0$.

Let's start with the difference-in-means estimator $\hat{\theta}_n$ using the data set "`cross_sell_email.csv`" in Canvas. We run the following code:

```r
# Load the necessary libraries
library(dplyr)

# Load the data
data <- read.csv("cross_sell_email.csv")
head(data)

# Keep only control and short email groups
data_s <- data %>%
  filter(cross_sell_email %in% c("short", "no_email"))

# Treatment indicator (A_s = 1 if "short" email, 0 if "control")
```

```r
data_s$A_s <- ifelse(data_s$cross_sell_email == "short", 1, 0)

# Run the regression of conversion on A_s
model_A_s <- lm(conversion ~ A_s, data = data_s)

# View summary of the regression model
summary(model_A_s)
```

Code Snippet 6.1: Difference-in-means

This code leads to the following value for our estimator: $\hat{\theta}_n = 0.08245$. What about if we use information on covariates? In this data set there are two covariates : age and gender. Let's call them $W$ and compute $\hat{\theta}_{n,\text{adj}}$ by running the two regressions previously described using the following code:

```r
# Covariate Adjustment
# Compute mean of covariates (age and gender)
W_mean <- colMeans(data_s %>% select(age, gender), na.rm = TRUE)

# Run regressions for treated (A_s = 1) and control (A_s = 0)
    groups
fit_t <- lm(conversion ~ age+gender, data = subset(data_s, A_s ==
     1))
fit_c <- lm(conversion ~ age+gender, data = subset(data_s, A_s ==
     0))

# Extract estimated coefficients
delta_1_hat <- coef(fit_t)[1]  # Intercept for treated
gamma_1_hat <- coef(fit_t)[-1] # Slope in treated group

delta_0_hat <- coef(fit_c)[1]  # Intercept for control
gamma_0_hat <- coef(fit_c)[-1] # Slope in control group

# Compute the adjusted treatment effect
theta_adj <- delta_1_hat - delta_0_hat
             + sum(W_mean * (gamma_1_hat - gamma_0_hat))

# Print results
cat("Covariate-Adjusted Estimator:", theta_adj, "\n")
```

Code Snippet 6.2: covariate-adjusted

This code leads to the following value for our estimator: $\hat{\theta}_{n,\text{adj}} = 0.08456$. This value can be alternatively obtained by running a regression with interactions, as follows.

```r
# Obtain the same estimator by using model with interactions
# Center the covariates to ensure E[W] = 0
data_s <- data_s %>%
  mutate(
    age_centered = age - mean(age, na.rm = TRUE),
    gender_centered = gender - mean(gender, na.rm = TRUE)
  )

# Run a single regression with interaction terms
fit_inter <- lm(conversion ~ A_s * (age_centered + gender_
    centered),
```

```
        data = data_s)

# View summary of the regression model
summary(fit_inter)
```

Code Snippet 6.3: Interactions

We conclude by noting that $\hat{\theta}_n = 0.08245$ and $\hat{\theta}_{n,\text{adj}} = 0.08456$ are very similar in this particular application, but not identical. If one were to compare their standard errors (not shown), we would see that these are also very similar. It appears to be the case that covariates in this case do not help improve efficiency much.

## 6.4 Key Concepts

- **Covariate-adjusted estimator:** estimating the ATE with covariates requires running two separate regressions of $Y$ on $(1, W)$, one for each treatment group, and then combining the estimates via $\hat{\theta}_{n,\text{adj}} = \hat{\delta}_{1,n} - \hat{\delta}_{0,n} + \bar{W}_n'(\hat{\gamma}_{1,n} - \hat{\gamma}_{0,n})$.
- **Efficiency gain from covariate adjustment:** $V_{\text{adj}} \leq V_{\text{dm}}$, with strict inequality whenever $W$ is linearly related to the potential outcomes. This holds whether or not the true conditional expectations $\mu_a(w)$ are linear.
- **Why not pool:** while a pooled regression of $Y$ on $(1, A, W)$ does correctly identify $\theta$ under random assignment, it does *not* achieve the optimal variance reduction because it forces $W$ to have the same coefficient in both treatment groups.
- **Interaction terms:** the covariate-adjusted estimator can be equivalently obtained from a single regression of $Y$ on $(1, A, W, AW)$ with demeaned covariates (so that $E[W] = 0$), where the coefficient on $A$ equals the ATE.

## 6.5 Concluding Remarks

The contents of this chapter are based on the notes by Ding [2023], the notes by Wager [2020], and the book by Facure [2023]. The book by Angrist and

Pischke [2008] is another reference that includes discussion on many of the topics we covered.

## 6.6 Problems

**Problem 6.1** *Re-do the empirical application in Section 6.3 focusing on the ATE of sending a long email versus no email.*

**Problem 6.2** *Show that $\hat{\theta}_n$ has conditional variance equal to :*

$$\text{Var}[\hat{\theta}_n \mid A_1, \ldots, A_n] = \frac{\text{Var}[Y(1)]}{n_1} + \frac{\text{Var}[Y(0)]}{n_0} \ .$$

**Problem 6.3** *For any random variable $Z$, let $\bar{Z}_{n,a} = \frac{1}{n_a}\sum_{i=1}^{n} Z_i I\{A = a\}$ be the average for units with $A = a$. Assume that the true conditional expectations are linear, so that*

$$\mu_a(w) = \delta_a + W'\gamma_a \ ,$$

*and that $\bar{W}_n = 0$ - the covariates are re-centered. Show that*

$$\hat{\theta}_{n,\text{adj}} = \hat{\theta}_n - (\bar{W}'_{n,1}\hat{\gamma}_{1,n} - \bar{W}'_{n,0}\hat{\gamma}_{0,n}) \ .$$

**Problem 6.4** *Assume the linear model for potential outcomes in (6.5)–(6.6) and that $E[W] = 0$.*

*(a) Starting from $Y = AY(1) + (1 - A)Y(0)$, show that $Y$ can be written as in (6.16) and verify the expressions for $\beta_0, \beta_1, \beta_2, \beta_3$, and $U$.*

*(b) Show that $E[U \mid A, W] = 0$.*

*(c) Explain why part (b) implies that $\beta_1$ in (6.16) can be interpreted as the ATE, and why this interpretation would fail without the interaction term $AW'\beta_3$.*

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

M. Facure. *Causal Inference in Python*. O'Reilly Media, Inc., 2023.

S. Wager. Causal inference. Standford University, 2020.

# 7

# *Selection on Observables*

In this chapter, we shift our attention to observational studies. These are settings where the treatment $A$ is not randomly assigned but, rather, is chosen by the units (individuals, families, firms, etc.). The key implication of an observational study is that the treatment, $A$, is typically **not** independent of the potential outcomes $Y(a)$. We will explore why this is the case below.

For our discussion, we use the following notation. We observe a binary treatment $A$, a real-valued outcome of interest $Y$, and some additional *pre-treatment* covariates, denoted by $W$, which take values in $\mathbf{R}^{d_w}$. Therefore, the observed variables are $(Y, A, W)$. Note that the term *pretreatment covariates* emphasizes that $W$ represents variables that are realized *before* the treatment is chosen or assigned.

Throughout the lecture, we will use the following notation:

$$\mu_a(w) := E[Y(a) \mid W = w] \tag{7.1}$$

to denote the conditional expectations of potential outcomes given $W$,

$$\sigma_a^2(w) := \text{Var}[Y(a) \mid W = w] \tag{7.2}$$

to denote the conditional variances of potential outcomes given $W$, and

$$\pi(w) := E[A \mid W = w] = P\{A = 1 \mid W = w\} \tag{7.3}$$

to denote the so-called propensity score.

The goal of this and the next chapter is to estimate causal parameters of interest (in particular, the ATE) under the assumption that, conditional on $W$, $A$ is "as-good-as-random" even if we do not control it. We will first review the assumptions and their consequences for identification, then examine three methods for estimating average treatment effects (ATE): matching (this chapter), regression, and propensity score weighting (next chapter).

## 7.1 Observational Studies and Selection Bias

Observational studies are studies in which the researcher does not control the assignment of the treatment. The researcher observes both the treatment and

the outcome and attempts to infer the causal effect of the treatment on the outcome. One immediate consequence of dealing with observational data is that the crucial assumption of treatment exogeneity, i.e.,

$$A \perp\!\!\!\perp (Y(a) : a \in \mathcal{A}) \ ,$$

is often difficult to defend.

The inability to control the assignment of the treatment does not mean that we cannot consider the following question: how would the study be conducted if it were possible to carry it out through controlled experimentation? Thought experiments are valuable for properly characterizing the counterfactual question of interest and for clarifying the assumptions required to answer it in terms of potential outcomes. Indeed, as we will see throughout this chapter, many of the ideas in causal inference with observational studies are deeply connected to those used in randomized experiments.

**Example 7.1 (Job Training Program)** LaLonde (1986) was interested in the causal effect of a job training program on earnings. He compared the results based on a randomized experiment to the results based on observational studies. LaLonde (1986) found that many traditional econometric methods for observational studies gave *quite different* estimates compared to the estimates based on the experimental data. Dehejia and Wahba [1999] re-analyzed the data using methods motivated by causal inference, and found that those methods can recover the experimental gold standard. Since then, this became a canonical example in causal inference with observational studies. ■

As we discussed in Chapter 2.3, random assignment is rarely compelling with observational data where agents choose $A$ typically maximizing some criterion function (utility, profits, etc.). At the time, we defined selection into the treatment state $A$ if

$$Y(a)|A = a \text{ is distributed } \textbf{differently} \text{ from } Y(a)|A = a' \text{ for } a \neq a' \ .$$

Consider the following concrete examples.

**Example 7.2 (Educational Attainment and Income)** Consider the effect of pursuing higher education (e.g., attending college) on income. People who choose to attend college may differ systematically from those who do not, such as being more motivated or having more financial resources. This means that $Y(0)$ may be systematically higher for those choosing $A = 1$ relative to those choosing $A = 0$. ■

**Example 7.3 (Health Interventions and Health Outcomes)** Consider a study examining the effect of a particular medical treatment (e.g., a new drug or therapy) on patient recovery rates. In an observational setting, patients who choose or are selected to receive the treatment might have different characteristics from those who do not, such as being sicker. This means that $Y(0)$ may

be systematically lower for those choosing $A = 1$ relative to those choosing $A = 0$. ■

**Example 7.4 (Subscription Services and Upselling)** In tech companies offering subscription-based services (e.g., streaming platforms, cloud services), there is often selection into treatment when customers opt for premium plans or additional services. For instance, suppose a company offers a basic subscription (e.g., Spotify's free plan) and a premium subscription (e.g., Spotify's premium plan). The selection into treatment occurs when customers who are more engaged with the platform or who have more disposable income are more likely to subscribe to the premium plan (the treatment). These customers may also be more likely to spend more money on additional services (e.g., buying concert tickets, purchasing exclusive content); the outcome. The issue here is that the customers who choose the premium plan may already have a systematically higher $Y(0)$, because of higher income or a greater affinity for the product, which may influence their likelihood to spend more. Thus, the company may observe higher spending among premium subscribers, but it could be difficult to separate the effect of the premium plan itself from the inherent differences in the types of customers who select into it. ■

For now we will focus on mean effects, and the parameter of interest could be the average treatment effect (ATE), $\theta := E[Y(1) - Y(0)]$, the average treatment effect on the treated (ATT),

$$\theta_{\mathrm{t}} := E[Y(1) - Y(0) \mid A = 1] \ ,$$

or the average treatment effect on the untreated (ATU),

$$\theta_{\mathrm{u}} := E[Y(1) - Y(0) \mid A = 0] \ .$$

For the time being, we ignore covariates and focus on the ATT and ATU. From the relationship $Y = Y(A)$, we immediately derive the following expressions for the ATT and ATU:

$$\begin{aligned} \theta_{\mathrm{t}} &= E[Y(1) \mid A = 1] - E[Y(0) \mid A = 1] \\ &= E[Y \mid A = 1] - E[Y(0) \mid A = 1] \ , \end{aligned}$$

and

$$\begin{aligned} \theta_{\mathrm{u}} &= E[Y(1) \mid A = 0] - E[Y(0) \mid A = 0] \\ &= E[Y(1) \mid A = 0] - E[Y \mid A = 0] \ . \end{aligned}$$

In these expressions for $\theta_{\mathrm{t}}$ and $\theta_{\mathrm{u}}$, the quantities $E[Y \mid A = 1]$ and $E[Y \mid A = 0]$ are *identified* from the observed data - that is, they are functions of the observed data. However, the counterfactual quantities $E[Y(0) \mid A = 1]$ and $E[Y(1) \mid A = 0]$ are *not* directly observable. These are called **counterfactuals**, as they represent the expected values of the potential outcomes corresponding to the treatment level that was *not* actually received.

We can then decompose the usual difference in means contrast in terms of the ATT or the ATU to obtain:

$$\begin{aligned} E[Y \mid A=1] - E[Y \mid A=0] &= \theta_{\text{t}} + E[Y(0) \mid A=1] - E[Y(0) \mid A=0] \\ &= \theta_{\text{u}} + E[Y(1) \mid A=1] - E[Y(1) \mid A=0] \ . \end{aligned}$$

This shows that the difference in means contrast is biased for either the ATT or the ATU in settings in which we cannot guarantee that $A \perp\!\!\!\perp (Y(1), Y(0))$. The bias terms, known as "selection bias", measure the differences in the means of the potential outcomes across the treatment and control groups, which would be expected to be different when agents choose $A$ with knowledge of the potential outcomes.

**Example 7.5 (Educational Attainment (cont.))** In this case, the selection bias term $E[Y(0) \mid A=1] - E[Y(0) \mid A=0]$ would be expected to be positive because the individuals who choose to attend college (those with $A=1$) are likely to have higher income potential even before considering the effect of education. The bias term here captures these differences in pre-treatment income potential, which would affect the observed outcomes. ■

**Example 7.6 (Health Interventions (cont.))** For a study on health outcomes, the selection bias term $E[Y(0) \mid A=1] - E[Y(0) \mid A=0]$ would be negative when sicker patients choose to receive a treatment more often. The bias term here captures these differences in pre-treatment health status, which would affect the observed outcomes. ■

## 7.2 Selection on Observables

Perhaps the simplest and most immediate relaxation of random assignment is to assume that the assignment of the treatment is independent of the potential outcomes conditional on the observed covariates $W$. This is known as selection on observables, conditional independence, strong ignorability, or unconfoundedness. We formally define it as follows:

**Assumption 7.1 (unconfoundedness)**

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid W \ . \tag{7.4}$$

The assumption states that conditional on $W$, treatment is as-good-as randomly assigned. This means that there could be selection into the treatment state $A$, but the selection is entirely driven by the covariates $W$: both $A$ and the potential outcomes may depend on $W$, but once we condition on $W$, no

residual confounding remains. There is an alternative version of this assumption that only requires $Y(a) \perp\!\!\!\perp A \mid W$ for all $a \in \mathcal{A}$, but for all practical purposes, the distinction between these two is immaterial, and so we work with (7.4).

**Example 7.7 (Educational Attainment and Income)** In this example, the treatment is whether an individual pursues higher education, and the outcome is their income. The observed covariates $W$ could include variables such as parental income, prior academic performance, or socio-economic status. The assumption of unconfoundedness implies that, conditional on these covariates, the decision to attend college is as-good-as-random. That is, after controlling for factors like parental income and academic performance, any remaining differences between college-goers and non-goers are **not** systematically related to their potential earnings. ■

**Example 7.8 (Subscription Services and Upselling)** In the context of subscription services, such as streaming platforms, $W$ might include demographic variables like age, income, or usage frequency (e.g., hours spent on the platform). The unconfoundedness assumption implies that, after controlling for these covariates, the decision to subscribe to a premium plan (the treatment) is as-good-as-random. This would mean that the observed differences in spending behavior between premium and non-premium subscribers with the same age, income, and usage frequency, are solely due to the effect of the premium plan. ■

Now that we hopefully understand what the assumption means, let's study its identifying power. Assumption 7.1 implies that

$$E[Y(a) \mid A = a', W] = E[Y(a) \mid W] \tag{7.5}$$

for all $a, a' \in \mathcal{A}$. As a result,

$$\begin{aligned} E[Y(0) \mid A = 1, W] - E[Y(0) \mid A = 0, W] = 0 \\ E[Y(1) \mid A = 1, W] - E[Y(1) \mid A = 0, W] = 0 \ . \end{aligned}$$

In other words, the differences in the means of the potential outcomes across the treatment and control groups are entirely due to the difference in the observed covariates $W$. So, given the same value of the covariates, the potential outcomes have the same means across the treatment and control groups.

The assumption in (7.4) delivers identification of the conditional expectations of potential outcomes via the conditional expectations of the observed outcomes conditional on $(A, W)$. That is, for all $a \in \mathcal{A}$,

$$\begin{aligned} \mu_a(w) &:= E[Y(a) \mid W = w] \\ &\overset{(1)}{=} E[Y(a) \mid A = a, W = w] \\ &\overset{(2)}{=} E[Y \mid A = a, W = w] \ , \end{aligned} \tag{7.6}$$

where (1) follows by (7.4), and (2) follows because $Y = Y(a)$ when $A = a$. Given the equality in (7.6), we use $\mu_a(w)$ to denote either $E[Y(a) \mid W = w]$ or $E[Y \mid A = a, W = w]$ in what follows.

Perhaps not surprisingly, the result in (7.6) implies that the ATE, ATT, and ATU are all identified in observational studies under selection on observables. In fact, it turns out that the *conditional mean contrast* $\Delta(W) := E[Y \mid A = 1, W] - E[Y \mid A = 0, W]$ identifies the *conditional* average treatment effect, defined as:

$$\theta(W) := E[Y(1) - Y(0) \mid W] = \mu_1(W) - \mu_0(W) \ . \tag{7.7}$$

We can easily prove this result using (7.6) as follows,

$$\begin{aligned} \Delta(W) &:= E[Y \mid A = 1, W] - E[Y \mid A = 0, W] \\ &= E[Y(1) \mid A = 1, W] - E[Y(0) \mid A = 0, W] \\ &= E[Y(1) \mid W] - E[Y(0) \mid W] \\ &= E[Y(1) - Y(0) \mid W] \\ &= \theta(W) \ . \end{aligned} \tag{7.8}$$

By the law of iterated expectations (LIE), this result automatically implies identification of the ATE $\theta$ under selection on observables,

$$\begin{aligned} E[\Delta(W)] &= E[\theta(W)] \\ &= E[E[Y(1) - Y(0) \mid W]] \\ &= E[Y(1) - Y(0)] \\ &= \theta \ , \end{aligned} \tag{7.9}$$

where the first equality follows from our proof (7.8), and the third by applying the LIE. This last expression motivates the two alternative approaches to estimate $\theta$ that we will discuss below: matching and regression. These approaches exploit Assumption 7.1 to identify the ATE and related parameters.

An important implication of selection on observables is that it imposes that the conditional versions of the treatment effects we previously discussed, i.e.,

$$\begin{aligned} \theta(W) &:= E[Y(1) - Y(0) \mid W] \\ \theta_{\mathrm{t}}(W) &:= E[Y(1) - Y(0) \mid W, A = 1] \\ \theta_{\mathrm{u}}(W) &:= E[Y(1) - Y(0) \mid W, A = 0] \ , \end{aligned}$$

are all the same,

$$\theta(W) = \theta_{\mathrm{t}}(W) = \theta_{\mathrm{u}}(W) \ .$$

This is not true for the unconditional counter-parts, since $\theta = E[\theta(W)]$ integrates over the distribution of $W$, $\theta_{\mathrm{t}} = E[\theta_{\mathrm{t}}(W) \mid A = 1]$ integrates using the conditional distribution $W|A = 1$, and $\theta_{\mathrm{u}} = E[\theta_{\mathrm{u}}(W) \mid A = 0]$ integrates using the conditional distribution $W|A = 0$. Since treatment is not randomly assigned, we expect all of these distributions to be different and so, in general, $\theta$, $\theta_{\mathrm{t}}$, and $\theta_{\mathrm{u}}$ are all different from each other.

## 7.3 Estimation of the ATE

Let $(Y, A, W)$ be a random vector where $Y$ takes values in $\mathbf{R}$, $A$ takes values in $\mathcal{A} = \{0, 1\}$, and $W$ takes values in $\mathbf{R}^{d_w}$. We denote by $P$ the distribution of $(Y, A, W)$ and assume we have access to a random sample of size $n$ from $P$ that we denote by

$$\{(Y_i, A_i, W_i) : 1 \leq i \leq n\} .$$

We assume unconfoundedness as in (7.4) and overlap, which we can state:

**Assumption 7.2 (Overlap)**

$$0 < P\{A = 1|W = w\} < 1, \quad a.s. \tag{7.10}$$

In words, overlap states that for each value of $W$, there are treated and control units with positive probability. This assumption rules out regions of the covariate space where all units are treated or all are untreated. Without it, $\mu_a(w) = E[Y \mid A = a, W = w]$ cannot be estimated at values of $w$ where no units with $A = a$ are observed, and the ATE is no longer identified.

Our goal is to estimate the ATE $\theta := E[Y(1) - Y(0)]$ or the ATT $\theta_{\mathrm{t}} := E[Y(1) - Y(0) \mid A = 1]$. We do so in three different ways (matching, regression, and weighting), one of which we discuss in this chapter.

### 7.3.1 Matching

Matching estimators are more easily understood when the covariates $W$ are discrete, taking values from a finite set $\mathcal{W}$ (think of $\mathcal{W}$ as categories, such as gender, age, or income brackets). In this case, the identification of $\theta$ follows from equation (7.9), with a similar argument applying to $\theta_{\mathrm{t}}$.

Intuitively, the assumption of selection on observables implies that, for each group of units characterized by the same value of $W$, we can compare treated and control units as if we had an experiment. That is, while we cannot compare all treated units with all control units directly — due to the presence of selection — we can indeed make such comparisons within groups that share the same values of $W$. For example, if $W$ includes age and gender only, a valid group would be units that share the same gender and age. Matching estimators exploit this intuition.

Mechanically, it is useful to go back to the proof in (7.9). There we proved that $\theta = E[\Delta(W)]$, so if we now use that $W$ is discrete, we obtain

$$\begin{aligned}
\theta &= E[\Delta(W)] \\
&= E\left[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]\right] \\
&= \sum_{w \in \mathcal{W}} \big(E[Y|A = 1, W = w] - E[Y|A = 0, W = w]\big) P\{W = w\} .
\end{aligned}$$

This formula shows that, for each $w \in \mathcal{W}$, we can compute a CATE via $\theta(W)$, to then aggregate back to the ATE by using the distribution of covariates $P\{W = w\}$.

Matching estimators are simply sample analogs of the previous derivation. To define this estimator properly, let

$$\hat{p}_n(w) := \frac{1}{n}\sum_{i=1}^{n} I\{W_i = w\}$$

denote the estimator of $p(w) := P\{W = w\}$ for all $w \in \mathcal{W}$, and let

$$n_{a,w} := \sum_{i=1}^{n} I\{A_i = a, W_i = w\}$$

denote the number of observations with $A = a$ and $W = w$. With this notation, the natural sample analog of $\theta$ given our population proof above, with discrete $W$, is given by

$$\hat{\theta}_{n,\text{mat}} := \sum_{w \in \mathcal{W}} \hat{p}_n(w)(\bar{Y}_{1,w} - \bar{Y}_{0,w}) \ , \tag{7.11}$$

where

$$\bar{Y}_{a,w} := \frac{1}{n_{a,w}}\sum_{i=1}^{n} Y_i I\{A_i = a, W_i = w\} \ .$$

In words, $\hat{\theta}_{n,\text{mat}}$ is a weighted average of the within-cell $W = w$ differences in averages between units that are treated and untreated. It is called a "matching" estimator because we "match" treated and untreated observations within each cell determined by $W$.

The beauty and simplicity of matching estimators when $W$ includes continuously distributed random variables quickly disappears. In this case, matching estimators impute the missing potential outcomes by using only the outcomes of nearest neighbors in the opposite treatment group. That is, the main idea is to find the "closest match" in the treatment group ($A = 1$) and control group ($A = 0$) with the same (or as close as possible) value of $W$, i.e., $W = w$. In that respect, matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in kernel regression. A point worth mentioning is that matching does not exhibit good asymptotic properties with many continuous covariates and, perhaps partly due to this reason, matching is rarely viewed as the go-to method in moderate to high-dimensional settings.

## 7.4 Empirical Illustration

Angrist [1998] studies the causal effects of voluntary military services in the

US on the later earning of soldiers. In this application, $Y$ is a labor market outcome like employment or earnings, the treatment $A$ denotes veteran status (i.e., participation in the military), and $W$ includes socioeconomic variables like race, year of birth, schooling, application year, and the Armed Forces Qualification Test (AFQT) scores.

| Race | Average earnings in in 1988-1991 | Differences in means by veteran status | Matching estimates |
|---|---|---|---|
| Whites | 14537 | 1233.4<br>(60.3) | -197.2<br>(70.5) |
| Non-whites | 11664 | 2449.1<br>(47.4) | 839.7<br>(62.7) |

TABLE 7.1: Uncontrolled and matching estimates of the effects of voluntary military service on earnings. Source: MHE p.73.

Column (3) in Table 7.1 shows the naive difference in means estimate. Column (4) in Table 7.1 shows the matching estimates for the ATT, which only differ from (7.11) by the fact that $\hat{p}_n(w)$ is replaced by an estimator of $P\{W = w \mid A = 1\}$. As we can see, the numbers in columns (3) and (4) are quite different and illustrate the presence of selection bias if one feels comfortable with Assumption 7.1 in this setting. Note that the comparison here is not between individuals that enrolled in the military with those that did not, but rather between veterans and non-veterans among those who applied to get into the all-volunteer forces between 1979 and 1982.

## 7.5 Key Concepts

- **Selection Bias:** in observational studies, the difference in means $E[Y \mid A = 1] - E[Y \mid A = 0]$ equals the ATT plus a selection bias term $E[Y(0) \mid A = 1] - E[Y(0) \mid A = 0]$, which is generally nonzero when agents choose $A$.
- **Unconfoundedness:** $(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid W$. Under this assumption, conditioning on $W$ removes all confounding, so the conditional mean contrast $\Delta(W) = E[Y \mid A = 1, W] - E[Y \mid A = 0, W]$ identifies the CATE $\theta(W)$, and the ATE follows by averaging over $W$.
- **Overlap:** $0 < P\{A = 1 \mid W = w\} < 1$ for all $w$. Without overlap,

some conditional expectations $\mu_a(w)$ cannot be estimated and the ATE is not identified.

- **ATE, ATT, and ATU:** under unconfoundedness, the conditional effects $\theta(W)$, $\theta_{\mathrm{t}}(W)$, and $\theta_{\mathrm{u}}(W)$ coincide, but their unconditional counterparts generally differ because each averages over a different distribution of $W$.
- **Matching Estimator:** compares treated and untreated units within cells sharing the same covariate values and aggregates via $\hat{\theta}_{n,\mathrm{mat}} = \sum_w \hat{p}_n(w)(\bar{Y}_{1,w} - \bar{Y}_{0,w})$.

## 7.6 Concluding Remarks

The material in this chapter borrows from several useful sources, including notes by Alex Torgovitsky, class notes by Stefan Wager Wager [2020], and publicly available notes by Peng Ding Ding [2023]. I want to particularly thank Alex for sharing his source notes with me. In addition to these resources, the paper by Guido Imbens Imbens [2004] provides a good review of many of the concepts covered in this chapter.

## 7.7 Problems

**Problem 7.1** *In e-commerce platforms, customers who purchase a product are often exposed to personalized recommendations based on their browsing or purchase history. Suppose that your goal is to identify the causal effect (say, the ATE) of engaging with recommendations (the treatment) on revenue from the customer on the platform (the outcome).*

1. *Suppose that users who are more likely to spend more money are more likely to click on and purchase recommended products. Explain formally (using mathematical notation) the implications of this behavior.*
2. *Suppose the e-commerce site recommends a high-end tech gadget to a customer based on their previous purchasing behavior. Customers who have previously purchased expensive items or who have more disposable income may be more likely to select and purchase these recommendations, while customers with a lower budget may ignore them. Assuming you observe*

*previous purchasing behavior, what type of assumption would allow you to identify the ATE given this narrative ?*

**Problem 7.2** *Show that Assumption 7.1 also identifies the conditional on $W$ distributions of potential outcomes.*

**Problem 7.3** *Provide a step by step proof that Assumption 7.1 identifies the ATT.*

**Problem 7.4** *Formally describe the matching estimator for the ATT.*

**Problem 7.5** *Suppose the covariate $W$ is binary, $W \in \{0,1\}$. Show that the matching estimator of the ATE defined in equation (7.11) can be simplified to:*

$$\hat{\theta}_{n,\text{mat}} = \frac{n(0)}{n}(\bar{Y}_{1,0} - \bar{Y}_{0,0}) + \frac{n(1)}{n}(\bar{Y}_{1,1} - \bar{Y}_{0,1}) \ ,$$

*and carefully define each term.*

**Problem 7.6** *Assume $W$ is binary and that Assumption 7.1 holds. Derive conditions under which the ATT equals the ATE. Interpret this result intuitively in terms of the relationship between treatment selection and covariates.*

**Problem 7.7** *Let $\pi := P\{A = 1\} \in (0,1)$.*

*(a) Show that $\theta = \pi\,\theta_{\text{t}} + (1-\pi)\,\theta_{\text{u}}$.*

*(b) Use part (a) together with the selection bias decompositions derived in the chapter to show that*

$$\begin{aligned}\theta = {} & E[Y \mid A=1] - E[Y \mid A=0] \\ & - \pi\big(E[Y(0) \mid A=1] - E[Y(0) \mid A=0]\big) \\ & - (1-\pi)\big(E[Y(1) \mid A=1] - E[Y(1) \mid A=0]\big) \ .\end{aligned}$$

*(c) Verify that the expression in part (b) reduces to $\theta = E[Y \mid A=1] - E[Y \mid A=0]$ under random assignment.*

**Problem 7.8** *Suppose that $A \perp\!\!\!\perp W$ (as in a randomized experiment) and that $W$ is discrete with support $\mathcal{W}$. Show that the population matching formula*

$$\theta = \sum_{w \in \mathcal{W}} \big(E[Y \mid A=1, W=w] - E[Y \mid A=0, W=w]\big)\, P\{W=w\}$$

*simplifies to $\theta = E[Y \mid A=1] - E[Y \mid A=0]$, the difference in means. Explain why this result is intuitive.*

## Bibliography

J. D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66 (2):249–288, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2669919.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

S. Wager. Causal inference. Standford University, 2020.

# 8

# *Selection on Observables II*

In this chapter, we continue the analysis of an observational study where we observe a binary treatment $A$, a real-valued outcome of interest $Y$, and some additional pretreatment covariates that we denote by $W$ and that take values in $\mathbf{R}^{d_w}$. Throughout this chapter, we will use the notation,

$$\mu_a(w) := E[Y(a) \mid W = w] \quad \text{and} \quad \sigma_a^2(w) := \text{Var}[Y(a) \mid W = w] \tag{8.1}$$

to denote the conditional expectations and variances of potential outcomes given $W$, and

$$\pi(w) := E[A \mid W = w] = P\{A = 1 \mid W = w\} \tag{8.2}$$

to denote the so-called *propensity score*. The main identifying assumption continues to be selection on observables, as in Assumption 7.1; i.e.,

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid W \ . \tag{8.3}$$

## 8.1 Regression

An alternative method to estimate the ATE, $\theta$, under unconfoundedness and overlap is known as regression or imputation. We start with the characterization of $\theta$ in (7.9), which we can re-write as follows

$$\theta = E[\mu_1(W) - \mu_0(W)] = \int (\mu_1(w) - \mu_0(w)) dF(w) \ , \tag{8.4}$$

where $F$ here denotes the CDF of $W$ - not assumed to be discrete. It follows from there that we can construct a consistent estimator of $\theta$ in two steps. In the first step we estimate the conditional expectations $\mu_1(w)$ and $\mu_0(w)$ nonparametrically. If we denote these estimators by $\hat{\mu}_{n,a}(w)$ for $a \in \mathcal{A}$, the second step simply involves taking a sample average of the difference,

$$\hat{\theta}_{n,\text{reg}} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_{n,1}\left(W_i\right) - \hat{\mu}_{n,0}\left(W_i\right) \right) \ . \tag{8.5}$$

We refer to this estimator as the "regression" estimator, but it is also usually referred to as the "imputation" estimator, since we essentially use the estimated functions $\hat{\mu}_{n,a}(w)$ to impute for the conditional means we do not get to observe. This estimator is consistent and asymptotically normal under certain conditions on the properties of the estimators $\hat{\mu}_{n,a}(w)$ for $a \in \mathcal{A}$. The details of these conditions are beyond the scope of this class.

It is important to note that the terminology "regression" in this context does not necessarily mean "linear" regression. In fact, the estimator in (8.5) is most often based on nonparametric estimators of $\hat{\mu}_{n,a}(w)$, including essentially any of the estimators you learned in Econ 386-1 like kernel regression and local polynomials. More recently, researchers have also relied on machine learning estimators like penalized estimators (e.g., lasso), neural nets, random forests, and so on. The important aspect is that $\hat{\theta}_{n,\text{reg}}$ requires consistent estimators of the true conditional means - as opposed to just linear approximations.

Relying on nonparametric estimators for $\mu_a(w)$ is certainly not the only way to approach this problem. It is not rare for researchers to instead rely on parametric models for $\mu_a(w)$, including a linear model of the form

$$\mu_a(w) = \delta_a + w'\gamma_a \ ,$$

where $\delta_a$ is a scalar and $\gamma_a$ is a $d_w$-dimensional vector of parameters. This is equivalent to assuming a linear model for potential outcomes; i.e.,

$$Y(0) = \delta_0 + W'\gamma_0 + \epsilon_0 \tag{8.6}$$
$$Y(1) = \delta_1 + W'\gamma_1 + \epsilon_1 \tag{8.7}$$

where $E[\epsilon_a \mid W] = 0$ and $E[\epsilon_a] = 0$. We could then proceed by estimating $\theta$ by running two separate least squares regression of $Y$ on $(1, W)$ for each group:

- LS of $Y$ on $(1, W)$ for units with $A_i = 1 \Rightarrow \hat{\mu}_{n,1}(W_i) = \hat{\delta}_{1,n} + W_i'\hat{\gamma}_{1,n}$.
- LS of $Y$ on $(1, W)$ for units with $A_i = 0 \Rightarrow \hat{\mu}_{n,0}(W_i) = \hat{\delta}_{0,n} + W_i'\hat{\gamma}_{0,n}$.

Here, we have denoted the least squares estimators of $(\delta_a, \gamma_a)$ by $(\hat{\delta}_{a,n}, \hat{\gamma}_{a,n})$. It follows that the regression estimator is given by:

$$\begin{aligned}
\hat{\theta}_{n,\text{reg}} &= \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mu}_{n,1}(W_i) - \hat{\mu}_{n,0}(W_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big( (\hat{\delta}_{1,n} + W_i'\hat{\gamma}_{1,n}) - (\hat{\delta}_{0,n} + W_i'\hat{\gamma}_{0,n}) \Big) \\
&= \hat{\delta}_{1,n} - \hat{\delta}_{0,n} + \left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)' (\hat{\gamma}_{1,n} - \hat{\gamma}_{0,n}) \ .
\end{aligned}$$

It follows that the regression estimator $\hat{\theta}_{n,\text{reg}}$ is numerically equivalent to the covariate-adjusted estimator $\hat{\theta}_{n,\text{adj}}$ discussed in Section 6.1.1. However, it is

crucial to understand that, while numerically equivalent, these two estimators are conceptually distinct and are valid under very different assumptions.

The covariate-adjusted estimator $\hat{\theta}_{n,\text{adj}}$ has the following features:

1. It relies on exogenous treatment: $A \perp\!\!\!\perp Y(a)$.
2. It requires only best linear approximations to $\mu_a(w)$.
3. The covariates are introduced for efficiency considerations.

On the other hand, the regression estimator $\hat{\theta}_{n,\text{reg}}$ in this section has the following features:

1. It relies on selection on observables: $A \perp\!\!\!\perp Y(a) \mid W$.
2. It assumes that $\mu_a(w)$ is indeed a linear function of $w$ for each $a$.
3. The covariates are *needed* in order to properly identify the ATE.

To reiterate, the use of linear regression to implement the estimator in (8.5) will only lead to a consistent estimator of $\theta$ if the linear model is correctly specified (i.e., if the potential outcomes are indeed linear, as assumed in the previous model). More generally, the "regression" estimator of the ATE in (8.5) requires a "good" estimator of $\mu_a(w)$ — either because the estimator is a reliable nonparametric estimator (including certain machine learning estimators) or because the estimator is assumed to be valid by assumption.

**Remark 8.1** As discussed in Section 6.1.1, the estimator $\hat{\theta}_{n,\text{reg}}$ can be obtained in a single step from a linear regression with interaction terms:

$$Y = \beta_0 + A\beta_1 + W'\beta_2 + AW'\beta_3 + U \ ,$$

after re-centering the covariates, i.e., $E[W]$. ■

## 8.2 The Role of the Propensity Score

An important result building on the selection on observables assumption shows that one need not condition simultaneously on *all* covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the *propensity score*. That is, the assumption in (8.3) implies the following condition,

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid \pi(W) \ . \tag{8.8}$$

To see this result, note that

$$\begin{aligned}
P\{A=1 \mid Y(0),Y(1),\pi(W)\} &= E[A \mid Y(0),Y(1),\pi(W)] \\
&= E\Big[E[A \mid Y(0),Y(1),\pi(W),W] \mid Y(0),Y(1),\pi(W)\Big] \\
&= E\Big[E[A \mid Y(0),Y(1),W] \mid Y(0),Y(1),\pi(W)\Big] \\
&= E\Big[E[A \mid W] \mid Y(0),Y(1),\pi(W)\Big] \\
&= E\Big[\pi(W) \mid Y(0),Y(1),\pi(W)\Big] \\
&= \pi(W) \ .
\end{aligned}$$

The first equality follows from Hint 5.1. The second equality follows from the law of iterated expectations (LIE), the third equality follows from the properties of conditional expectations (i.e., $E[X|Z,g(Z)] = E[X|Z]$), the fourth equality follows from selection on observables, the fifth equality follows from the definition of the propensity score, and the last equality follows from the properties of conditional expectations (i.e., $E[X|X] = X$).

Similarly, we can show that

$$P\{A=1 \mid \pi(W)\} = \pi(W) \ .$$

This result, due to Rosenbaum and Rubin [1983], shows that the propensity score is a sufficient statistic for the treatment assignment mechanism. In other words, it contains all the relevant information in $W$ for the treatment assignment. Although the original covariates $W$ can be multidimensional, the propensity score $\pi(W)$ is a one-dimensional scalar variable bounded between 0 and 1. Thus, the propensity score reduces the dimensionality of the original covariates while preserving the "ignorability" condition.

The result we just proved has two important implications. First, we can use the propensity score as a dimension-reduction technique by conditioning on $\pi(W)$ instead of $W$, thereby reducing the conditioning set from dimension $d_w$ to just 1. Second, this result leads to alternative characterizations of the ATE and ATT parameters that rely on *weighting*, rather than matching or regression. We discuss these two implications in more detail below.

### 8.2.1 Propensity Score Stratification

The implication of (8.8) is that if we can partition the observations into groups with the same values of the propensity score, $\pi(W)$, then we can identify and consistently estimate parameters such as the ATE and ATT using arguments similar to those in the previous chapter. However, there are two challenges with this intuition. First, the propensity score is generally unknown and must be estimated from the observed data. Second, when $W$ contains continuously distributed covariates, the propensity score may take a continuum of values in $[0,1]$. For simplicity, we begin by assuming that:

1. the propensity score is known, and
2. $\pi(W)$ only takes $K$ possible values $\{\pi_1, \ldots, \pi_K\}$.

The second condition may arise, for example, when $W$ is discrete, so that $\pi_k := P\{A = 1 \mid W = w_k\}$ for each $w_k \in \mathcal{W}$.

The identifying assumption in this case reduces to

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid \pi_k \quad \text{for} \quad k = 1, \ldots, K \ .$$

This situation is often called a "stratified" randomized experiment, since it essentially means that we have $K$ independent experiments within each stratum determined by the propensity score. We can then identify $\theta$ using the following argument,

$$\begin{aligned}
\theta &= E[Y(1) - Y(0)] \\
&= E[E[Y(1) - Y(0) \mid \pi(W)]] \\
&= E[E[Y \mid A = 1, \pi(W)] - E[Y \mid A = 0, \pi(W)]] \\
&= \sum_{k=1}^{K} (E[Y|A = 1, \pi(W) = \pi_k] - E[Y|A = 0, \pi(W) = \pi_k]) P\{\pi(W) = \pi_k\}
\end{aligned}$$

which parallels the one we derived in Section 7.3.1. This result shows that the ATE can be identified by a weighted average of the differences in the conditional expectations of the potential outcomes within strata defined by the propensity score, where the weights are given by $P\{\pi(W) = \pi_k\}$.

We can easily define the natural sample analog associated with propensity score stratification after introducing some additional notation. Let

$$\hat{p}^{\pi}_{n,k} := \frac{1}{n} \sum_{i=1}^{n} I\{\pi(W_i) = \pi_k\}$$

denote the estimator of $P\{\pi(W) = \pi_k\}$ for all $k \leq K$, and let

$$n_{a,k} := \sum_{i=1}^{n} I\{A_i = a, \pi(W_i) = \pi_k\}$$

denote the number of observations with $A = a$ in the $k$th stratum. With this notation, we obtain

$$\hat{\theta}_{n,\text{pss}} := \sum_{k=1}^{K} \hat{p}^{\pi}_{n,k} (\bar{Y}_{1,k} - \bar{Y}_{0,k}) \tag{8.9}$$

where

$$\bar{Y}_{a,k} := \frac{1}{n_{a,k}} \sum_{i=1}^{n} Y_i I\{A_i = a, \pi(W_i) = \pi_k\} \ .$$

In words, $\hat{\theta}_{n,\text{pss}}$ is a weighted average of the within stratum differences in averages between units that are treated and untreated.

In the particular case where $W$ is discrete, both of the difficulties mentioned earlier can be addressed simultaneously. First, we can estimate $\{\pi(w) : w \in \mathcal{W}\}$ nonparametrically by

$$\hat{\pi}_n(w) := \frac{\sum_{i=1}^n I\{A_i = 1, W_i = w\}}{\sum_{i=1}^n I\{W_i = w\}} \quad \forall w \in \mathcal{W} \, . \tag{8.10}$$

Second, because $W$ is discrete, we typically have sufficient observations for each value of $W$ so that both treated and untreated units are present in each stratum. Consequently, $\hat{\theta}_{n,\text{pss}}$ can be defined in the same way as before, with $\hat{\pi}_n(w)$ replacing $\pi(w)$.

In general, the propensity score is not known and $W$ is not discrete. In this case researchers often fit a statistical model for $P\{A = 1 \mid W\}$ (for example, a logistic model or a nonparametric model) to obtain the estimated propensity score $\hat{\pi}_n(W)$. This estimated propensity score can take as many values as the sample size, but we can discretize it to approximate the simple case above. For example, we can discretize the estimated propensity score by its $K$ quantiles, or sort the values of $\hat{\pi}_n(W_i)$ and then split the sample into $K$ evenly sized strata using the sorted propensity score. Once we obtain the $K$ strata, we proceed in the same way we previously did to obtain $\hat{\theta}_{n,\text{pss}}$. This makes the final estimator dependent on the model specification for the propensity score, and so using an incorrect model for $\pi(W)$ would lead to bias in $\hat{\theta}_{n,\text{pss}}$. An important practical question is how to choose $K$. If $K$ is too small, it is expected that exogeneity of $A$ would not appropriately hold. If $K$ is too large, then we may not have enough units within each stratum, with some strata only having treated or control units. Therefore, we face a trade-off in practice. A good data dependent rule to choose $K$ has not been yet established. These issues are beyond the scope of our class.

### 8.2.2 Inverse Probability Weighting

The propensity score leads to an alternative characterization of the ATE and ATT that exhibits certain benefits relative to the one we derived in Section

7.3.1. To see this, consider the following argument,

$$\begin{aligned} E\left[\frac{AY}{\pi(W)}\right] &\overset{(1)}{=} E\left[E\left[\frac{1}{\pi(W)}YA \mid W\right]\right] \\ &\overset{(2)}{=} E\left[\frac{1}{\pi(W)}E\left[YA \mid W\right]\right] \\ &\overset{(3)}{=} E\left[\frac{1}{\pi(W)}E\left[Y(1)A \mid W\right]\right] \\ &\overset{(4)}{=} E\left[\frac{1}{\pi(W)}E[A \mid W]E[Y(1) \mid W]\right] \\ &\overset{(5)}{=} E\left[\frac{1}{\pi(W)}\pi(W)E[Y(1) \mid W]\right] \\ &\overset{(6)}{=} E[Y(1)] \ , \end{aligned}$$

where (1) follows from the LIE; (2) uses that $\pi(W)$ is a function of $W$, so $1/\pi(W)$ can be pulled out of $E[\cdot \mid W]$; (3) follows because $YA = Y(1)A$ by consistency; (4) follows from the selection on observables assumption in (7.4); (5) follows from the definition of the propensity score and Hint 5.1; and (6) follows from the LIE. Similarly, we can show that

$$E\left[\frac{Y(1-A)}{1-\pi(W)}\right] = E[Y(0)] \ .$$

Combining these two results, we obtain the following alternative characterization of $\theta$,

$$\theta = E\left[\frac{AY}{\pi(W)}\right] - E\left[\frac{(1-A)Y}{1-\pi(W)}\right] \ . \tag{8.11}$$

This characterization is known as the inverse probability weighting (IPW) characterization of the ATE and shows that the ATE can be identified by a weighted average of the observed outcomes, where the weights are inversely proportional to the propensity score. Since the propensity score is in the denominator, this alternative characterization explicitly requires the same *overlap* assumption we previously invoked when we proved identification of $F_a(y)$ under selection on observables. We re-state the assumption here,

$$0 < \pi(W) < 1 \quad \text{a.s.} \tag{8.12}$$

Looking at the representation in (8.11), we can see immediately that in order to exploit this representation we would need to either know the propensity score $\pi(w)$ or be able to estimate it consistently. Denote by $\hat{\pi}_n(w)$ a consistent estimator of $\pi(w)$ and define,

$$\hat{\theta}_{n,\text{ipw}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i Y_i}{\hat{\pi}_n(W_i)} - \frac{(1-A_i)\,Y_i}{1-\hat{\pi}_n(W_i)}\right) \ . \tag{8.13}$$

We note again that in experiments where treatment is randomly assigned conditional on covariates (e.g., strata), the propensity score may be reasonably assumed to be known. In other settings that would not be the case, and so the consistency of $\hat{\theta}_{n,\text{ipw}}$ would fundamentally depend on whether $\pi(w)$ can be estimated consistently or not.

## 8.3 Empirical Illustration

Dehejia and Wahba [1999] used a subsample of the CPS combined with an experimental sample to evaluate the performance of propensity score–based methods. In their design, treated observations were drawn from the experiment, while control observations were obtained from the CPS — thereby converting an experimental setting into an observational one. The experiment involved a labor training program, and the goal was to assess the effect of the program on participants' 1978 income (the final year of the program). Notably, when the average treatment effect (ATE) was estimated using a naive difference-in-means estimator, the result was $-\$8,948$. In contrast, methods exploiting the propensity score indicated that the treatment increased yearly income by $\$1,794$ for the participants.

Our goal is to use the data set of this application to implement the methods discussed in this chapter. The variables in the data set are:

| | |
|---|---|
| **re78** | Income in year 1978 (outcome) |
| **treat** | (1: Treated, 0: Control) (treatment) |
| **age** | age |
| **black** | race (1: Black, 0: otherwise) |
| **hispanic** | race (1: Hispanic, 0 otherwise) |
| **nodegree** | (1: Dropped out High School, 0: otherwise) |
| **education** | Years of Education |
| **re74** | Income in year 1974 |
| **re75** | Income in year 1975 |

The authors use a logistic regression of the treatment status variable on the covariates as a model for the propensity score. Since we do not cover this method in class, the data set `dw_pscore.csv` has this propensity score already estimated in the column `pscore` - if you are curious, the R code `7-1-pscore-logit.R` contains the code to generate this data set, which estimates the logistic regression and then keeps only the observations with $\pi_n(W_i) \in [0.1, 0.9]$. The resulting dataset has 490 observations in total, with 140 being treated.

To compute $\hat{\theta}_{n,\text{pss}}$, we first need to choose the number of strata $K$ and then stratify the propensity score. As mentioned earlier, determining the optimal value of $K$ is beyond the scope of this class, so we simply consider two options:

$K \in \{5, 10\}$. Given $K$, we stratify the propensity score using the following function:

```
# Load data
dw <- read.csv("dw_pscore.csv", header = TRUE)

# Function to add strata based on the propensity score
stratify_pscore <- function(data, n_strata, pscore = pscore) {
  data %>% mutate(strata = cut({{ pscore }},
                       breaks = seq(min({{ pscore }}), max({{ pscore
    }}),
                       length.out = n_strata + 1),
                       include.lowest = TRUE))
}
```

Code Snippet 8.1: Stratify propensity score

Figure 8.1 displays histograms of treated and control units for each stratum. The coloring scheme clearly shows that overlap holds, as each stratum contains both treated and untreated observations. Not surprisingly, strata with low values of the propensity score have many untreated units and few treated units, whereas those with high propensity scores have many treated units relative to controls.

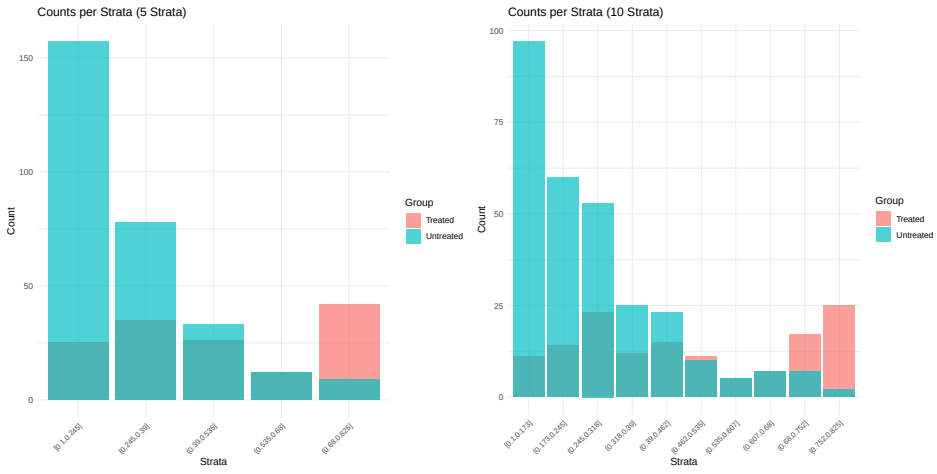



FIGURE 8.1: Number of treated and untreated units within strata.

We can now estimate $\theta$ using $\hat{\theta}_{n,pss}$ in (8.9). The code below accomplishes this in R and Table 8.1 presents the results for both values of $K$.

```
# Function to compute the ATE from stratified data
compute_ate <- function(data, treat, outcome) {
  data %>%
    group_by(strata) %>%
    summarise(diff = mean({{ outcome }}[{{ treat }} == 1])
                     - mean({{ outcome }}[{{ treat }} == 0]),
```

```
                weight = n() / nrow(data)) %>%
    ungroup() %>%
    summarise(ATE = sum(diff * weight)) %>%
    pull(ATE)
}

# Estimate ATE with K=5 strata
dw <- stratify_pscore(dw, n_strata = 5)
ate_5 <- compute_ate(dw, treat, re78)
```

Code Snippet 8.2: Propensity score stratification

| K | 5 | 10 |
|---|---|---|
| $\hat{\theta}_{n,\text{pss}}$ | \$1,421 | \$1,346 |

TABLE 8.1: $\hat{\theta}_{n,\text{pss}}$ for different values of $K$.

It is interesting to decompose the numbers within each stratum to get a sense of the magnitude of the estimated treatment effect within each stratum, as well as the numbers of treated and control units within each stratum. The following output illustrates this for the case $K = 5$

```
  strata        ate_s weight n_treat n_untreat
1 [0.1,0.245]   -475  0.424       25       157
2 (0.245,0.39] 1402  0.263       35        78
3 (0.39,0.535] 4504  0.138       26        33
4 (0.535,0.68] 3210  0.0559      12        12
5 (0.68,0.825] 3815  0.119       42         9
```

Code Snippet 8.3: Summary within strata

We conclude that our estimates are closer to the experimental estimate of \$1,794 in the original paper, than they are to the naive difference-in-means estimate of $-\$8,948$. The code `7-2-logit-pss.R` reproduces all of the results in this section.

## 8.4 Scope of Selection on Observables

Selection on observables continues to be a popular assumption in many social sciences and, more recently, in industry applications. But it is nevertheless an assumption that is difficult to digest in most economic applications, as inherent unobservables (preferences, private info, expectations) tend to play a role in how people make decisions and these, in turn, tend to depend on potential outcomes. That is, the idea that observationally identical people behave differently due to a coin flip is difficult to defend. In the era of big data, researchers often tend to feel more re-assured about selection on observables

by "controlling for more". That is, with large sets of covariates we can include a lot of information into $W$. However, it can be shown that controlling for more covariates may increase the bias of estimators relative to using a subset of them. There is also an existing tension with overlap. That is, if we could perfectly explain $A$ with $W$ then $P\{A = 1|W\}$ would be either 0 or 1 and we wouldn't have the required variation. Better methods for choosing observables will not solve these problems, so the ability of machine learning techniques to provide identification guarantees is limited. However, taking the identifying assumption as given and focusing on estimation accuracy then it is indeed the case the more modern ML techniques could lead to better estimators than the ones we discussed in this chapter. We will cover one of these improvements in the next chapter.

## 8.5 Key Concepts

- **Regression Estimator:** uses consistent estimates of $\mu_a(w)$ to impute missing potential outcomes. Numerically equivalent to the covariate-adjusted estimator, but valid under different assumptions: it requires selection on observables and a correctly specified model for $\mu_a(w)$.
- **Propensity Score:** $\pi(w) = P\{A = 1 \mid W = w\}$. Under selection on observables, $(Y(a)) \perp\!\!\!\perp A \mid \pi(W)$, reducing the conditioning set from $d_w$ dimensions to one.
- **Propensity Score Stratification:** partitions the sample into strata based on the propensity score and compares treated and control units within each stratum. Relies on consistent estimation of $\pi(w)$.
- **Inverse Probability Weighting (IPW):** re-weights observations by the inverse of $\pi(W)$ or $1 - \pi(W)$ to identify the ATE. Requires both consistent estimation of $\pi(w)$ and overlap.
- **Scope of Selection on Observables:** controlling for more covariates does not necessarily reduce bias and can conflict with the overlap assumption.

## 8.6 Concluding Remarks

The material in this chapter borrows from several useful sources, including notes by Alex Torgovitsky, class notes by Wager [2020], publicly available notes by Ding [2023], and the book by Angrist and Pischke [2008]. In addition to these resources, the paper by Imbens [2004] provides a good review of many of the concepts covered in this chapter.

## 8.7 Problems

**Problem 8.1** *The estimator in (8.5) imputes each treated group with its respective conditional expectation, $\hat{\mu}_{n,a}(w)$. However, it is easy to see that the amount of imputation can be limited by using the observed outcome whenever available (i.e., using $Y_i$ when $A_i = 1$ instead of $\hat{\mu}_{n,1}(w)$). This leads to the following variation of the "regression" estimator:*

$$\tilde{\theta}_{n,\text{reg}} := \frac{1}{n}\sum_{i=1}^{n} A_i \left(Y_i - \hat{\mu}_{n,0}(W_i)\right) + \frac{1}{n}\sum_{i=1}^{n}(1 - A_i)\left(\hat{\mu}_{n,1}(W_i) - Y_i\right) \ . \quad (8.14)$$

*Assume that $\hat{\mu}_{n,a}(w) \xrightarrow{p} \mu_a(w)$ for all $w \in \mathcal{W}$ and $a \in \mathcal{A}$.*

1. *Denote by $\tilde{\theta}^*_{n,\text{reg}}$ the version of the estimator that uses the true conditional expectations, $\mu_a(w)$, instead of the estimated ones. Show that $\tilde{\theta}^*_{n,\text{reg}}$ is unbiased for $\theta$.*
2. *Show that $\tilde{\theta}^*_{n,\text{reg}}$ is consistent for $\theta$.*
3. *Between $\tilde{\theta}_{n,\text{reg}}$ and $\hat{\theta}_{n,\text{reg}}$, which estimator do you expect to have better properties? Use your intuition rather than providing a formal answer.*

**Problem 8.2** *Show that $E\left[\frac{Y(1-A)}{1-\pi(W)}\right] = E[Y(0)]$.*

**Problem 8.3** *Using the same data set we used in Section 8.3, compute the inverse probability weighted estimator, $\hat{\theta}_{n,\text{ipw}}$, and show it leads to an estimated value of \$916. The IPW estimator is also known as the Horvitz-Thompson estimator.*

**Problem 8.4** *One somewhat unappealing feature of the IPW estimator is that the weights do not necessarily add to 1. This implies that this estimator is not invariant to location shifts on the outcomes (i.e., adding a constant to*

*$Y$). A variation that ensure that the weights add up to 1 consists in simply re-normalizing the weights as follows :*

$$\tilde{\theta}_{n,\text{ipw}} := \frac{\sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}_n(W_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-A_i)Y_i}{1-\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n} \frac{1-A_i}{1-\hat{\pi}_n(W_i)}} \ .$$

*Using the same data set we used in Section 8.3, compute the estimator above and show it leads to an estimated value of* $\$1,326$. *This estimator is also known as the Hájek estimator and tends to perform better in finite samples.*

**Problem 8.5** *Using the IPW characterization in (8.11), show that if treatment is randomly assigned (i.e., $A \perp\!\!\!\perp (Y(0), Y(1))$) and $\pi := P\{A = 1\}$, then the IPW estimator $\hat{\theta}_{n,\text{ipw}}$ reduces to the simple difference in means $\bar{Y}_1 - \bar{Y}_0$.*

**Problem 8.6** *Show that the ATT can be written as*

$$\theta_{\text{t}} = E\left[Y \cdot \frac{A - \pi(W)}{P\{A = 1\}(1 - \pi(W))}\right] \ .$$

Hint: *start from the definition $\theta_{\text{t}} = E[Y(1) - Y(0) \mid A = 1]$ and use the LIE and selection on observables to express each term using observed quantities.*

## Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2669919.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

S. Wager. Causal inference. Standford University, 2020.

# Part II

# Panel Data and RDD

# 9

# *Panel Data*

In many empirical settings, we observe the same units—individuals, firms, states, cities—repeatedly over multiple time periods. Data of this form is called *panel data* (or longitudinal data), and it has a distinctive structure: each observation is indexed by a unit $i$ and a time period $t$, giving rise to data of the form

$$\{(Y_{i,t}, X_{i,t}) : i = 1, \ldots, n,\ t = 1, \ldots, T\} \ .$$

We motivate the ideas of this chapter with the following running example.

**Example 9.1 (Traffic fatalities and alcohol taxes)** There are approximately 40,000 highway traffic fatalities each year in the United States, and roughly one-fourth of fatal crashes involve a driver who was drinking. Levitt and Porter [2001] estimate that as many as 25% of drivers on the road between 1 a.m. and 3 a.m. have been drinking, and that a legally drunk driver is at least 13 times more likely to cause a fatal crash than a sober one. A natural policy question is: how effective are government policies designed to discourage drunk driving at actually reducing traffic deaths?

This application, drawn from Chapter 10 of Stock and Watson [2003], will serve as our running example throughout this chapter. The data set is a panel of 48 U.S. states observed annually from 1982 to 1988 ($n = 48$ states, $T = 7$ years). The outcome variable $Y$ is the *fatality rate*—the number of annual traffic deaths per 10,000 people in the state—and the main covariate $X$ is the real tax on a case of beer (in 1988 dollars), which serves as a proxy for alcohol policy. ■

Why is this structure useful? Consider a model where the outcome $Y_{i,t}$ depends on observed covariates $X_{i,t}$ and also on a variable $W_i$ that varies across units but does *not* change over time. In the traffic fatalities example, $W_i$ could represent cultural attitudes toward drinking and driving in state $i$—something that plausibly differs across states but changes slowly enough to be approximately constant over the sample period. The model is then

$$Y_{i,t} = \alpha_0 + X_{i,t}'\beta + \gamma W_i + U_{i,t} \ , \tag{9.1}$$

where $U_{i,t}$ represents idiosyncratic shocks that vary across both units and time. The key difficulty is that $W_i$ is typically *unobserved* and *correlated* with $X_{i,t}$. For example, cultural attitudes toward alcohol may influence both the

beer tax a state adopts and its fatality rate. If we ignore $W_i$ and run a cross-sectional regression of $Y$ on $X$, the omitted variable $W_i$ gets absorbed into the error term, and the resulting OLS estimator is inconsistent for $\beta$—this is the familiar omitted variable bias we discussed earlier in the course.

**Example 9.2 (Traffic fatalities, cont.)** As a first pass, consider ignoring the panel structure and running a cross-sectional regression of the fatality rate on the beer tax using data from a single year. The code below does this for 1982 and 1988 separately.

```
library(haven)
library(dplyr)

# Read the Stata file (Stock & Watson fatality data)
data <- read_dta("fatality.dta")

# Create fatality rate: annual traffic deaths per 10,000 people
data$fatality.rate <- 10000 * data$mrall

# --- Cross-sectional OLS using only 1982 data ---
data_1982 <- data %>% filter(year == 1982)
lm(formula = fatality.rate ~ 1 + beertax, data = data_1982)

# --- Cross-sectional OLS using only 1988 data ---
data_1988 <- data %>% filter(year == 1988)
lm(formula = fatality.rate ~ 1 + beertax, data = data_1988)
```

Code Snippet 9.1: Cross-sectional regressions for 1982 and 1988

The estimated slope coefficients are *positive* in both years: $\hat{\beta}_1 \approx 0.15$ in 1982 and $\hat{\beta}_1 \approx 0.44$ in 1988. That is, higher beer taxes are associated with *more* traffic fatalities, not fewer—contradicting the prediction from economic theory. This is precisely the type of omitted variable bias described above: unobserved state-level factors $W_i$, such as cultural attitudes toward drinking, are correlated with both the beer tax and the fatality rate, biasing the cross-sectional estimate. ■

Panel data offers an elegant solution to this specific type of endogeneity. Since $W_i$ does not change over time, its effect $\gamma W_i$ is the same in every period for a given unit. If we can difference it away—by comparing the *same unit* across time—we eliminate the bias without needing to observe $W_i$ directly. To formalize this idea, we define

$$\eta_i := \alpha_0 + \gamma W_i \ ,$$

which captures the intercept together with all unobserved, unit-specific, time-invariant factors. Note that $X_{i,t}$ does *not* include a constant, as it is absorbed into $\eta_i$. The model becomes

$$Y_{i,t} = X_{i,t}'\beta + \eta_i + U_{i,t} \ . \tag{9.2}$$

The unobserved random variables $\eta_i$ are commonly called *fixed effects*. Even

though $\eta_i$ may be correlated with $X_{i,t}$ (so that $E[X_{i,t}\eta_i] \neq 0$), we will see that the panel structure allows us to consistently estimate $\beta$ under certain restrictions on $U_{i,t}$.

To see the main idea, consider the simplest case: two time periods ($T = 2$). The model gives us

$$\begin{aligned} Y_{i,1} &= X_{i,1}'\beta + \eta_i + U_{i,1} \\ Y_{i,2} &= X_{i,2}'\beta + \eta_i + U_{i,2} \ . \end{aligned}$$

Note that $\beta$ is a constant parameter that does not change over time, and the fixed effect $\eta_i$ appears identically in both equations. Taking first differences eliminates $\eta_i$:

$$\begin{aligned} Y_{i,2} - Y_{i,1} &= (X_{i,2} - X_{i,1})'\beta + U_{i,2} - U_{i,1} \\ \Delta Y_i &= \Delta X_i'\beta + \Delta U_i \ , \end{aligned}$$

and we have removed the unobserved individual effect in the process. For least squares on the differenced equation to deliver a consistent estimator of $\beta$, we need $E[\Delta X_i \Delta U_i] = 0$. Expanding this condition,

$$E[\Delta X_i \Delta U_i] = E[X_{i,2}U_{i,2}] + E[X_{i,1}U_{i,1}] - E[X_{i,2}U_{i,1}] - E[X_{i,1}U_{i,2}] \ . \quad (9.3)$$

For this expression to be zero, it is not enough to assume $E[X_{i,t}U_{i,t}] = 0$ for each $t$ (the standard orthogonality assumption). We also need $E[X_{i,2}U_{i,1}] = E[X_{i,1}U_{i,2}] = 0$, meaning that the covariates in one time period are uncorrelated with the idiosyncratic shocks in other time periods. These conditions taken together are called *strict exogeneity*. Under strict exogeneity and the assumption that $E[\Delta X_i \Delta X_i']$ is invertible, running least squares of $\Delta Y_i$ on $\Delta X_i$ consistently estimates

$$\beta = E[\Delta X_i \Delta X_i']^{-1} E[\Delta X_i \Delta Y_i] \ .$$

Before we proceed to formalize and extend these ideas, there are a few aspects worth keeping in mind. First, observing the same units over multiple time periods allows us to control for unobserved factors that are *constant over time*—the $\eta_i$. The trick we just used would not work if $\eta_i$ were allowed to change over time. Second, the requirement that $E[\Delta X_i \Delta X_i']$ is invertible means that we need $X$ to *change over time*; hence, the differencing approach does not allow us to estimate coefficients of variables that are constant over time, since such variables are removed by the transformation in the same way $\eta_i$ is removed. Finally, strict exogeneity is a stronger assumption than simply requiring $E[X_{i,t}U_{i,t}] = 0$ for each $t$. Cases where $X_{i,2}$ is a decision variable chosen by an agent who has already observed $U_{i,1}$ may seriously violate $E[X_{i,2}U_{i,1}] = 0$. This type of dynamic concern is distinct from omitted variable bias and could arise even if $E[X_{i,t}U_{i,t}] = 0$ holds for every $t$.

## 9.1 Fixed Effects

### 9.1.1 First Differences

Let $(Y, X, \eta, U)$ be distributed as described above and denote by $P$ the distribution of

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}) \ . \tag{9.4}$$

We assume that we have a random sample of size $n$, so that the observed data is given by $\{(Y_{i,t}, X_{i,t}) : 1 \leq i \leq n, 1 \leq t \leq T\}$. Note that while the sampling process is i.i.d. across $i$, we are being completely agnostic about the dependence across time for a given unit $i$. We then consider

$$Y_{i,t} = X_{i,t}'\beta + \eta_i + U_{i,t}, \ i = 1, \ldots, n \quad t = 1, \ldots, T \ , \tag{9.5}$$

under the assumptions on $X_{i,t}$ and $U_{i,t}$ that we formalize below. Now define

$$\Delta X_{i,t} = X_{i,t} - X_{i,t-1}$$

for $t \geq 2$, and proceed analogously with the other random variables. Note again that $\Delta\eta_i = 0$. Applying this transformation to (9.5), we get

$$\Delta Y_{i,t} = \Delta X_{i,t}'\beta + \Delta U_{i,t}, \ i = 1, \ldots, n \quad t = 2, \ldots, T \ . \tag{9.6}$$

It follows that a regression of $\Delta Y_{i,t}$ on $\Delta X_{i,t}$ provides a consistent estimator of $\beta$ if the following two assumptions hold,

**FD1**. $E[U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$,

**FD2**. $\sum_{t=2}^{T} E[\Delta X_{i,t}\Delta X_{i,t}'] < \infty$ is invertible.

FD1 is a sufficient condition for $E[\Delta U_{i,t}|\Delta X_{i,t}] = 0$. FD2 fails if some component of $X_{i,t}$ does not vary over time. The first-difference estimator then takes the form

$$\hat{\beta}_n^{\text{fd}} = \left( \sum_{1 \leq i \leq n} \sum_{2 \leq t \leq T} \Delta X_{i,t}\Delta X_{i,t}' \right)^{-1} \left( \sum_{1 \leq i \leq n} \sum_{2 \leq t \leq T} \Delta X_{i,t}\Delta Y_{i,t} \right) \ . \tag{9.7}$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1}, \ldots, X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fd}}$ is not asymptotically efficient and that a different transformation of the data delivers an estimator with a lower asymptotic variance under those assumption. We will discuss this further after describing this alternative transformation.

**Example 9.3 (Traffic fatalities, cont.)** Returning to the traffic fatalities application (Example 9.1), we now apply the first-difference approach using only 1982 and 1988 ($T = 2$). For each state we compute the change in the fatality rate and the change in the beer tax, then regress one on the other.

```r
library(haven)
library(dplyr)

data <- read_dta("fatality.dta")
data$fatality.rate <- 10000 * data$mrall

# Build first-difference data (1988 minus 1982)
fd <- data %>%
  filter(year %in% c(1982, 1988)) %>%
  arrange(state, year) %>%
  group_by(state) %>%
  summarise(
    d_fatality = diff(fatality.rate),
    d_beertax  = diff(beertax),
    .groups = "drop"
  )

# OLS on differenced data (no intercept needed in theory,
# but lm includes one by default)
model_fd <- lm(d_fatality ~ d_beertax, data = fd)
summary(model_fd)
```

Code Snippet 9.2: First-difference regression (1982 vs. 1988)

The estimated coefficient on the change in the beer tax is $\hat{\beta}_1^{\text{fd}} \approx -1.04$, which is now *negative*, consistent with the prediction from economic theory that higher alcohol taxes reduce traffic fatalities. The sign reversal relative to the cross-sectional results in Example 9.2 illustrates the power of the differencing approach: cultural attitudes toward drinking and driving affect the *level* of fatalities in a state, but if those attitudes did not change appreciably between 1982 and 1988, they are eliminated by differencing. Any remaining changes in fatalities must be driven by other sources—such as changes in the beer tax. ■

### 9.1.2 Deviations from Means

An alternative transformation to remove the individual effects $\eta_i$ from (9.5) is the so-called de-meaning technique. In order to define this formally, let

$$\dot{X}_{i,t} = X_{i,t} - \bar{X}_i \quad \text{where} \quad \bar{X}_i = \frac{1}{T} \sum_{1 \le t \le T} X_{i,t} \ ,$$

and define $\dot{Y}_{i,t}$ and $\dot{U}_{i,t}$ analogously. Note that $\dot{\eta}_i = 0$ for all $i = 1, \ldots, n$. Applying this transformation to (9.5), we get

$$\dot{Y}_{i,t} = \dot{X}_{i,t}'\beta + \dot{U}_{i,t}, \quad i = 1, \ldots, n \quad t = 1, \ldots, T \ . \tag{9.8}$$

It follows that a regression of $\dot{Y}_{i,t}$ on $\dot{X}_{i,t}$ provides a consistent estimator of $\beta$ if the following two assumptions hold,

**FE1**. $E[U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$,

**FE2**. $\sum_{t=1}^{T} E[\dot{X}_{i,t}\dot{X}'_{i,t}] < \infty$ is invertible.

FE1, which is the same strict exogeneity condition in FD1, is a sufficient condition for $E[\dot{U}_{i,t}|\dot{X}_{i,t}] = 0$. As before, FE2 fails if some component of $X_{i,t}$ does not vary over time. The de-meaning estimator (commonly known as the fixed effect estimator) or dummy variable estimator takes the form

$$\hat{\beta}_n^{\text{fe}} = \left( \sum_{1\le i\le n} \sum_{1\le t\le T} \dot{X}_{i,t}\dot{X}'_{i,t} \right)^{-1} \left( \sum_{1\le i\le n} \sum_{1\le t\le T} \dot{X}_{i,t}\dot{Y}_{i,t} \right) . \tag{9.9}$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1}, \ldots, X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fe}}$ is asymptotically efficient. We discuss this in the next section.

**Example 9.4 (Traffic fatalities, cont.)** We now apply the fixed effects estimator to the full panel of 48 states over all 7 years (1982–1988). In R, the `feols` function from the `fixest` package provides a convenient way to estimate fixed effects models. The syntax `fatality.rate ~ beertax | state` tells R to regress the fatality rate on the beer tax, including state fixed effects. The option `cluster = ~state` produces cluster-robust standard errors.

```r
library(haven)
library(dplyr)
library(fixest)

data <- read_dta("fatality.dta")
data$fatality.rate <- 10000 * data$mrall

# --- FE regression using all 7 years (1982--1988) ---
# State fixed effects with cluster-robust standard errors
fe_model <- feols(fatality.rate ~ beertax | state,
                  data = data, cluster = ~state)
summary(fe_model)

# --- Manual within transformation (for illustration) ---
demeaned <- data %>%
  group_by(state) %>%
  mutate(
    y_dot = fatality.rate - mean(fatality.rate),
    x_dot = beertax       - mean(beertax)
  ) %>%
  ungroup()

fe_manual <- lm(y_dot ~ x_dot - 1, data = demeaned)
summary(fe_manual)
```

Code Snippet 9.3: Fixed effects regression with all 7 years

The estimated effect of the beer tax is $\hat{\beta}_1^{\text{fe}} \approx -0.66$ with a cluster-robust standard error of approximately 0.29, yielding a t-statistic of about $-2.25$

(significant at the 5% level). The sign is negative, as in the first-difference approach. The magnitude differs because we are now using all 7 time periods rather than just two. Recall that when $T = 2$, the fixed effects and first-difference estimators are numerically identical (see Problem 9.1). The code also illustrates the manual within transformation: demeaning $Y$ and $X$ within each state and running OLS on the demeaned data produces the same point estimate as `feols`, confirming that the fixed effects estimator is simply OLS on de-meaned data. ■

### 9.1.3 Two-Way Fixed Effects

The model in (9.5) includes unit fixed effects $\eta_i$ to absorb unobserved factors that vary across units but are constant over time. In many applications, however, there are also common shocks that affect all units in a given time period. For instance, over the 1982–1988 period, cars were becoming safer and seat belt usage was rising, changes that affected all states simultaneously. If such time-varying common factors are correlated with the beer tax, the fixed effects estimator will still suffer from omitted variable bias.

To address this, we can extend the model by including *time fixed effects* $\gamma_t$:

$$Y_{i,t} = X_{i,t}'\beta + \eta_i + \gamma_t + U_{i,t}, \quad i = 1, \ldots, n \quad t = 1, \ldots, T \ . \tag{9.10}$$

The unit effects $\eta_i$ capture unobserved heterogeneity across units that is constant over time, while the time effects $\gamma_t$ capture unobserved shocks common to all units in period $t$. Together, they give the model its name: the *two-way fixed effects* (TWFE) model.

To eliminate both sets of fixed effects, we apply a two-way de-meaning transformation. For a generic variable $W_{i,t}$, define the residual from projecting onto unit and time effects as

$$\tilde{W}_{i,t} = W_{i,t} - \bar{W}_i - \bar{W}_t + \bar{W} \ , \tag{9.11}$$

where

$$\bar{W}_i := \frac{1}{T}\sum_{t=1}^{T} W_{i,t} \ , \quad \bar{W}_t := \frac{1}{n}\sum_{i=1}^{n} W_{i,t} \ , \quad \bar{W} := \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} W_{i,t} \ .$$

Subtracting $\bar{W}_i$ removes the unit mean (eliminating $\eta_i$), subtracting $\bar{W}_t$ removes the time mean (eliminating $\gamma_t$), and adding back $\bar{W}$ corrects for double-counting the grand mean. Applying this transformation to (9.10) yields

$$\tilde{Y}_{i,t} = \tilde{X}_{i,t}'\beta + \tilde{U}_{i,t} \ ,$$

and the two-way fixed effects estimator is

$$\tilde{\beta}_n^{\text{fe}} = \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{i,t}\tilde{X}_{i,t}'\right)^{-1} \left(\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{i,t}\tilde{Y}_{i,t}\right) \ . \tag{9.12}$$

In R, including time fixed effects with `feols` is straightforward: the specification `fatality.rate ~ beertax  state + year|` adds year dummies alongside state dummies.

### 9.1.4 Asymptotic Properties

Deriving an asymptotic approximation for estimators in panel data models involves two elements that were not present with cross-sectional data. First, the data is i.i.d. across $i$ but may be dependent across time. This is, we may suspect that $X_{i,t}$ and $X_{i,s}$ for $t \neq s$ may not be independent. Second, the data has two indices now: the number of units (denoted by $n$) and the number of time periods (denoted by $T$). We will definitely need $nT \to \infty$ to get a useful asymptotic approximation, but we may achieve this by all sort of different assumptions about how $n$ and/or $T$ grow. The two standard approximations are $n \to \infty$ and $T$ fixed (the so-called short panels) and $n \to \infty$ and $T \to \infty$ (the so-called large panels). Many commonly used panels in applied research include thousands of units ($n$ large) and few time periods ($T$ small) so we will focus on short panels first and discuss large panels later in class.

Under asymptotics where $n \to \infty$ and fixed $T$, we can show that $\hat{\beta}_n^{\text{fe}}$ and $\hat{\beta}_n^{\text{fd}}$ are asymptotically normal using similar arguments to those we use before, provided we assume

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}, U_{i,1}, \ldots, U_{i,T})$$

are i.i.d. across $i = 1, \ldots, n$. Start by writing

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}} - \beta) = \left( \frac{1}{n} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} \right) .$$

In order to make this expression more tractable, we use two tricks. First, note that

$$\sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} - \bar{U}_i \sum_{1 \leq t \leq T} \dot{X}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} , \qquad (9.13)$$

where the last step follows from $\sum_{1 \leq t \leq T} \dot{X}_{i,t} = 0$. We can therefore replace $\dot{U}_{i,t}$ with $U_{i,t}$. Second, let $\dot{X}_i = (\dot{X}_{i,1}, \ldots, \dot{X}_{i,T})'$ be a $T \times k$ vector of stacked observations for unit $i$, and define $U_i$ in the same way. Using this notation, we can write

$$\dot{X}_i' \dot{X}_i = \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \quad \text{and} \quad \dot{X}_i' U_i = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_i . \qquad (9.14)$$

Combining (9.13) and (9.14), we obtain

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}} - \beta) = \left( \frac{1}{n} \sum_{1 \leq i \leq n} \dot{X}_i' \dot{X}_i \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \dot{X}_i' U_i \right) .$$

By the law of large numbers and FE2,

$$\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i \xrightarrow{P} \Sigma_{\dot{X}} \equiv E[\dot{X}_i'\dot{X}_i] = \sum_{1\leq t\leq T} E[\dot{X}_{i,t}\dot{X}_{i,t}'] \ .$$

In addition, by the central limit theorem and FE1,

$$\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\dot{X}_i'U_i \xrightarrow{d} N(0,\Omega), \quad \text{where} \quad \Omega = \mathrm{Var}[\dot{X}_i'U_i] = E[\dot{X}_i'U_iU_i'\dot{X}_i] \ .$$

Combining these results with the CMT we get

$$\sqrt{n}(\hat{\beta}_n^{\mathrm{fe}} - \beta) \xrightarrow{d} N(0, \mathbb{V}^{\mathrm{fe}}) \tag{9.15}$$

where

$$\mathbb{V}^{\mathrm{fe}} = \Sigma_{\dot{X}}^{-1}\Omega\Sigma_{\dot{X}}^{-1} \ . \tag{9.16}$$

Historically, researchers often assumed that $U_{i,t}$ was serially uncorrelated with variance independent of $X_{i,t}$ (i.e. homoskedastic). The default standard errors in Stata are still based on these assumptions. However, these assumptions are difficult to justify for most economic data, which is often strongly autocorrelated and heteroskedastic. One faces basically the same trade-off as with heteroskedasticity in the cross-sectional case. The most common strategy is to use the fully robust consistent estimator of the asymptotic variance,

$$\hat{\mathbb{V}}^{\mathrm{fe}} = \left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\hat{U}_i\hat{U}_i'\dot{X}_i\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1} \ ,$$

where $\hat{U}_i = \dot{Y}_i - \dot{X}_i\hat{\beta}_n^{\mathrm{fe}}$. This is what Stata computes when one uses the `cluster(unit)` option to `xtreg` where unit is the variable that indexes $i$. This estimator is an appealing generalization of White's (1980) heteroskedasticity consistent covariance matrix estimator that allows for arbitrary intertemporal correlation patterns and heteroskedasticity across individuals. As we will see later in class, this estimator is generally known as a cluster covariance estimator (CCE) and is consistent as $n \to \infty$, i.e., $\hat{\mathbb{V}}^{\mathrm{fe}} \xrightarrow{P} \mathbb{V}^{\mathrm{fe}}$.

A cluster-robust standard error for the $j$th component of $\hat{\beta}_n^{\mathrm{fe}}$ is then obtained as

$$\widehat{\mathrm{se}}(\hat{\beta}_{n,j}^{\mathrm{fe}}) = \sqrt{\frac{\hat{\mathbb{V}}_{n,[j,j]}^{\mathrm{fe}}}{n}} \ .$$

Note that because the model does not include a constant (it is absorbed by the fixed effects), the $j$th diagonal element of $\hat{\mathbb{V}}_n^{\mathrm{fe}}$ directly corresponds to the asymptotic variance of $\hat{\beta}_{n,j}^{\mathrm{fe}}$.

**Remark 9.1** Panel data traditionally deals with units over time. However, we

can think about other cases where the data has a two-dimensional index and where we believe that one of the indices may exhibit within group dependence. For example, it could be that we observe "employees" within "firms", or "students" within "schools", or "families" in metropolitan statistical areas (MSA), etc. Cases like these are similar but not identical to panel data. To start, units are not "repeated" in the sense that each unit is potentially observed only once in the sample. In addition, these are cases where "$T$" is usually large and "$n$" is small. For example, we typically observe many students (which may be dependent within a school) and few schools. We will study these cases later in the class. ■

## 9.2 Key Concepts

- **Panel Data:** Data where the same units (individuals, states, firms, etc.) are observed over multiple time periods. The model $Y_{i,t} = X_{i,t}'\beta + \eta_i + U_{i,t}$ includes unit-specific fixed effects $\eta_i$ that capture unobserved heterogeneity constant over time.
- **First-Difference Estimator:** Eliminates fixed effects by taking differences across consecutive time periods: $\Delta Y_{i,t} = \Delta X_{i,t}'\beta + \Delta U_{i,t}$. Loses one time period but is particularly useful when $T = 2$.
- **Fixed Effects (Within) Estimator:** Eliminates fixed effects by subtracting unit-specific means: $\dot{Y}_{i,t} = \dot{X}_{i,t}'\beta + \dot{U}_{i,t}$. When $T = 2$, it is numerically identical to the first-difference estimator.
- **Strict Exogeneity:** The assumption $E[U_{i,t} \mid X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t$, which is stronger than contemporaneous exogeneity. It rules out feedback from past shocks to future covariates.
- **Two-Way Fixed Effects:** Extends the model to include both unit and time fixed effects, $Y_{i,t} = X_{i,t}'\beta + \eta_i + \gamma_t + U_{i,t}$, controlling for both unit-specific and time-specific unobserved heterogeneity.
- **Cluster-Robust Standard Errors:** Standard errors that allow for arbitrary serial correlation and heteroskedasticity within units, obtained from the cluster covariance estimator $\hat{\mathbb{V}}_n^{\text{fe}}$.

## 9.3 Concluding Remarks

The material in this chapter borrows from several useful sources, including Stock and Watson [2003] (Chapter 10), lecture notes kindly shared by Alex Torgovitsky, and the books by Hansen [2022] and Wooldridge [2025]. The traffic fatalities data set and codes are available on Canvas.

## 9.4 Problems

**Problem 9.1** *Show that when $T = 2$, the FE and FD estimators of $\beta$ are numerically the same.*

**Problem 9.2** *Show that a projection of a random variable $W_{i,t}$ on fixed and time effects is equal to $\bar{W}_i + \bar{W}_t - \bar{W}$, where $\bar{W}_i$ is an average over time, $\bar{W}_t$ is an average over individuals, and $\bar{W}$ is a full sample average.*

**Problem 9.3** *Consider the following two research settings.*

1. *A researcher wants to estimate the effect of per-pupil spending ($X_{i,t}$) on average standardized test scores ($Y_{i,t}$) using a panel of school districts ($i = 1, \ldots, n$) observed annually over 5 consecutive years.*

   (a) *Propose at least two specific unobserved variables that could play the role of $W_i$ in equation (9.1). For each, explain why it is plausibly correlated with both per-pupil spending and test scores.*

   (b) *Are these variables plausibly constant over the 5-year sample period? Would a fixed effects approach be appropriate in this setting?*

2. *An insurance company studies the effect of the number of annual doctor visits ($X_{i,t}$) on a health outcome index ($Y_{i,t}$) using a panel of individuals observed annually over 20 years.*

   (a) *Propose at least two specific unobserved variables that could play the role of $W_i$. Are these variables plausibly constant over a 20-year period? Explain.*

   (b) *Discuss whether the fixed effects assumption that $W_i$ does not change over time is convincing in this context. Give a concrete example of how a violation of this assumption could bias the fixed effects estimator.*

**Problem 9.4** *Consider the dynamic panel data model*

$$Y_{i,t} = \rho Y_{i,t-1} + \eta_i + U_{i,t}, \quad t = 1, 2 \ ,$$

*where* $|\rho| < 1$, $E[U_{i,t}] = 0$, $\text{Var}[U_{i,t}] = \sigma_u^2$, $U_{i,1} \perp U_{i,2}$, *and* $E[U_{i,t} \mid Y_{i,0}, \eta_i] = 0$ *for* $t = 1, 2$. *The initial observation* $Y_{i,0}$ *is taken as given.*

1. *Show that strict exogeneity fails in this model. Specifically, let* $X_{i,t} = Y_{i,t-1}$ *denote the regressor at time* $t$ *and verify that* $E[X_{i,1}U_{i,1}] = 0$ *but* $E[X_{i,2}U_{i,1}] \neq 0$.

2. *Since* $T = 2$, *the FE and FD estimators coincide (Problem 9.1). Write the first-differenced equation and show that* $E[\Delta X_i \cdot \Delta U_i] = -\sigma_u^2$, *where* $\Delta X_i = Y_{i,1} - Y_{i,0}$ *and* $\Delta U_i = U_{i,2} - U_{i,1}$.

3. *What does part (b) imply about the consistency of the FE/FD estimator of* $\rho$*? Is the estimator biased upward or downward? Explain briefly.*

**Problem 9.5** *Suppose you have panel data on* $n$ *units observed in three years: 2010, 2015, and 2016, and that the model in (9.5) holds. Your friend suggests the following estimation strategy: subtract the 2010 observation from each of the other two years to obtain*

$$Y_{i,t} - Y_{i,2010} = (X_{i,t} - X_{i,2010})'\beta + (U_{i,t} - U_{i,2010}), \quad t \in \{2015, 2016\} \ ,$$

*and then run OLS of* $(Y_{i,t} - Y_{i,2010})$ *on* $(X_{i,t} - X_{i,2010})$ *using the pooled observations for* $t = 2015$ *and* $t = 2016$. *Is this a valid approach for estimating* $\beta$*? Explain why or why not.*

## Bibliography

B. Hansen. *Econometrics*. Princeton University Press, 2022.

S. D. Levitt and J. Porter. How dangerous are drinking drivers? *Journal of Political Economy*, 109(6):1198–1237, 2001.

J. Stock and M. Watson. *Introduction to Econometrics*. Prentice Hall, New York, 2003.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach 8th ed.* Cengage learning, 2025.

# 10

# *Difference in Differences*

Today we return to the problem of evaluating the causal effect of a treatment or program on an outcome of interest, denoted by $Y$. Unlike earlier approaches based on randomized experiments, selection-on-observables assumptions, or instrumental variables, our focus here is on causal inference in group-level panel data settings where treatment variation comes from natural experiments.

In the social sciences, natural experiments often arise from policy changes that affect some groups but not others—for example, when one state changes its minimum wage while a neighboring state does not. These settings are attractive because they often involve large and policy-relevant populations. At the same time, treatment assignment is not chosen by the researcher and is therefore typically not random, so simple comparisons between treated and untreated groups are generally not credible.

Difference-in-Differences (DiD) addresses this problem by comparing outcome changes over time in treated and untreated groups. We begin with the classic two-group, two-period setup, where the logic of the method is most transparent, and then move to a more general framework. We also study the connection between DiD and two-way fixed effects regressions, paying particular attention to the conditions under which that regression has a causal interpretation.

## 10.1 Two Groups and Two Periods

The simplest setup to describe the DiD approach is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. To be specific, let

$$\{(Y_{g,t}, D_{g,t}) : g \in \{s, n\} \text{ and } t \in \{1, 2\}\} \tag{10.1}$$

be the observed data, where $Y_{g,t} \in \mathbf{R}$ denotes the outcome of interest for group $g$ at time $t$ and $D_{g,t} \in \{0, 1\}$ the treatment status of group $g$ at time $t$. A group could be a location, a group of firms, a family, etc. In this simple setup, group $s$ switches from untreated to treated from period 1 to period 2,

while group $n$ is untreated in both periods. Thus, $D_{g,t} = 1$ if and only if $g = s$ and $t = 2$. Let also $Y_{g,t}(1)$ and $Y_{g,t}(0)$ denote the counterfactual outcomes of group $g$ at time $t$ under treatment and without treatment, respectively.

A natural causal parameter of interest is the average treatment effect on the treated (ATT). In this simple setup, there is only one treated group-period cell, namely $(g, t) = (s, 2)$. Therefore,

$$\theta_{\text{att}} = E[Y_{g,t}(1) - Y_{g,t}(0) \mid D_{g,t} = 1] = E[Y_{s,2}(1) - Y_{s,2}(0)] \ . \tag{10.2}$$

Of course, we may also be interested in other causal parameters, including the average treatment effect (ATE). However, in the framework we study today, identifying the ATE would require additional assumptions on the exogeneity of $D_{g,t}$ that are typically difficult to defend in the type of natural experiments where DiD tools are appropriate. For the moment, consider the following well-known example as an illustration of the $2 \times 2$ setup of this section.

**Example 10.1** On April 1, 1992, New Jersey raised the state minimum wage from \$4.25 to \$5.05. Card and Krueger [1994] collected data on employment at fast food restaurants in New Jersey in February 1992 ($t = 1$) and again in November 1992 ($t = 2$) to study the effect of increasing the minimum wage on employment. They also collected data from the same type of restaurants in eastern Pennsylvania, just across the river. The minimum wage in Pennsylvania stayed at \$4.25 throughout this period. In our notation, New Jersey would be the group $s$ that "switched," $Y_{g,t}$ would be the employment rate in location $g$ at time $t$, and $D_{g,t} = 1$ would indicate exposure to the higher minimum wage in group $g$ at time $t$. ■

The identification strategy of DiD relies on a parallel-trends assumption: in the absence of the treatment, both locations would have experienced the same average outcome evolution. Mathematically,

$$E[Y_{s,2}(0) - Y_{s,1}(0)] = E[Y_{n,2}(0) - Y_{n,1}(0)] \ , \tag{10.3}$$

i.e., both groups have "common trends" in the absence of a treatment. In the present setup, and as we show formally later in (10.9), this condition is equivalent to the additive structure

$$E[Y_{g,t}(0)] = \eta_g + \gamma_t \ , \tag{10.4}$$

where $\eta_g$ and $\gamma_t$ are (non-random) group and time effects. This additive structure for non-treated potential outcomes implies that $E[Y_{g,2}(0) - Y_{g,1}(0)] = \gamma_2 - \gamma_1 \equiv \gamma$, which is constant across groups. Note that this assumption, together with (10.2), implies that

$$E[Y_{s,2}(1)] = \theta_{\text{att}} + \eta_s + \gamma_2 \ . \tag{10.5}$$

That is, the observed mean outcome for the treated group in period 2 equals

the untreated counterfactual mean for that cell plus the ATT. In the context of the previous example, this assumption says that in the absence of a minimum wage change, expected employment is determined by the sum of a time-invariant state effect and a period effect that is common across states. Before we discuss the identifying power of this structure, we discuss two natural (but unsuccessful) approaches that may come to mind.

### 10.1.1 Pre and post comparison

A natural approach to identify $\theta_{\text{att}}$ in (10.2) would be to compare $Y_{s,2}$ and $Y_{s,1}$; that is, using outcomes before and after the policy change for the treated group alone. This approach delivers,

$$E[\Delta Y_{s,2}] = E[Y_{s,2}(1) - Y_{s,1}(0)] = \theta_{\text{att}} + \gamma \ ,$$

where $\Delta Y_{s,2} = Y_{s,2} - Y_{s,1}$ and $\gamma = \gamma_2 - \gamma_1$. Clearly, this approach does not identify $\theta_{\text{att}}$ in the presence of time trends, i.e., $\gamma \neq 0$. In the context of Example 10.1, the employment rate in New Jersey may have been going up (or down) in the absence of a policy change (the treatment), and so before and after comparisons confound the time trend as being part of the treatment effect. Unless one is willing to assume $\gamma = 0$, this approach does not identify the ATT.

### 10.1.2 Treatment and control comparison

A second natural approach to identify $\theta_{\text{att}}$ in (10.2) would be to compare $Y_{s,2}$ and $Y_{n,2}$; that is, using outcomes from both groups in the second time period. This approach delivers,

$$E[Y_{s,2} - Y_{n,2}] = E[Y_{s,2}(1) - Y_{n,2}(0)] = \theta_{\text{att}} + \eta \ ,$$

where $\eta = \eta_s - \eta_n$. Clearly, this approach does not identify $\theta_{\text{att}}$ in the presence of persistent group differences, i.e., $\eta \neq 0$. In the context of Example 10.1, the employment rate in New Jersey and Pennsylvania may be idiosyncratically different in the absence of a policy change, and so comparing these two states confounds these permanent differences with the treatment effect. Unless one is willing to assume $\eta = 0$, this approach does not identify the ATT.

### 10.1.3 Taking both differences

The DiD approach exploits the common-trends assumption in (10.3) to identify $\theta_{\text{att}}$. The idea is to remove the time trend $\gamma$ by comparing the before-after change in the treated group to the before-after change in the control group. That is, we take the difference between $\Delta Y_{s,2}$ and $\Delta Y_{n,2}$ to obtain

$$\begin{aligned} E[\Delta Y_{s,2} - \Delta Y_{n,2}] &= E[Y_{s,2}(1) - Y_{s,1}(0)] - E[Y_{n,2}(0) - Y_{n,1}(0)] \\ &= \theta_{\text{att}} + \gamma - \gamma = \theta_{\text{att}} \ . \end{aligned} \tag{10.6}$$

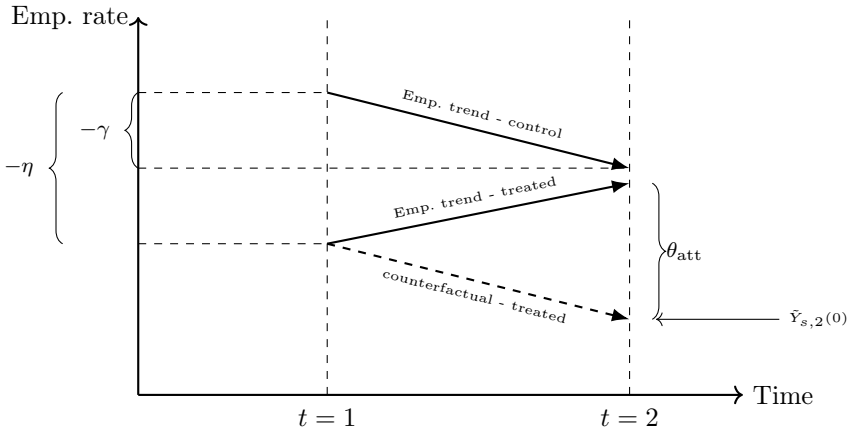



FIGURE 10.1: Causal effects in the DD model

Thus, this approach identifies $\theta_{\text{att}}$ by differencing out the common time trend.

An equivalent interpretation is to compare $(Y_{s,2} - Y_{n,2})$ and $(Y_{s,1} - Y_{n,1})$, that is, the treatment-control comparison after and before the policy change. This is because

$$\begin{aligned} E[(Y_{s,2} - Y_{n,2}) - (Y_{s,1} - Y_{n,1})] &= E[Y_{s,2}(1) - Y_{n,2}(0)] - E[Y_{s,1}(0) - Y_{n,1}(0)] \\ &= \theta_{\text{att}} + \eta - \eta = \theta_{\text{att}} \ . \end{aligned} \tag{10.7}$$

Under the maintained structure, subtracting the pre-period treated-control difference removes the persistent group effect $\eta$.

A final interpretation is that the DiD approach constructs the missing counterfactual outcome $Y_{s,2}(0)$ by combining observed untreated outcomes. Specifically, it starts from the treated group's untreated outcome in period 1 and then adds the untreated change observed in the control group. The resulting constructed counterfactual is

$$\tilde{Y}_{s,2}(0) = Y_{s,1}(0) + Y_{n,2}(0) - Y_{n,1}(0),$$

so that

$$\begin{aligned} E[\tilde{Y}_{s,2}(0)] &= E[Y_{s,1}(0) + Y_{n,2}(0) - Y_{n,1}(0)] \\ &= \eta_s + \gamma_1 + \eta_n + \gamma_2 - (\eta_n + \gamma_1) \\ &= \eta_s + \gamma_2 \ . \end{aligned}$$

Therefore, $E[Y_{s,2} - \tilde{Y}_{s,2}(0)] = \theta_{\text{att}}$, which delivers a valid identification strategy. Figure 10.1 illustrates this idea.

## 10.2 Standard Framework in DiD Models

The $2 \times 2$ case makes the logic of DiD transparent, but most applications use more than two groups or more than two periods. Additional periods allow researchers to examine pre-treatment trends, additional groups can improve precision, and policy changes often occur at different dates across locations. We therefore move to a group-level panel dataset in which groups $g$ take values in $G$, the set of groups, and periods $t$ take values in $T$, the set of time periods. For simplicity, we assume that the panel is balanced: the outcome and treatment of each group are observed in every period, which is why $T$ is not indexed by $g$. We denote the number of groups by $|G|$ and the number of time periods by $|T|$. Typically, groups are locations, such as states, counties, or municipalities, but a group could also be a subset of individuals defined by time-invariant characteristics, or even a single individual or firm.

We focus on the case where treatment is assigned at the $(g, t)$ level, as when the treatment is a county-level law or regulation, such as a minimum-wage change. We do not consider fuzzy designs, where treatment may vary within a $(g, t)$ cell and $D_{g,t}$ would denote average treatment in that cell. We also restrict attention to the case of a binary treatment, so that $D_{g,t} \in \{0, 1\}$. Several of the results below extend, with modest modifications, to treatments with multiple values or to continuous treatments, but those cases are beyond the scope of this class.

In terms of notation, let $D_g := (D_{g,t} : t \in T)$ denote the treatment path of group $g$, and let $D^{(n)} = (D_g : g \in G)$ denote the collection of treatment paths across groups. We refer to $D^{(n)}$ as the design of the study. In addition, for any $(d_1, \ldots, d_{|T|}) \in \{0, 1\}^{|T|}$, let

$$Y_{g,t}(d_1, \ldots, d_{|T|}) \tag{10.8}$$

denote the *potential outcome* of group $g$ at time $t$, and let $Y_{g,t} = Y_{g,t}(D_g)$ denote the *observed* outcome of group $g$ at time $t$. This dynamic potential-outcome framework allows a group's outcome at time $t$ to depend on its past and future treatments. Soon we will introduce assumptions that simplify this notation.

**Remark 10.1** In most of the literature on DiD, the study design $D^{(n)}$ is implicitly conditioned upon. This is in line with the focus on non-randomized natural experiments, where the researcher does not control treatment assignment and must take it as given. We follow this convention here and interpret all assumptions below as conditional on the design. Concretely, whenever we write $E[X]$, this should be understood as $E[X \mid D^{(n)}]$. Leaving this conditioning implicit greatly alleviates the notational burden. Conditional on the design, the potential outcomes are the only source of randomness left. ■

The DiD approach relies on assumptions restricting how potential outcomes depend on treatment, together with a parallel-trends assumption. We introduce these in stages.

**Assumption 10.1 (No anticipation)** *For all* $(d_1, \ldots, d_{|T|})$ *and* $g \in G$,

$$Y_{g,t}(d_1, \ldots, d_{|T|}) = Y_{g,t}(d_1, \ldots, d_t) \ .$$

**Assumption 10.2 (No dynamics)** *For all* $(d_1, \ldots, d_t)$ *and* $g \in G$,

$$Y_{g,t}(d_1, \ldots, d_t) = Y_{g,t}(d_t) \ .$$

Assumption 10.1 requires that a group's potential outcomes do not depend on future treatments. This assumption may fail, for example, when a policy change is announced in advance and agents adjust their behavior before the policy takes effect. Assumption 10.2 requires that a group's current outcome does not depend on its past treatments. This assumption is violated, for example, when the length of exposure to treatment affects the outcome. Under these two assumptions, and with a binary treatment, each cell $(g, t)$ has two potential outcomes: $Y_{g,t}(0)$ if group $g$ is untreated at $t$, and $Y_{g,t}(1)$ if group $g$ is treated at $t$. Thus, the notation reduces to the standard two-potential-outcomes setup at each $(g, t)$ cell.

The last substantive assumption we impose is parallel trends. We will use two versions of this assumption. The first is the standard formulation, appropriate when Assumptions 10.1–10.2 hold, so that potential outcomes take the form $Y_{g,t}(d_t)$.

**Assumption 10.3 (Parallel-trends)** *For all* $t \geq 2$, $E[Y_{g,t}(0) - Y_{g,t-1}(0)]$ *does not vary across* $g \in G$.

Assumption 10.3 is the standard parallel-trends condition in a setting where potential outcomes depend only on contemporaneous treatment status.

Parallel trends is an assumption about the untreated counterfactual path, so it cannot be tested directly in periods when some groups are treated. In practice, researchers try to make it plausible by choosing comparison groups exposed to similar economic shocks, by controlling for clear differences in environment when appropriate, and by examining whether treated and untreated groups followed similar trends before treatment began. Similar pre-treatment trends do not prove that Assumption 10.3 holds after treatment, but large pre-treatment differences in trends are an important warning sign.

An important implication of Assumptions 10.1–10.3 is the following additively separable representation for untreated mean potential outcomes:

$$E[Y_{g,t}(0)] = \eta_g + \gamma_t \ . \tag{10.9}$$

Finally, when we discuss inference in DiD models, we will also impose a

cross-group independence condition on potential outcomes (though not necessarily identical distributions). That is, we assume that the vectors $\{Y_{g,t}(d_t) : t \in T\}$ are independent across $g \in G$.

## 10.3 The Two-Way Fixed Effects Estimator

In the simple $2 \times 2$ model we previously discussed, the DiD contrast,

$$E[\Delta Y_{s,2} - \Delta Y_{n,2}] \ ,$$

can be equivalently obtained as the coefficient in a two-way fixed effects (TWFE) regression of $Y_{g,t}$ on treatment, group fixed effects, and period fixed effects. Motivated by this fact, researchers have also estimated TWFE regressions in more general designs with many locations and periods, variation in treatment timing, treatments switching on and off, and/or non-binary treatments. TWFE regressions are therefore widely used in empirical work.

Let the observed data be given by $\{(Y_{g,t}, D_{g,t}) : g \in G, t \in T\}$. The TWFE regression is given by the following specification,

$$Y_{g,t} = \eta_g + \gamma_t + \beta^{\text{fe}} D_{g,t} + U_{g,t} \ , \tag{10.10}$$

where $\eta_g$ is a group fixed effect, $\gamma_t$ is a time fixed effect, $\beta^{\text{fe}}$ is the coefficient of interest, and $U_{g,t}$ is a projection error. It is important to understand that these regressions are not usually interpreted as a literal linear model for $Y_{g,t}$. Rather, they are viewed as a mechanical way to compute an estimand $\beta^{\text{fe}}$ that, hopefully, admits an interesting causal interpretation. In particular, we know from the previous section that this is indeed the case in the $2 \times 2$ model, where $\beta^{\text{fe}}$ equals the ATT.

Despite its popularity, whether the TWFE estimand $\beta^{\text{fe}}$ equals the ATT (or related parameters) depends delicately on the design $D^{(n)}$. Some of the important considerations are: (a) whether all groups that are treated get treated in the same time period or not, (b) whether treated groups remain treated or may switch back to untreated, and (c) whether the treatment $D_{g,t}$ is binary, continuous, or multi-valued. For our purposes here, we discuss only two binary-treatment designs. The first, which we call the basic design, has no variation in treatment timing. The second, known as the staggered design, allows treatment timing to vary across groups but assumes that treatment is an absorbing state. More complex designs only exacerbate some of the issues that we illustrate with the staggered design.

### 10.3.1 Basic DiD Design

We start with the simplest case, where all treated groups are first treated in the same time period and remain treated thereafter.

**Definition 10.1 (Basic Design)** *The treatment assignment takes the form* $D_{g,t} = I\{t \geq t_g^*\}$*, where* $t_g^* \in \{t^*, \infty\}$ *for all* $g \in G$*, for some* $t^* \geq 2$*, and there exists* $g, g' \in G$ *such that* $t_g^* = t^*$ *and* $t_{g'}^* = \infty$*.*

That is, each group is either treated at the same date $t^*$ or never treated (normalized at $\infty$ for simplicity). In this design we can then partition $G = G_1 \cup G_0$ and $T = T_1 \cup T_0$ as follows:

$$\begin{aligned} G_1 &= \{g \in G : t_g^* = t^*\} = \text{the set of treated groups} \\ G_0 &= \{g \in G : t_g^* = \infty\} = \text{the set of control groups} \\ T_1 &= \{t \in T : t \geq t^*\} = \text{the set of treated time periods} \\ T_0 &= \{t \in T : t < t^*\} = \text{the set of control time periods} \end{aligned}$$

Using the Frisch-Waugh-Lovell decomposition, one can show that the TWFE estimator $\hat{\beta}^{\text{fe}}$ of $\beta^{\text{fe}}$ in (10.10) equals $\hat{\theta}^{\text{did}}$, where

$$\hat{\theta}^{\text{did}} = \frac{1}{|G_1|} \sum_{g \in G_1} \hat{\Delta}_g - \frac{1}{|G_0|} \sum_{g \in G_0} \hat{\Delta}_g \ , \tag{10.11}$$

and

$$\hat{\Delta}_g = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{g,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{g,t} \ . \tag{10.12}$$

The estimator $\hat{\theta}^{\text{did}}$ is the difference between the average outcome change in the treated groups before and after treatment and the average outcome change in the control groups before and after the treatment date. Moreover, in the basic design, the TWFE estimator is unbiased for the ATT under parallel trends, where the ATT is the average, over treated group-period cells, of the average treatment effect on the treated. This does not require treatment effects to be homogeneous. In particular, treatment effects may still vary across treated groups and periods.

### 10.3.2 Staggered DiD Design

In the basic design of Definition 10.1, the TWFE estimator $\hat{\beta}^{\text{fe}}$ is a simple DiD estimator, and so it identifies the ATT under a parallel-trends assumption. However, such a desirable property does not necessarily translate to more general designs and, in general, $\hat{\beta}^{\text{fe}}$ may fail to estimate an ATT-like parameter if treatment effects vary across groups and time periods.

For the purposes of this class, we restrict attention to the case where treatment continues to be binary, treatment is an absorbing state, but the timing of treatment varies across groups. This is known as the staggered design.

**Definition 10.2 (Staggered Design)** *The treatment assignment takes the form* $D_{g,t} = I\{t \geq t_g^*\}$*, with* $\min_{g \in G : t_g^* \geq 2} t_g^* < \max_{g \in G} t_g^*$*.*

Again, $t_g^*$ is the first date at which group $g$ becomes treated and, once treated, group $g$ remains treated thereafter. If $g$ never becomes treated over the study period, we let $t_g^* > |T|$, e.g., $t_g^* = \infty$. Note that $t_g^*$ may be equal to 1, meaning that group $g$ is always treated. The condition $\min_{g \in G: t_g^* \geq 2} t_g^* < \max_{g \in G} t_g^*$ rules out the basic design by requiring variation in treatment timing among groups that are untreated in period 1. If that condition fails, then the design collapses to the basic design or otherwise does not generate the timing variation that is central to the staggered setting.

Let

$$TE_{g,t} := E[Y_{g,t}(1) - Y_{g,t}(0)]$$

denote the expected treatment effect in cell $(g,t)$ of moving the treatment from 0 to 1. Theorem 1 in de Chaisemartin and D'Haultfoeuille (2020) De Chaisemartin and d'Haultfoeuille [2020] shows that under the assumptions we have made so far, $\hat{\beta}^{\text{fe}}$ is unbiased for a weighted sum of the $TE_{g,t}$. That is,

$$E[\hat{\beta}^{\text{fe}}] = \frac{1}{|G||T|} \sum_{g \in G} \sum_{t \in T} W_{g,t} TE_{g,t} \ , \tag{10.13}$$

where

$$W_{g,t} := \frac{\hat{V}_{g,t} D_{g,t}}{\frac{1}{|G||T|} \sum_{g' \in G} \sum_{t' \in T} \hat{V}_{g',t'} D_{g',t'}} \tag{10.14}$$

and $\hat{V}_{g,t}$ denotes the residual from a regression of $D_{g,t}$ on group and period fixed effects.

Note that it follows from the definition that the weights $W_{g,t}$ add up to 1, i.e.,

$$\frac{1}{|G||T|} \sum_{g \in G} \sum_{t \in T} W_{g,t} = 1 \ .$$

However, it is not true that the weights are non-negative independently of the design of the treatment assignment $D^{(n)}$. The result in the previous section implies that in the basic design $W_{g,t} = 1$ for all $(g,t) \in G_1 \times T_1$ and so, not only are the weights non-negative, but they are also all equal to one. This leads to $\hat{\beta}^{\text{fe}}$ being unbiased for the ATT. In general, however, the result in (10.13) implies that $\hat{\beta}^{\text{fe}}$ identifies a weighted average of $TE_{g,t}$ with weights that may potentially be *negative*.

In general, the weights in (10.13) need not be non-negative. Therefore, even if all group-time treatment effects $TE_{g,t}$ are positive, the TWFE estimand may still be smaller than some of those effects and may even have the opposite sign. This is one of the main reasons why TWFE can be difficult to interpret causally in staggered designs.

## 10.4 Key Concepts

- The Difference-in-Differences (DiD) approach estimates causal effects by comparing changes in outcomes over time between treated and untreated groups.
- Parallel trends assumption: in the absence of treatment, the treated and control groups would have experienced the same outcome evolution.
- DiD does not rely on before-after comparisons alone; it identifies treatment effects by comparing outcome changes in treated groups to outcome changes in control groups.
- DiD is often implemented via two-way fixed effects (TWFE) regressions, which control for time-invariant group effects and common time shocks.
- In the basic design, the TWFE estimator identifies the ATT under parallel trends.
- In staggered adoption designs, TWFE estimates a weighted average of group-time treatment effects, where some weights may be negative.

## 10.5 Concluding Remarks

I would like to thank Clément de Chaisemartin for sharing useful material that made it possible to write these notes. A large fraction of these notes are based on his book in progress, de Chaisemartin and d'Haultfoeuille [2026], and the original paper De Chaisemartin and d'Haultfoeuille [2020]. Many other important references are available in those papers.

## 10.6 Problems

**Problem 10.1** *Prove* (10.9)

**Problem 10.2** *Consider the basic design of Definition 10.1. Let*

$$\hat{\Delta}_g = \frac{1}{|T_1|}\sum_{t\in T_1} Y_{g,t} - \frac{1}{|T_0|}\sum_{t\in T_0} Y_{g,t}$$

*and*

$$\hat{\theta}^{\text{did}} = \frac{1}{|G_1|}\sum_{g\in G_1}\hat{\Delta}_g - \frac{1}{|G_0|}\sum_{g\in G_0}\hat{\Delta}_g \ .$$

*Assume that for every group $g$ and period $t$, $E[Y_{g,t}(0)] = \eta_g + \gamma_t$ and that for treated cells, $E[Y_{g,t}(1) - Y_{g,t}(0)] = \theta_{g,t}$.*

*(a) Show that*

$$E[\hat{\Delta}_g] = \frac{1}{|T_1|}\sum_{t\in T_1}\theta_{g,t} + \left(\frac{1}{|T_1|}\sum_{t\in T_1}\gamma_t - \frac{1}{|T_0|}\sum_{t\in T_0}\gamma_t\right)$$

*for $g \in G_1$.*

*(b) Show that*

$$E[\hat{\Delta}_g] = \left(\frac{1}{|T_1|}\sum_{t\in T_1}\gamma_t - \frac{1}{|T_0|}\sum_{t\in T_0}\gamma_t\right)$$

*for $g \in G_0$.*

*(c) Deduce that*

$$E[\hat{\theta}^{\text{did}}] = \frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1}\theta_{g,t} \ .$$

*(d) Explain why this shows that, in the basic design, DiD identifies the ATT even when treatment effects are heterogeneous across treated groups and periods.*

**Problem 10.3** *Consider a staggered design with three groups and three periods. Suppose treatment paths are given by*

$$D_A = (0,1,1), \qquad D_B = (0,0,1), \qquad D_C = (0,0,0) \ .$$

*(a) Identify the first treatment date $t_g^*$ for each group.*

*(b) For each treated cell $(g,t)$, list all groups that could potentially serve as controls for group $g$ at time $t$.*

**Problem 10.4** *Consider the same setting as in Problem 10.3. Let*

$$\bar{D}_{g\cdot} = \frac{1}{3}\sum_{t=1}^{3} D_{g,t} \qquad \bar{D}_{\cdot t} = \frac{1}{3}\sum_{g\in G} D_{g,t} \qquad \bar{D}_{\cdot\cdot} = \frac{1}{9}\sum_{g\in G}\sum_{t=1}^{3} D_{g,t} \ .$$

*Using the Frisch-Waugh-Lovell decomposition, the residual from regressing $D_{g,t}$ on group and period fixed effects is*

$$\hat{V}_{g,t} = D_{g,t} - \bar{D}_{g\cdot} - \bar{D}_{\cdot t} + \bar{D}_{\cdot\cdot} \ .$$

*(a) Compute $\hat{V}_{g,t}$ for every $(g,t)$ cell.*

*(b) Compute the weights*

$$W_{g,t} := \frac{\hat{V}_{g,t} D_{g,t}}{\frac{1}{|G||T|} \sum_{g' \in G} \sum_{t' \in T} \hat{V}_{g',t'} D_{g',t'}} \ .$$

*for every treated cell $(g,t)$.*

*(c) Which treated cells receive negative weights, if any?*

## Bibliography

D. Card and A. B. Krueger. Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84:772–793, 1994.

C. De Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9): 2964–96, 2020.

C. de Chaisemartin and X. d'Haultfoeuille. *Causal Inference with Differences-in-Differences: Credible Answers to Hard Questions*. SSRN, February 2026. Available at SSRN: https://ssrn.com/abstract=4487202.

# 11

# *Regression Discontinuity Design*

We have discussed how to analyze the causal effect of a treatment on outcomes of interest in experimental settings. Because many interventions of interest to economists cannot be randomly assigned, we now turn to research designs for the rigorous study of non-experimental interventions. In this lecture, we study the Regression Discontinuity (RD) design. In an RD design, each unit has a score, and treatment is assigned to units whose score exceeds a known cutoff or threshold, while units with scores below the cutoff are not treated. The key feature of the design is that the probability of receiving treatment changes abruptly at the threshold. If units cannot perfectly sort around that threshold, this discontinuous change can be used to identify a local causal effect of the treatment on the outcome of interest, using units just below the cutoff as a comparison group for units just above it.

Today we focus on the canonical sharp RD design, which has the following features: (i) the running variable is one-dimensional and continuously distributed, (ii) there is a single cutoff, and (iii) compliance with treatment assignment is perfect, so all units with scores equal to or above the cutoff receive treatment, while all units with scores below the cutoff receive control instead. We will discuss the assumptions needed to identify the RD treatment effect, methods for estimation and inference, and the graphical intuition behind the design.

## 11.1 Sharp RD: Identification

Consider the following setting: there are $n$ units, indexed by $i = 1, 2, \ldots, n$, each unit has a score $X_i$. $c$ is a known cutoff: units with $X_i \geq c$ are assigned to the treatment condition, and units with $X_i < c$ are assigned to the control condition. Thus the treatment assignment can be defined as $A_i = I\{X_i \geq c\}$. In other words, the conditional probability of receiving treatment given the score, $P\{A_i = 1 \mid X_i = x\}$, changes from 0 to 1 at $X_i = c$, as illustrated by Figure 11.1.

We adopt the potential outcomes framework and assume that each unit has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding to the outcomes that would be observed under the treatment or control conditions. The observed

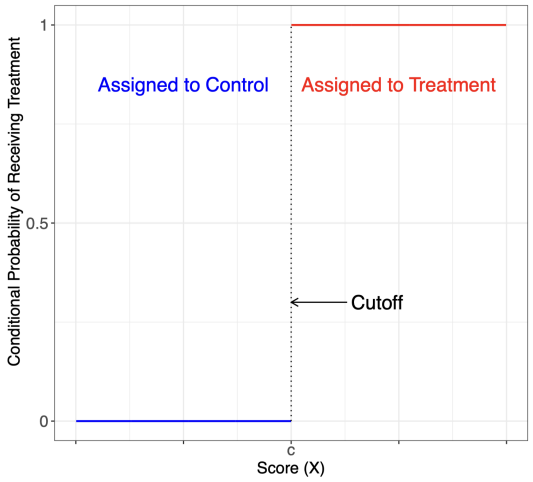



FIGURE 11.1: Conditional Probability of Receiving Treatment in the Sharp RD Design

outcome is

$$Y_i = (1 - A_i) \cdot Y_i(0) + A_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } X_i < c \\ Y_i(1) & \text{if } X_i \geq c \end{cases}.$$

The fundamental problem of causal inference occurs because we only observe the outcome under control, $Y_i(0)$, for those units whose score is below the cutoff, and we only observe the outcome under treatment, $Y_i(1)$, for those above the cutoff. As shown in Figure 11.2, the regression function $E[Y_i(1) \mid X_i = x]$ is observed for values of the score to the right of the cutoff, represented with the solid red line. However, to the left of the cutoff, all units are untreated, and therefore $E[Y_i(1) \mid X_i = x]$ is not observed (represented by a dashed red line). A similar phenomenon occurs for $E[Y_i(0) \mid X_i = x]$, which is observed for values of the score to the left of the cutoff (solid blue line), but unobserved for $x \geq c$ (dashed blue line). Thus, the observed average outcome given the score is

$$E[Y_i \mid X_i = x] = \begin{cases} E[Y_i(0) \mid X_i = x] & \text{if } x < c \\ E[Y_i(1) \mid X_i = x] & \text{if } x \geq c \end{cases}. \tag{11.1}$$

The average treatment effect at a given value of the score,

$$E[Y_i(1) \mid X_i = x] - E[Y_i(0) \mid X_i = x] \ ,$$

is the vertical distance between the two regression curves at that value. This distance cannot be directly estimated because we never observe both curves for the same value of $x$. However, the cutoff $c$ is special because it is the point at which treatment status changes discontinuously. To build intuition,

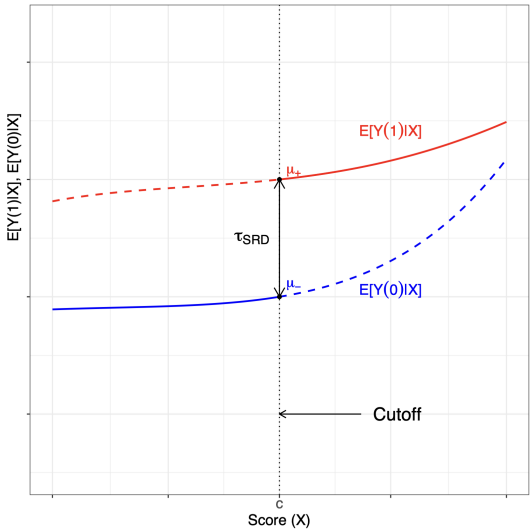



FIGURE 11.2: RD Treatment Effect in Sharp RD Design

compare units with scores just above $c$ to units with scores just below $c$. The former receive treatment, while the latter receive control. If the average potential outcomes vary smoothly with the score at the cutoff, then units on either side of the cutoff are very similar except for their treatment status. In that case, the difference between the right-hand and left-hand limits of the observed regression function at $c$ recovers the vertical distance between the two potential-outcome regression functions at the cutoff. This motivates the Sharp RD treatment effect

$$\theta_{\text{srd}} \equiv E\left[Y_i(1) - Y_i(0) \mid X_i = c\right] = E\left[Y_i(1) \mid X_i = c\right] - E\left[Y_i(0) \mid X_i = c\right].$$

The assumption of comparability between units with very similar values of the score but on opposite sides of the cutoff can be formalized as the continuity of the potential-outcome regression functions at $x = c$.

**Assumption 11.1 (Continuity)** *The conditional expectations* $E\left[Y_i(1) \mid X_i = x\right]$ *and* $E\left[Y_i(0) \mid X_i = x\right]$ *are continuous in* $x$ *at* $x = c$.

This assumption says that, in the absence of the discontinuous treatment rule, average potential outcomes would evolve smoothly as the score crosses the cutoff. In practice, the main concern is sorting or manipulation: if units can precisely choose whether to fall just above or just below the threshold, then units on the two sides may differ in ways other than treatment status.

Under Assumption 11.1, we have

$$\begin{aligned} E\left[Y_i(1) - Y_i(0) \mid X_i = c\right] &= E\left[Y_i(1) \mid X_i = c\right] - E\left[Y_i(0) \mid X_i = c\right] \\ &= \lim_{x \downarrow c} E\left[Y_i(1) \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i(0) \mid X_i = x\right] \\ &= \lim_{x \downarrow c} E\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i \mid X_i = x\right], \quad (11.2) \end{aligned}$$

where the second equality is due to the continuity assumption, and the last equality is due to (11.1). The last expression is the difference between the right-hand and left-hand limits of the observed average outcomes as the score converges to the cutoff, which is identifiable from the data.

The Sharp RD parameter presented above can be interpreted as causal in the sense that it captures the average difference in potential outcomes under treatment versus control. However, this average difference is defined at a single point on the support of the score, and as a result is local in nature. In a sharp RD design, $\theta_{\text{srd}}$ is an ATE conditional on $X_i = c$, not an average effect for all units in the population. This locality is both a strength and a limitation: identification relies on relatively weak smoothness assumptions, but the resulting effect is most directly relevant for units near the threshold.

## 11.2 Estimation Under Local Linearity

By (11.2), to estimate the sharp RD effect $\theta_{\text{srd}} = E[Y_i(1) - Y_i(0) \mid X_i = c]$, we must approximate the limits

$$\lim_{x \downarrow c} E[Y_i \mid X_i = x] \quad \text{and} \quad \lim_{x \uparrow c} E[Y_i \mid X_i = x].$$

In modern empirical work, RD effects are typically estimated using nonparametric local polynomial methods, often combined with kernel weighting and data-driven bandwidth selection. More broadly, researchers may also use other flexible smoothing or machine learning methods to approximate the conditional expectation functions near the cutoff. However, because our goal is to introduce the logic of RD without requiring much background in nonparametrics, we will study the simplest version of this idea: local linear regression on each side of the cutoff within a small bandwidth. This linearity assumption is not required for identification of the treatment effect, but it provides a tractable and intuitive estimation strategy.

We also note that the choice of the bandwidth $h$ below (which determines the local neighborhood) is the most important consideration when applying the method we introduce in this section. A smaller bandwidth uses observations closer to the cutoff and therefore reduces bias from misspecifying the regression function, but it also leaves fewer observations and increases sam-

pling variability. A larger bandwidth has the opposite trade-off. A formal discussion on how to choose this parameter is beyond the scope of our class.

**Assumption 11.2 (Local Linearity)** *There exists a bandwidth $h > 0$ such that*

$$E[Y_i(0) \mid X_i = x] = \beta_0^- + \beta_1^-(x - c), \quad \text{for all } x \in [c - h, c)$$
$$E[Y_i(1) \mid X_i = x] = \beta_0^+ + \beta_1^+(x - c), \quad \text{for all } x \in [c, c + h] \ .$$

For simplicity, this assumption states that the conditional expectation functions for the untreated and treated potential outcomes are exactly linear within a neighborhood of the cutoff $c$. The parameters $\beta_1^-$ and $\beta_1^+$ allow the slope of the regression function to differ on each side of the cutoff, while the constants $\beta_0^-$ and $\beta_0^+$ capture the levels at the cutoff.

Under Assumption 11.2, we have

$$\lim_{x \uparrow c} E[Y_i \mid X_i = x] = \beta_0^- + \beta_1^-(c - c) = \beta_0^-,$$

$$\lim_{x \downarrow c} E[Y_i \mid X_i = x] = \beta_0^+ + \beta_1^+(c - c) = \beta_0^+.$$

That is, the intercept terms in the regression capture precisely the conditional expectations we need to characterize the RD treatment effect. It follows that

$$\theta_{\text{srd}} = \beta_0^+ - \beta_0^- \ .$$

This result provides a direct estimation strategy: estimate the constants $\beta_0^+$ and $\beta_0^-$ using linear regression on each side of the cutoff, and take their difference. Concretely, consider the following two step estimator:

Step 1 Run a regression of $Y$ on $\tilde{X}_i = (1, X_i - c)'$ using *only* observations to the left of the cutoff $c$ and within the tolerance of the bandwidth $h$. This leads to

$$\begin{pmatrix} \hat{\beta}_0^- \\ \hat{\beta}_1^- \end{pmatrix} = \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i' I\{c - h \le X_i < c\} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i Y_i I\{c - h \le X_i < c\} \ .$$

Step 2 Run a regression of $Y$ on $\tilde{X}_i = (1, X_i - c)'$ using *only* observations to the right of the cutoff $c$ and within the tolerance of the bandwidth $h$. This leads to

$$\begin{pmatrix} \hat{\beta}_0^+ \\ \hat{\beta}_1^+ \end{pmatrix} = \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i' I\{c \le X_i \le c + h\} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i Y_i I\{c \le X_i \le c + h\} \ .$$

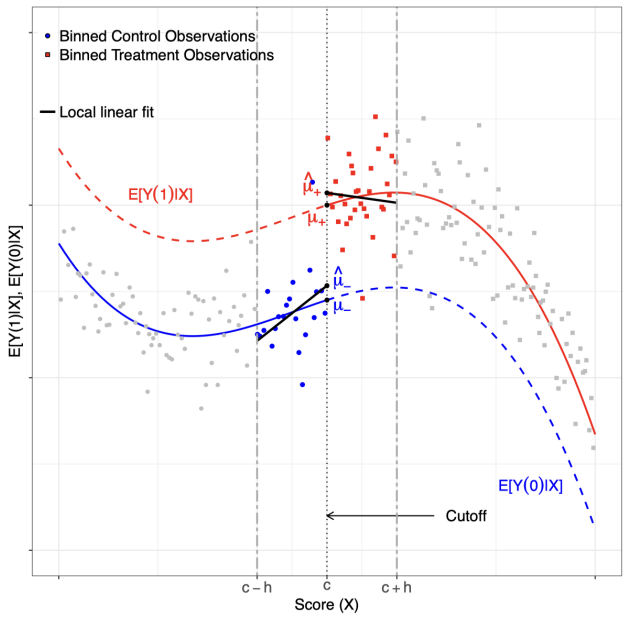



FIGURE 11.3: RD Estimation with Local Linear Regression

**Definition 11.1 (Local Linear RD Estimator)** *Let $\hat{\beta}_0^+$ and $\hat{\beta}_0^-$ be the intercepts from linear regressions of $Y_i$ on a constant and $(X_i - c)$, estimated using observations with $X_i \in [c, c+h]$ and $X_i \in [c-h, c)$, respectively. Then,*

$$\hat{\theta}_{\text{srd}} = \hat{\beta}_0^+ - \hat{\beta}_0^- .$$

A graphical representation of local linear RD estimation is given in Figure 11.3, where a linear regression is run on each side of the cutoff but only using observations within the area determined by the bandwidth; observations outside bandwidth $h$ are not used in the estimation. The figure denotes fitted values at the cutoff by $\mu_-$ and $\mu_+$; these correspond to the intercepts $\beta_0^-$ and $\beta_0^+$ in Definition 11.1.

The slope terms $\beta_1^+$ and $\beta_1^-$ allow for smooth trends in outcomes near the cutoff and help reduce bias in the estimated intercepts relative to simple averages on each side of the threshold. In practice, researchers often use local polynomial methods, such as local linear or local quadratic regression, and may also weight observations by their distance from the cutoff.

## 11.3 Validity Checks

Assumption 11.1 is not directly testable because it concerns potential outcomes that are not both observed at the cutoff. Nevertheless, RD papers usually report falsification checks that would be expected to fail if there were precise sorting around the threshold. Two standard checks are especially useful. First, researchers examine whether the density of the running variable has a discontinuity at the cutoff; a jump in the density suggests that units may be manipulating the score to end up on one side. Second, researchers test whether predetermined covariates, such as baseline demographics or lagged outcomes, change discontinuously at the cutoff. Such jumps would suggest that units just above and just below the cutoff are not comparable even before treatment.

## 11.4 Empirical Example

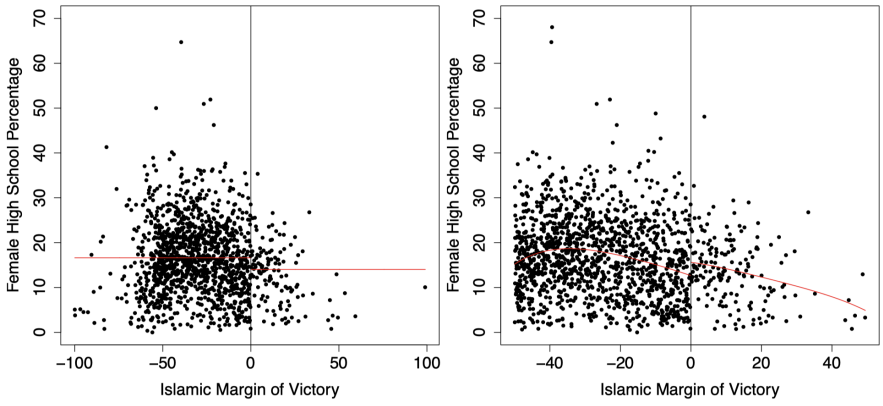
We now introduce an empirical example, originally analyzed by Meyersson [2014]. The study employs a Sharp RD design based on close elections in Turkey to evaluate the impact of Islamic party mayors on various outcomes. The unit of analysis is the municipality, and the running variable is the Islamic margin of victory—defined as the vote share of the top Islamic party minus that of the top secular opponent. The cutoff for treatment is zero: municipalities with a margin above zero elected an Islamic mayor in 1994 ($A = 1$); those below elected a secular mayor ($A = 0$). The outcome of interest is the educational attainment of women, measured as the percentage of women aged 15–20 in 2000 who had completed high school.

We summarize the main variables:

- $Y$: high school completion rate among women aged 15–20 in 2000.
- $X$: Islamic margin of victory in the 1994 mayoral election.
- $A$: treatment indicator, equal to 1 if the Islamic party won the 1994 election, and 0 otherwise.

A simple comparison of municipalities by party may not be valid, as places electing Islamic mayors may differ systematically from those electing secular mayors in ways that also affect women's education. For example, municipalities with stronger religious conservatism may be both more likely to elect an Islamic mayor and less likely to invest in female schooling. A naive comparison would therefore mix the causal effect of party control with these pre-existing differences.

The RD design addresses this concern by focusing on municipalities with




(a) Raw Comparison of Means

(b) Local Comparison of Means

FIGURE 11.4: Municipalities with Islamic Mayor versus Municipalities with Secular Mayor. Panel (a) compares the full sample; Panel (b) focuses on close elections near the cutoff.

election outcomes very close to the cutoff at zero. Municipalities just above the cutoff barely elected an Islamic mayor, while municipalities just below the cutoff barely elected a secular mayor. If potential outcomes vary smoothly with the Islamic margin of victory at the threshold, then these municipalities are comparable except for the identity of the winning party. The discontinuity in outcomes at the cutoff can therefore be interpreted as the causal effect of electing an Islamic mayor for municipalities with very close elections.

This logic is illustrated in Figure 11.4. Panel (a) shows that, in the full sample, municipalities with an Islamic mayor have lower average female high-school completion. However, this raw difference is not causal because it reflects broad differences across municipalities. Panel (b) instead focuses on observations near the cutoff and fits flexible trends on each side. The key object is the jump at zero: this discontinuity captures the local treatment effect of electing an Islamic mayor in close elections.

To make the discontinuity at the cutoff easier to see, Figure 11.5 presents a binned RD plot of the same relationship. Such plots are popular in empirical RD papers because they make the regression functions on either side of the threshold easier to visualize. However, they do not alter the parameter of interest or the underlying estimator. Instead, they serve only as a graphical representation of the same local discontinuity, displayed in a way that makes the jump at the cutoff more transparent.

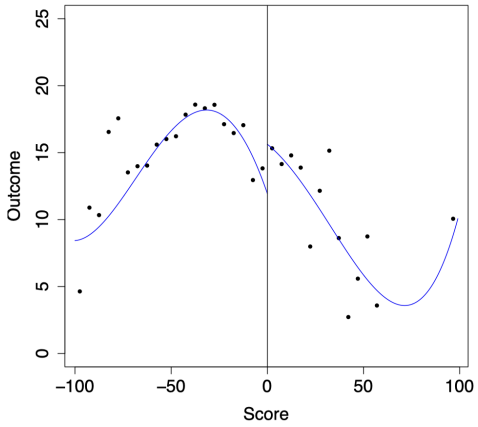



FIGURE 11.5: Binned RD Plot for the Meyersson Data

## 11.5 Key Concepts

- In a sharp RD design, treatment assignment changes deterministically at a known cutoff value of a running variable.
- The key comparison is between units just below and just above the cutoff, who are assumed to be similar except for treatment status.
- The parameter identified by the design is local: $\theta_{\text{srd}} = E[Y_i(1) - Y_i(0) \mid X_i = c]$, the average treatment effect at the cutoff.
- Identification requires the conditional expectations of the treated and untreated potential outcomes to be continuous at the cutoff.
- Under local linearity near the cutoff, the RD effect can be estimated as the difference between the fitted intercepts from separate regressions on the two sides of the threshold.
- In practice, the choice of bandwidth matters because the estimator is based on observations in a neighborhood around the cutoff.
- Common validity checks examine whether the density of the running variable or predetermined covariates jump at the cutoff.

## 11.6 Concluding Remarks

The material today heavily borrows from two books on RDD: Cattaneo et al. [2019] and Cattaneo et al. [2024], as well as slides shared by Matias Cattaneo. I want to particularly thank Matias for sharing his slides and code with me.

## 11.7 Problems

**Problem 11.1** *Consider a sharp RD design with cutoff $c = 0$. Suppose potential outcomes satisfy*

$$E[Y_i(0) \mid X_i = x, T_i = t] = x + \delta t, \qquad E[Y_i(1) \mid X_i = x, T_i = t] = \tau + x + \delta t,$$

*where $T_i \in \{0, 1\}$ is an unobserved type and $\tau$ and $\delta$ are constants.*

*Treatment is assigned by the rule $A_i = I\{X_i \geq 0\}$, so the observed outcome is $Y_i = (1 - A_i)Y_i(0) + A_i Y_i(1)$. Suppose that near the cutoff,*

$$P\{T_i = 1 \mid X_i = x\} = \begin{cases} p_-, & x < 0, \\ p_+, & x \geq 0, \end{cases}$$

*with $p_+ \neq p_-$.*

1. *What is the true treatment effect at the cutoff?*
2. *Compute $\lim_{x \uparrow 0} E[Y_i \mid X_i = x]$.*
3. *Compute $\lim_{x \downarrow 0} E[Y_i \mid X_i = x]$.*
4. *Show that the observed jump at the cutoff is*

$$\lim_{x \downarrow 0} E[Y_i \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i \mid X_i = x] = \tau + \delta(p_+ - p_-).$$

5. *Explain why the RD design fails to identify the causal effect in this example, even though treatment assignment still changes discontinuously at the cutoff.*

**Problem 11.2** *In an RD design, a standard validity check is to test whether predetermined covariates show a discontinuity at the cutoff. Let $Z_i$ be a covariate determined before treatment assignment, such as a baseline demographic characteristic.*

1. *If units just above and just below the cutoff are comparable, what should we expect about*

$$\lim_{x \downarrow c} E[Z_i \mid X_i = x] \qquad \textit{and} \qquad \lim_{x \uparrow c} E[Z_i \mid X_i = x]?$$

2. *Suppose a researcher finds a large and statistically significant jump in $Z_i$ at the cutoff. What does this suggest about the validity of the RD design?*

3. *In the Meyersson application, suggest one predetermined covariate that would be informative as a validity check, and explain why a jump in that variable at the electoral cutoff would be concerning.*

**Problem 11.3** *Suppose the cutoff is $c = 0$, and within a symmetric bandwidth $[-h, h]$ the observed conditional expectation of the outcome is*

$$E[Y_i \mid X_i = x] = \begin{cases} \beta_0^- + \beta_1^- x, & -h \le x < 0, \\ \beta_0^+ + \beta_1^+ x, & 0 \le x \le h. \end{cases}$$

*Assume that the running variable is uniformly distributed on $[-h, 0)$ and on $[0, h]$, with the same number of observations on each side of the cutoff.*

*A researcher proposes the following simple estimator:*

$$\tilde{\theta} = \bar{Y}_+ - \bar{Y}_-,$$

*where $\bar{Y}_+$ is the average outcome for observations with $X_i \in [0, h]$ and $\bar{Y}_-$ is the average outcome for observations with $X_i \in [-h, 0)$.*

1. *Show that*

$$E[\bar{Y}_+] = \beta_0^+ + \beta_1^+ \frac{h}{2} \qquad \textit{and} \qquad E[\bar{Y}_-] = \beta_0^- - \beta_1^- \frac{h}{2}.$$

2. *Deduce that*

$$E[\tilde{\theta}] = (\beta_0^+ - \beta_0^-) + \frac{h}{2}(\beta_1^+ + \beta_1^-).$$

3. *Recall that the sharp RD estimand is $\theta_{\text{srd}} = \beta_0^+ - \beta_0^-$. Explain why $\tilde{\theta}$ generally does not recover $\theta_{\text{srd}}$.*

4. *Under what special condition would $\tilde{\theta}$ coincide with $\theta_{\text{srd}}$?*

5. *In one or two sentences, explain why local linear regression is preferable to this simple difference-in-means estimator.*

## Bibliography

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press, 2019.

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press, 2024.

E. Meyersson. Islamic rule and the empowerment of the poor and pious. *Econometrica*, 82(1):229–269, 2014.

# Part III

# Causality and Endogeneity

# 12

# *Endogeneity*

In the previous chapters, selection on observables meant that, after conditioning on observed covariates, treatment assignment could be treated as unrelated to the relevant potential outcomes. This assumption is often difficult to justify in economic applications. Broadly speaking, unobserved confounders, such as preferences, private information, or expectations, tend to be the rule rather than the exception. In such settings, concerns about endogeneity, selection, and related issues naturally arise. More generally, if our goal is to identify the **causal effect** of a variable $X$ on an outcome $Y$, the presence of unobserved confounders that are correlated with $X$ (hence the name *confounders*) complicates matters. In these cases, it becomes difficult to disentangle whether the observed relationship is driven by $X$ itself or by some omitted variable $U$, as illustrated in Figure 12.1 below.

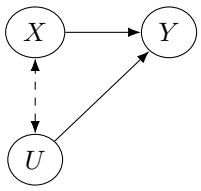



FIGURE 12.1: Unobserved factors are confounded with observed factors

Understanding the implications of endogeneity is essential for both theoretical and applied econometric work. We start our discussion of endogeneity within the context of the linear model, and then move on to more general settings.

## 12.1 Endogeneity in Linear Regression

Let $(Y, X, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$ and $X$ takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of $X$ is constant and equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let

$\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

In contrast to our earlier discussion, we do not assume that $E[XU] = 0$. Any $X_j$ such that $E[X_jU] = 0$ is said to be *exogenous*; any $X_j$ such that $E[X_jU] \neq 0$ is said to be *endogenous*. By normalizing $\beta_0$ if necessary, we assume $X_0$ is exogenous. Note that it must be the case that we are interpreting this regression as a causal model.

Note that when $E[XU] \neq 0$ we have that

$$E[XY] = E[XX']\beta + E[XU]$$

which implies that

$$E[XX']^{-1}E[XY] = \beta + E[XX']^{-1}E[XU] \ .$$

The results from the previous class showed that the least squares estimator $\hat{\beta}_n$ of $\beta$ converges in probability to $E[XX']^{-1}E[XY]$. It follows that

$$\hat{\beta}_n \xrightarrow{P} \beta + E[XX']^{-1}E[XU] \ , \tag{12.1}$$

and is therefore inconsistent for $\beta$ when $E[XU] \neq 0$, i.e., under endogeneity.

We now briefly review some common ways in which endogeneity may arise. In many of these cases, we focus on the simple case with a single regressor ($k = 1$), where the model takes the form

$$Y = \beta_0 + \beta_1 X_1 + U \ ,$$

and the asymptotic bias in (12.1) simplifies to

$$\hat{\beta}_{1,n} \xrightarrow{P} \beta_1 + \frac{\mathrm{Cov}[X_1, U]}{\mathrm{Var}[X_1]} \ . \tag{12.2}$$

### 12.1.1 Omitted Variables

To illustrate one source of endogeneity, consider a linear model with two covariates:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \ ,$$

where we interpret this as a causal model and assume exogeneity holds; that is,

$$E[U] = E[X_1U] = E[X_2U] = 0 \ .$$

Suppose, however, that one of the relevant covariates, $X_2$, is unobserved. A classic example of this situation arises in labor economics, where $Y$ denotes wages, $X_1$ is years of education, and $X_2$ captures unobserved ability.

Because $X_2$ is unobserved, we are forced to estimate a misspecified model omitting this covariate:

$$Y = \beta_0^* + \beta_1^* X_1 + U^* \ ,$$

where the new error term $U^*$ absorbs the influence of the omitted variable. Under the standard omitted variable framework, we can express the parameters of the reduced-form model as:

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_2 E[X_2] \ , \\ \beta_1^* &= \beta_1 \ , \\ U^* &= \beta_2 (X_2 - E[X_2]) + U \ . \end{aligned}$$

Note that we have normalized $\beta_0^*$ so that $E[U^*] = 0$. This representation makes clear that even if the original error term $U$ is uncorrelated with $X_1$, the new composite error term $U^*$ will generally not be:

$$E[X_1 U^*] = \beta_2 \operatorname{Cov}(X_1, X_2) \neq 0 \ ,$$

unless $X_1$ and $X_2$ are uncorrelated, or $\beta_2$ equals zero. In that case, $X_1$ is endogenous in the reduced-form model, and OLS estimation of $\beta_1^*$ will generally be biased and inconsistent for the causal parameter $\beta_1$.

Using the expression in (12.2), it follows immediately that running a regression of $Y$ on $(1, X_1)$ leads to a LS estimator of the slope coefficient satisfying:

$$\hat{\beta}_{1,n}^* \xrightarrow{P} \frac{\operatorname{Cov}[X_1, Y]}{\operatorname{Var}[X_1]} = \frac{\operatorname{Cov}[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U]}{\operatorname{Var}[X_1]} = \beta_1 + \beta_2 \frac{\operatorname{Cov}[X_1, X_2]}{\operatorname{Var}[X_1]} \ ,$$

where the term

$$\beta_2 \frac{\operatorname{Cov}[X_1, X_2]}{\operatorname{Var}[X_1]} \ ,$$

is usually referred to as *omitted variable bias*.

### 12.1.2 Measurement Error

Partition $X$ into $X_0$ and $X_1$, where $X_0 = 1$ and $X_1$ takes values in $\mathbf{R}$. Partition $\beta$ analogously. We note that while we focus on the scalar case, the result we discuss in this section extends to the case where $X_1$ takes values in $\mathbf{R}^k$. In this notation,

$$Y = \beta_0 + \beta_1 X_1 + U \ .$$

We are interpreting this regression as a causal model and are willing to assume that $E[XU] = 0$, but $X_1$ is *not* observed. Instead, $\hat{X}_1$ is observed, where

$$\hat{X}_1 = X_1 + V \ .$$

Assume $E[V] = 0$, $\operatorname{Cov}[X_1, V] = 0$, and $\operatorname{Cov}[U, V] = 0$. We may therefore rewrite this model as

$$Y = \beta_0^* + \beta_1^* \hat{X}_1 + U^* \ ,$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 \\ \beta_1^* &= \beta_1 \\ U^* &= -\beta_1 V + U \ . \end{aligned}$$

In this model,

$$E[\hat{X}_1 U^*] = -\beta_1 E[\hat{X}_1 V] = -\beta_1 E[V^2] = -\beta_1 \operatorname{Var}[V] \ ,$$

so $\hat{X}_1$ is typically endogenous. Using the expression in (12.2), it follows that running a regression of $Y$ on $\hat{X}_1$ produces an estimator $\hat{\beta}^*_{1,n}$ with the property that

$$\begin{aligned} \hat{\beta}^*_{1,n} &\xrightarrow{P} \frac{\operatorname{Cov}[\hat{X}_1, Y]}{\operatorname{Var}[\hat{X}_1]} \\ &= \frac{\operatorname{Cov}[\hat{X}_1, \beta_0^* + \beta_1^* \hat{X}_1 + U^*]}{\operatorname{Var}[\hat{X}_1]} \\ &= \beta_1 + \frac{\operatorname{Cov}[\hat{X}_1, U^*]}{\operatorname{Var}[\hat{X}_1]} \\ &= \beta_1 + \frac{E[\hat{X}_1 U^*]}{\operatorname{Var}[\hat{X}_1]} \\ &= \beta_1 \left( 1 - \frac{\operatorname{Var}[V]}{\operatorname{Var}[\hat{X}_1]} \right) \ , \end{aligned}$$

where the second equality follows by using the fact that the covariance between a random variable and a constant is zero, the third equality follows from the fact that $E[U^*] = E[-\beta_1 V + U] = 0$, and the fourth equality from the previous derivation. Importantly, note that the regression coefficient is biased towards zero when the regressor of interest is measured with the so-called classical random errors. If $\beta_1 > 0$, the probability limit is smaller than $\beta_1$; if $\beta_1 < 0$, it is larger than $\beta_1$ but still closer to zero. Indeed, in the extreme case where $\hat{X}_1 = V$, it follows that $\hat{\beta}^*_{1,n} \xrightarrow{P} 0$.

### 12.1.3 Simultaneity

A third source of endogeneity arises when the regressor and the outcome are jointly determined. The classic example is a market in which price and quantity are determined at the same time by supply and demand. If we observe many pairs $(P, Q)$ of equilibrium prices and quantities, those points do not necessarily trace out a demand curve or a supply curve. They reflect both curves moving around at the same time.

To make this concrete, suppose demand and supply are given by

$$Q^d = \beta_0^d + \beta_1^d P + U^d \ ,$$
$$Q^s = \beta_0^s + \beta_1^s P + U^s \ ,$$

where $U^d$ is an unobserved demand shock and $U^s$ is an unobserved supply shock. In equilibrium, $Q^d = Q^s$, so the observed price $P$ must adjust until the two equations agree. Solving the two equations gives

$$P = \frac{\beta_0^d - \beta_0^s + U^d - U^s}{\beta_1^s - \beta_1^d} \ .$$

This expression shows why price is endogenous. A positive demand shock raises the equilibrium price, so price is correlated with the unobserved demand shock. A supply shock also changes the equilibrium price. Thus, if we regress observed quantity $Q$ on observed price $P$, the slope generally mixes movements along demand with shifts in demand and supply. It need not recover either the demand slope $\beta_1^d$ or the supply slope $\beta_1^s$.

This is exactly the kind of problem for which instrumental variables are useful. To identify demand, we would like a variable that shifts supply but does not directly shift demand—for example, a weather or input-cost shock. Such a variable moves prices for reasons unrelated to the demand shock, allowing us to isolate variation in price that is more plausibly exogenous for the demand equation.

## 12.2 Instrumental Variables

To address the challenge posed by endogeneity, that is, the failure of the condition $E[XU] = 0$ in the model

$$Y = X'\beta + U \ ,$$

we introduce an additional random vector called *instruments*, denoted by $Z$. Specifically, we assume:

$$Z = (Z_0, Z_1, \ldots, Z_\ell)' \in \mathbf{R}^{\ell+1}, \quad \text{with } \ell + 1 \geq k + 1 \ .$$

We assume that any exogenous component of $X$ is contained in $Z$ (the so-called included instruments). In particular, we assume that the first component of $Z$ is constant and equal to one, i.e., $Z = (Z_0, Z_1, \ldots, Z_\ell)'$ with $Z_0 = 1$. However, since there are endogenous components in $X$, it must be the case that $Z$ includes variables that are *not* components of $X$ (the so-called excluded instruments).[1] A leading example we will return to in the next chapter

[1] Here "excluded" means excluded from the vector of regressors $X$. In applied IV discussions, the exclusion restriction often refers more broadly to the assumption that the instrument affects $Y$ only through the endogenous regressor.

is an experiment with imperfect compliance, where random assignment is not the same as the treatment actually received.

We impose the following conditions on $Z$:

**Instrument Exogeneity:** The instruments are uncorrelated with the structural error term $U$:

$$E[ZU] = 0 \ .$$

This ensures that any correlation between $Z$ and $Y$ operates through $X$, not through omitted variables in $U$.

**Instrument Relevance (Rank Condition):** The instruments must be sufficiently correlated with the endogenous regressors. Specifically, we require:

$$\text{rank}(E[ZX']) = k + 1 \ .$$

This ensures that the system of moment conditions has a unique solution for $\beta$.

**Order Condition:** A necessary condition for relevance is that the number of instruments (excluding the intercept) is at least as large as the number of endogenous regressors:

$$\ell \geq k \ .$$

The order condition is necessary but not sufficient for relevance; the rank condition is the operative requirement for identification.

**Regularity Conditions:** We also assume the following expectations are finite:

$$E[ZX'] < \infty \quad \text{and} \quad E[ZZ'] < \infty \ ,$$

and that there is no perfect collinearity among the components of $Z$.

These properties are essential for identification of the causal effect $\beta$ in the presence of endogeneity. Throughout this chapter, we maintain the linear homogeneous-effects interpretation of the structural equation.

Using the identity $U = Y - X'\beta$, we can pre-multiply both sides by $Z$ and take expectations to obtain:

$$E[ZY] = E[ZX']\beta \ . \tag{12.3}$$

The system of equations in (12.3) characterizes the population moment condition implied by instrumental variables. When $\ell + 1 \geq k + 1$, the system may be over-identified—that is, the number of equations exceeds the number of unknowns. To obtain an explicit solution for $\beta$, we consider two broad cases: (a) the *just-identified* case where $\ell = k$, and (b) the *over-identified* case where we must exploit additional structure, such as full-rank conditions on $E[ZX']$, to uniquely identify $\beta$. We begin by analyzing the just-identified case, which is both conceptually simpler and pedagogically useful.

### 12.2.1 The "just" identified case: $\ell = k$

Instrumental variables are easier to understand when the number of instruments matches the number of regressors — i.e., when $\ell = k$. In this just-identified setting, the system of equations

$$E[ZY] = E[ZX']\beta$$

has the same number of equations as unknowns. Provided that the matrix $E[ZX']$ is invertible, we can solve directly for $\beta$ without further technical assumptions. In particular, note that

$$\beta = (E[ZX'])^{-1}E[ZY] \ . \tag{12.4}$$

In this case, we say that $\beta$ is *exactly identified*. This case serves as a useful stepping stone for understanding instrumental variables estimation and is often the first one tackled in applied work. For this reason, we begin our formal development of IV methods by analyzing this special case.

An especially tractable and instructive case arises when $\ell = k = 1$, so that $X = (1, X_1)'$, $Z = (1, Z_1)'$, and the model is:

$$Y = \beta_0 + \beta_1 X_1 + U$$

with the assumption $E[U] = E[Z_1 U] = 0$. In this case it can be shown (see Problem 12.1) that

$$\beta_1 = \frac{\text{Cov}(Z_1, Y)}{\text{Cov}(Z_1, X_1)} \ . \tag{12.5}$$

This expression makes the roles of the two key IV conditions particularly transparent. The denominator reflects the *relevance* condition, requiring that the instrument $Z_1$ be correlated with the endogenous regressor $X_1$. This prevents situations with zero denominators. The numerator reflects the correlation between the instrument and the outcome, which — under the *exogeneity* condition $E[Z_1 U] = 0$ — is driven solely by the causal pathway from $Z_1$ to $X_1$ to $Y$.

These relationships can be summarized graphically via a directed acyclic graph (DAG) in Figure 12.2. The instrument $Z_1$ is *relevant* because it is correlated with $X_1$, and it is valid because it does not have a direct path to $Y$ except through $X_1$ - *excluded* - and because it is uncorrelated with the structural error term $U$ - *exogenous*.

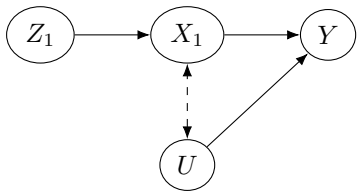



FIGURE 12.2: An instrument $Z_1$ that is both *relevant* and *excluded*.

### 12.2.2 The "over" identified case: $\ell > k$

When $\ell > k$, the system of equations in (12.3) is over-identified—that is, the number of equations exceeds the number of unknowns. To uniquely identify $\beta$, we introduce the following useful Lemma.

**Lemma 12.1** *Suppose there is no perfect collinearity in $Z$ and let $\Pi$ be such that $BLP(X|Z) = \Pi' Z$; i.e.,*

$$X = \Pi' Z + V \quad \text{with} \quad E[ZV'] = 0 \ .$$

*Then, (a) $E[ZX']$ has rank $k+1$ if and only if $\Pi$ has rank $k+1$; and (b) if $\Pi$ has rank $k+1$, the matrix $\Pi' E[ZX']$ is invertible.*

The lemma above provides a crucial insight: First, by part (a) we learn that the rank condition on $E[ZX']$ holds if and only if $\Pi$ has full column rank, where $\Pi$ collects the regression coefficients of a regression of the variables in $X$ onto the variables $Z$. Second, by part (b) we learn that even though $E[ZX']$ is not a square matrix when $\ell > k$, the matrix $\Pi' E[ZX']$ is not only square, but also *invertible* whenever the rank condition holds.

The second insight is precisely what allows us to obtain a unique solution for $\beta$, since

$$E[ZY] = E[ZX']\beta \quad \text{which implies} \quad \Pi' E[ZY] = \Pi' E[ZX']\beta$$

has a unique solution by virtue of the matrix $\Pi' E[ZX']$ being invertible. This condition, known as the **rank condition**, guarantees that the instruments provide sufficiently rich and independent variation to pin down each element of $\beta$. If the rank condition fails, then some components of $\beta$ remain unidentified.

To build intuition, consider the following illustrative special case. Suppose $k = \ell = 2$ and that only one regressor, $X_2$, is endogenous. Assume that $Z_1 = X_1$ (i.e., $X_1$ is its own instrument), and that $Z_2$ is a valid external instrument for $X_2$. In this case, the projection matrix $\Pi'$ takes the form:

$$\Pi' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \pi_0 & \pi_1 & \pi_2 \end{pmatrix},$$

where the rows correspond to the constant and the two components of $X$, and the columns correspond to the components of $Z = (1, X_1, Z_2)$. The rank of $\Pi'$ must be $k+1 = 3$ for identification. This implies in particular that $\pi_2 \neq 0$; that is, the instrument $Z_2$ must contribute nontrivially to the prediction of $X_2$ even after controlling for the constant and for $X_1$.

In econometric parlance, we say that $Z_2$ is a **relevant instrument** for $X_2$: it shifts $X_2$ in the population in a way that is not redundant with the variation explained by $X_1$. Relevance is not simply a technical condition—without it, we would be trying to explain variation in $Y$ using components of $X$ that we cannot isolate using exogenous variation.

This example generalizes naturally to models with multiple endogenous regressors and multiple instruments. In all such cases, two key requirements must be satisfied for identification via instrumental variables:

(i) **Instrument exogeneity**: The instruments must be uncorrelated with the error term $U$, i.e., $E[ZU] = 0$.

(ii) **Instrument relevance**: The instruments must contain sufficient variation to predict the endogenous components of $X$, conditional on the exogenous components.

Together, these conditions ensure that we can isolate variation in $X$ that is both exogenous and informative—thus allowing us to identify the structural effect $\beta$.

All in all, the strength of Lemma 12.1 lies in its ability to turn the system of moment conditions

$$E[ZY] = E[ZX']\beta$$

into an explicit solution for $\beta$. By pre-multiplying both sides by the matrix $\Pi'$, we obtain:

$$\Pi' E[ZY] = \Pi' E[ZX']\beta \ .$$

Then, assuming the invertibility of $\Pi' E[ZX']$—which is guaranteed by the rank condition—we can solve for $\beta$:

$$\beta = (\Pi' E[ZX'])^{-1} \Pi' E[ZY] \ . \tag{12.6}$$

This expression provides the foundation for identification in instrumental variables models, even in overidentified systems.

## 12.3 Some examples of IVs in practice

Instrumental variables are widely used in applied work to address endogeneity concerns, and over time a variety of instruments have been proposed across different fields. Table 12.1 collects a few well-known examples. Here we briefly discuss three of them to illustrate how the key IV conditions—relevance and exogeneity—are evaluated in context.

In their study on the effect of fertility on female labor supply, Angrist and Evans [1998] use the occurrence of twins at the second birth as an instrument (though this is not the instrument they recommend in the end). The treatment variable $A$ (the endogenous $X_1$ in our notation) is an indicator for having more than two children, and the outcome $Y$ is a labor market measure for the mother. The idea is that having twins increases fertility in a way that is plausibly exogenous to labor supply decisions. While the instrument appears relevant—twins mechanically increase the likelihood of having more than two

children—its exogeneity has been questioned. For instance, twinning rates vary with maternal age, and twins affect not only the number but also the spacing of children, which may have independent effects on labor outcomes.

A cleaner example comes from Angrist [1990], who studies the causal effect of military service on later-life outcomes using the Vietnam draft lottery as an instrument. The treatment $A$ is veteran status, and the instrument $Z$ is a dummy for receiving a low lottery number. The instrument is relevant: individuals with low numbers were more likely to serve. In addition, since lottery numbers were randomly assigned, it is tempting to conclude that $Z$ is exogenous. However, exogeneity is a stronger condition than randomness: it requires that $Z$ be independent of unobserved determinants of the outcome. In this case, there would be concerns if individuals reacted to their draft status—by enrolling in school, moving abroad, or otherwise altering life decisions—which could create correlation between the instrument and $U$.

A third example is from Sarsons [2015], who studies the relationship between income shocks and religious conflict in India. Rainfall shocks are used as an instrument for income, leveraging the fact that agricultural income is sensitive to rain. Exogeneity is often assumed on the basis that rainfall is naturally random, but Sarsons shows that this assumption may not hold. In particular, she finds that even in districts with irrigation dams—where income is largely insulated from rainfall—rainfall still predicts conflict. This suggests that rainfall may affect conflict through non-income channels, violating the exclusion restriction.

These examples highlight the central challenge of IV analysis: while relevance is often straightforward to test, exogeneity and exclusion must be argued carefully based on institutional knowledge and empirical evidence.

## 12.4 Key Concepts

- **Endogeneity** occurs when a regressor is correlated with the error term, leading to inconsistent OLS estimates.
- Common sources include **omitted variables**, **measurement error**, and **simultaneity**.
- **Instrumental Variables (IV)** resolve endogeneity by using instruments $Z$ that satisfy:
  - **Relevance**: $Z$ must predict $X$ .
  - **Exogeneity**: $Z$ must be uncorrelated with $U$.

TABLE 12.1: Examples of instruments used in practice

| Outcome Variable | Endogenous Variable | Instrumental Variable(s) | Reference |
|---|---|---|---|
| Labor supply | Disability insurance replacement rates | Region and time variation in benefit rules | Gruber (2000) |
| Labor supply | Fertility | Sibling-Sex composition | Angrist and Evans (1998) |
| Education, Labor supply | Out-of-wedlock fertility | Occurrence of twin births | Bronars and Grogger (1994) |
| Wages | Unemployment insurance tax rate | State laws | Anderson and Meyer (2000) |
| Earnings | Years of schooling | Region and time variation in school construction | Duflo (2001) |
| Earnings | Years of schooling | Proximity to college | Card (1995) |
| Earnings | Years of schooling | Quarter of birth | Angrist and Krueger (1991) |
| Earnings | Veteran status | Cohort dummies | Imbens and van der Klaauw (1995) |
| Earnings | Veteran status | Draft lottery number | Angrist (1990) |
| Achievement test scores | Class size | Discontinuities in class size due to maximum class-size rule | Angrist and Lavy (1999) |
| College enrollment | Financial aid | Discontinuities in financial aid formula | van der Klaauw (1996) |
| Health | Heart attack surgery | Proximity to cardiac care centers | McClellan, McNeil and Newhouse (1994) |
| Crime | Police | Electoral cycles | Levitt (1997) |
| Employment and Earnings | Length of prison sentence | Randomly assigned federal judges | Kling (1999) |
| Birth weight | Maternal smoking | State cigarette taxes | Evans and Ringel (1999) |

- **Just-identified case**: when $\ell = k$, $\beta$ is identified by the equation

$$\beta = (E[ZX'])^{-1}E[ZY] \ .$$

- **Over-identified case**: when $\ell > k$, $\beta$ is identified by the equation

$$\beta = (\Pi' E[ZX'])^{-1}\Pi' E[ZY] \ .$$

## 12.5 Concluding Remarks

The material today borrows from notes by Azeem Shaikh that he kindly shared with me. In general, the topic of linear IV and endogeneity is covered in most standard sources, including Hansen [2022] and Wooldridge [2010].

## 12.6 Problems

**Problem 12.1** *Suppose* $Y = \beta_0 + \beta_1 X_1 + U$*, where* $E[U] = 0$ *and* $E[Z_1 U] = 0$*. Show that under these conditions,*

$$\beta_1 = \frac{\mathrm{Cov}(Z_1, Y)}{\mathrm{Cov}(Z_1, X_1)} \ .$$

Hint: *Instead of relying on matrix algebra, note that* $E[UZ_1] = 0$ *is the same as* $\mathrm{Cov}(U, Z_1) = 0$*. Use this last condition in combination with the identity* $U = Y - \beta_0 - \beta_1 X_1$ *to obtain the result - recalling that the covariance between a random variable and a constant is always zero.*

**Problem 12.2** *Consider the measurement error setting in Section 12.1.2. Assume that* $\beta_1 > 0$ *and show that* $\frac{\mathrm{Cov}[\hat{X},Y]}{\mathrm{Var}[\hat{X}]} \leq \beta_1 \leq \frac{\mathrm{Var}[Y]}{\mathrm{Cov}[\hat{X},Y]}$*. Interpret the upper and lower bounds in terms of coefficients from a regression.*

**Problem 12.3** *Consider the linear model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \ ,$$

*where* $E[U] = E[X_1 U] = E[X_2 U] = 0$*. Suppose the researcher omits* $X_2$ *and runs a regression of* $Y$ *on* $(1, X_1)$*. Let* $\hat{\beta}^*_{1,n}$ *denote the resulting LS estimator of the slope.*

*(a) Write the asymptotic bias of* $\hat{\beta}^*_{1,n}$ *as a product of two terms, each with a clear interpretation.*

*(b) Determine the sign of the asymptotic bias in each of the four cases given by* $\beta_2 > 0$ *or* $\beta_2 < 0$*, and* $\mathrm{Cov}(X_1, X_2) > 0$ *or* $\mathrm{Cov}(X_1, X_2) < 0$*.*

*(c) Apply your reasoning to the canonical example in which* $Y$ *is wages,* $X_1$ *is years of schooling, and* $X_2$ *is unobserved ability. State what sign you expect for the asymptotic bias of LS and explain whether LS over- or under-estimates the causal return to schooling.*

*(d) Now suppose the omitted variable is labor-market network strength, such as having a parent in a high-paying occupation, and that it is positively correlated with both wages and schooling. Does this change your answer to part (c) qualitatively? Explain.*

**Problem 12.4** *State whether each of the following claims is true, false, or partially true. Justify briefly. A counterexample is sufficient to disprove a claim.*

*(a) If the rank condition fails, then the order condition fails.*

*(b) Suppose* $X = (1, X_1, X_2)'$ *with both* $X_1$ *and* $X_2$ *endogenous, and* $Z = (1, Z_1, Z_2)'$ *with* $\text{Cov}(Z_1, X_1) \neq 0$ *and* $\text{Cov}(Z_2, X_2) \neq 0$. *Then the rank condition is automatically satisfied.*

*(c) If the rank condition fails for* $X$ *on* $Z$, *the IV estimand* $(E[ZX'])^{-1}E[ZY]$ *does not exist.*

*For any claim you mark partially true, identify a sufficient condition under which the claim becomes true.*

## Bibliography

J. D. Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The american economic review*, pages 313–336, 1990.

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

H. Sarsons. Rainfall and conflict: A cautionary tale. *Journal of development Economics*, 115:62–72, 2015.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

# 13

# *Estimation under Endogeneity: IV and TSLS*

In the previous chapter, we introduced the problem of endogeneity and developed the key identifying conditions for instrumental variables (IV): exogeneity and relevance. We characterized the population solution for the causal parameter $\beta$ under both just- and over-identified systems using moment conditions implied by valid instruments.

In this chapter, we move from identification to estimation. We begin with the simplest case where the number of instruments equals the number of endogenous regressors and derive the IV estimator from first principles. We then extend this framework to the over-identified case and introduce the Two-Stage Least Squares (TSLS) estimator. Our goal is to build intuition for how these estimators work, where they come from, and how they relate to the identification results discussed previously. We postpone formal analysis of their statistical properties to the next chapter.

## 13.1 Where Do Instrumental Variables Come From?

Before deriving the IV estimator, it is useful to begin with a setting in which the logic of instrumental variables is especially transparent: a randomized experiment with imperfect compliance. Suppose we want to estimate the effect of a treatment on an outcome. Let $A = 1$ indicate that an individual actually receives the treatment, and let $Y$ denote the outcome of interest. If treatment receipt were randomly assigned, then comparing average outcomes between those with $A = 1$ and those with $A = 0$ would be a natural way to estimate the causal effect of the treatment.

In many experiments, however, the researcher does not directly control treatment receipt. Instead, the researcher controls assignment to treatment. Let $Z_1 = 1$ denote assignment to the treatment group and $Z_1 = 0$ denote assignment to the control group. If everyone complied perfectly, then $A = Z_1$, and assignment would coincide with treatment receipt. In that case, the difference in average outcomes by assignment status,

$$E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0] \ ,$$

would also be the difference in average outcomes by treatment status.

With imperfect compliance, this equality breaks down. Some individuals assigned to treatment may not take up the treatment, and some individuals assigned to control may obtain the treatment through other channels. In that case, $Z_1$ is not the same as $A$. Nevertheless, $Z_1$ can still be useful. Random assignment makes $Z_1$ plausibly unrelated to the unobserved determinants of the outcome, while the fact that assignment changes the probability of treatment receipt makes $Z_1$ informative about $A$. In the language of the previous chapter, $Z_1$ is a candidate instrument for $A$: it is plausibly exogenous and relevant.

A concrete example is the Oregon Medicaid lottery. In 2008, Oregon used a lottery to allocate a limited number of opportunities to apply for Medicaid coverage. Winning the lottery did not itself guarantee Medicaid coverage: some lottery winners did not enroll, and some non-winners obtained coverage through other channels. Thus lottery selection is not the same as Medicaid enrollment. But lottery selection is randomly assigned and strongly affects the probability of enrolling in Medicaid. In our notation, lottery selection plays the role of $Z_1$, while actual Medicaid enrollment plays the role of $A$.

This example already contains the main idea behind IV. If assignment affects the outcome only through its effect on treatment receipt, then the effect of assignment on the outcome should be proportional to the effect of assignment on treatment receipt. Formally, in a simple homogeneous-effects model,

$$Y = \beta_0 + \beta_1 A + U \ ,$$

where $\beta_1$ is the causal effect of treatment, random assignment implies that $Z_1$ is uncorrelated with $U$. Taking conditional means by assignment status gives

$$E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0] = \beta_1 \{E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0]\} \ .$$

Therefore,

$$\beta_1 = \frac{E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0]}{E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0]} \ .$$

The numerator is the effect of assignment on the outcome, often called the intent-to-treat effect. The denominator is the effect of assignment on treatment receipt, or the first stage. The IV estimand rescales the assignment effect by the amount of compliance generated by the assignment.

This discussion answers, in a simple setting, where instrumental variables come from. An instrument is a source of variation that shifts the endogenous variable of interest while being otherwise unrelated to the unobserved determinants of the outcome. Random assignment with imperfect compliance provides the cleanest example: assignment is not the treatment, but it generates exogenous variation in treatment receipt. Other instruments often come from natural experiments or institutional rules, such as the highway plan used in the empirical illustration below. The rest of this chapter shows how this same idea is expressed algebraically through moment conditions and how it leads to the IV and TSLS estimators.

## 13.2 The Instrumental Variables (IV) Estimator

Let $(Y, X, Z, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$, $X$ takes values in $\mathbf{R}^{k+1}$, and $Z$ takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in $Z$, and that the rank of $E[ZX']$ is $k+1$.

We now discuss estimation of $\beta$. In order to do so, we assume that we have access to a random sample of size $n$ from the distribution of $(Y, X, Z)$, denoted by $P$. Concretely, we let

$$(Y_i, X_i, Z_i)_{i=1}^n$$

be an i.i.d. sequence of random variables with distribution $P$.

We first consider the case in which $k = \ell$. By analogy with the expression we derived for $\beta$ in (12.4), that is,

$$\beta = (E[ZX'])^{-1}E[ZY] \ ,$$

under these assumptions, the natural estimator of $\beta$ is given by

$$\hat{\beta}_{n,\text{iv}} = \left(\frac{1}{n}\sum_{1\le i\le n} Z_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\le i\le n} Z_i Y_i\right) \ . \tag{13.1}$$

This estimator is called the *instrumental variables* (IV) estimator of $\beta$. Note that $\hat{\beta}_{n,\text{iv}}$ satisfies

$$\frac{1}{n}\sum_{1\le i\le n} Z_i(Y_i - X_i'\hat{\beta}_{n,\text{iv}}) = 0 \ .$$

In particular, $\hat{U}_i = Y_i - X_i'\hat{\beta}_{n,\text{iv}}$ satisfies

$$\frac{1}{n}\sum_{1\le i\le n} Z_i\hat{U}_i = 0 \ ,$$

which means that the instruments are uncorrelated with the residuals from the regression in each finite sample of size $n$ by construction.

Similarly to our discussion in the previous chapter, an especially tractable and instructive case arises when $\ell = k = 1$, so that $X = (1, X_1)'$, $Z = (1, Z_1)'$, and the model is:

$$Y = \beta_0 + \beta_1 X_1 + U$$

with the assumption $E[U] = E[Z_1 U] = 0$. In this case we learned that the expression for $\beta_1$ simplifies to:

$$\beta_1 = \frac{\text{Cov}(Z_1, Y)}{\text{Cov}(Z_1, X_1)} \ . \tag{13.2}$$

It follows that a similar simplification applies to the IV estimator of $\beta_1$, as in this case

$$\hat{\beta}_{1,\text{iv}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})Y_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})X_{1,i}} \ , \tag{13.3}$$

where $\bar{Z}_{1,n} := \frac{1}{n}\sum_{i=1}^{n} Z_{1,i}$ is the sample average of the instrument $Z_1$. In this case, a useful interpretation arises if we divide both the numerator and the denominator by $\frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})^2$, i.e.,

$$\hat{\beta}_{1,\text{iv}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})Y_i / \frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})^2}{\frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})X_{1,i} / \frac{1}{n}\sum_{i=1}^{n}(Z_{1,i} - \bar{Z}_{1,n})^2} \ . \tag{13.4}$$

This expression reveals that the IV estimator of the slope coefficient $\beta_1$ is simply the ratio of the regression slope of $Y$ on $Z_1$ (the so-called *reduced form*) to the regression slope of $X_1$ on $Z_1$ (the so-called *first stage*).

To see this in a different way, write the model as

$$Y = \beta_0 + \beta_1 X_1 + U$$

and

$$X_1 = \pi_0 + \pi_1 Z_1 + V \ .$$

It is important to note that this second equation for $X_1$ is *not* an assumption but rather a mechanical projection of $X_1$ on $Z_1$. In other words, we interpret the coefficients $(\pi_0, \pi_1)$ according to the second interpretation of least squares (the BLP interpretation) and, by construction, $E[ZV] = 0$; in particular, $E[V] = 0$. In contrast, the equation for $Y$ is viewed as a homogeneous causal model and so $(\beta_0, \beta_1)$ admit the third interpretation of least squares we have previously discussed (the causal interpretation).

Replacing the second equation into the first one delivers

$$Y = \beta_0^* + \beta_1^* Z_1 + U^*$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_1\pi_0 \\ \beta_1^* &= \beta_1\pi_1 \\ U^* &= U + \beta_1 V \ . \end{aligned}$$

It follows that $E[ZU^*] = 0$ and so a regression of $Y$ on $(1, Z_1)$ would identify the slope coefficient $\beta_1^* = \beta_1\pi_1$. In addition, the estimated slope from a regression of $X_1$ on $(1, Z_1)$ converges to $\pi_1$. We conclude:

$$\frac{\text{slope of the reduced form}}{\text{slope of the first stage}} = \frac{\beta_1^*}{\pi_1} = \frac{\beta_1\pi_1}{\pi_1} = \beta_1 \ . \tag{13.5}$$

The IV estimator is nothing other than the sample analog of this insight, therefore delivering a consistent estimator of $\beta_1$ (a property we prove in the next chapter). Note that the IV estimand is predicated on the notion that the first stage slope is not zero ($\pi_1 \neq 0$), which is just another way to state our rank/relevance condition in this simple case.

### 13.2.1 Empirical Illustration

Baum-Snow [2007] studies a classic question in urban economics: What explains the growth of suburbs in U.S. cities? One explanation, grounded in land use theory, is that improvements in transportation—specifically, faster commuting times—make suburban living more attractive. Other factors include shifting amenities, racial preferences, school desegregation, and crime in central cities.

To isolate the causal effect of transportation infrastructure, Baum-Snow examines the impact of highway construction between 1950 and 1990. He finds that roughly one-third of the suburbanization observed in this period can be attributed to the expansion of the highway system. Estimating this effect raises a natural endogeneity concern: are highways built because cities grow, or do cities grow because highways are built? For example, economically successful cities may build more highways to accommodate growth, or rising crime may spur suburban demand and justify road expansion. In such cases, observed correlations between highways and suburbanization confound cause and effect.

TABLE 13.1: First Stage Results

| | *Dependent variable:* |
|---|---|
| | Change in N. Highways, 1950 - 1990 |
| Planned highways in 1950 | 0.510*** |
| | (0.074) |
| 1950 Center City Radius | 0.306*** |
| | (0.072) |
| Change in (log) income | −0.939 |
| | (1.819) |
| Change in Metropolitan Area Population | 0.856*** |
| | (0.279) |
| Constant | 0.463 |
| | (1.231) |
| Observations | 139 |
| $R^2$ | 0.503 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

The identification strategy in Baum-Snow [2007] exploits the 1947 Federal Highway Plan, which laid out a national system of interstate highways to connect major cities. Crucially, local cities had little say in the planned routes. As a result, cities that were "assigned" more connections in the 1947 plan were effectively required to build more highway infrastructure, even if they had not chosen to do so on their own. This provides a source of exogenous variation

in highway construction. The author then runs the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \sum_{j=2}^{k} \beta_j X_{j,i} + U_i \ ,$$

where $Y_i$ is the change in central-city population in city $i$ between 1950 and 1990, $X_{1,i}$ is the change in the number of highways over the same period, and $X_{j,i}$ are control variables such as changes in income, initial population, and other demographic trends. Because cities may build highways in response to expected growth, the regressor $X_{1,i}$ is believed to be endogenous, and so $E[X_{1,i}U_i] \neq 0$.

To address this concern, the author uses as an instrument $Z_{1,i}$: the number of interstate highways assigned to city $i$ in the 1947 plan. The first-stage regression is given by:

$$X_{1,i} = \pi_0 + \pi_1 Z_{1,i} + \sum_{j=2}^{k} \pi_j X_{j,i} + V_i \ .$$

Table 13.1 shows the first-stage results. The instrument appears to be *relevant*: conditional on controls, $Z_{1,i}$ is strongly positively correlated with $X_{1,i}$. Table 13.2 reports OLS and IV results. The LS coefficient on $X_1$ is smaller in magnitude than the IV coefficient, suggesting that LS attenuates the negative suburbanization effect of highway construction toward zero. The IV estimate is larger in magnitude, and plausibly causal: more highways caused a strong decrease in urban center population. The full `R` code used to produce the tables in this section is available on Canvas.

TABLE 13.2: Regression Results

| | *Dependent variable:* | |
|---|---|---|
| | Change pop. in the urban center of the city, 1950-1990 | |
| | *OLS* | *Instrumental Variable (IV)* |
| | (1) | (2) |
| Change in N. Highways, 1950 - 1990 | $-0.059^{***}$ | $-0.123^{***}$ |
| | (0.014) | (0.029) |
| 1950 Center City Radius | $0.080^{***}$ | $0.113^{***}$ |
| | (0.014) | (0.020) |
| Change in (log) income | 0.084 | 0.048 |
| | (0.335) | (0.362) |
| Change in Metropolitan Area Population | $0.363^{***}$ | $0.424^{***}$ |
| | (0.053) | (0.062) |
| Observations | 139 | 139 |
| $R^2$ | 0.395 | 0.296 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

```
# Load data
```

```r
dw <- read_dta("baumsnow.dta")

# Select preferred sample (zones considered urban centers)
dw_filtered <- dw %>% filter(pop50 > .1 & cpop50 > .05)

# Run the first stage model
first_stage_model <- lm(draccld ~ rays_planc + rcarea + Dlincomeh
     + dlpop, data = dw_filtered)

# Run OLS and IV
ols_model <- lm(dlcpop2 ~ draccld + rcarea + Dlincomeh + dlpop,
    data = dw_filtered)

iv_model1 <- ivreg(dlcpop2 ~ draccld + rcarea + Dlincomeh + dlpop
     | rays_planc + rcarea + Dlincomeh + dlpop, data = dw_filtered
    )
```

Code Snippet 13.1: First Stage, OLS, and IV

## 13.3 The Two-Stage Least Squares (TSLS) Estimator

Now consider the case in which $\ell > k$. The expressions derived for $\beta$ in this case involved $\Pi$, where $\text{BLP}(X|Z) = \Pi' Z$. That is,

$$\beta = (\Pi' E[ZX'])^{-1} \Pi' E[ZY] \ .$$

An estimate of $\Pi$ can be obtained by OLS. More precisely, since $\Pi = E[ZZ']^{-1} E[ZX']$, a natural estimator of $\Pi$ is

$$\hat{\Pi}_n = \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right) \ .$$

This is mechanically the same as running a separate regression of each component in $X$ onto $Z$.

With this estimator of $\Pi$, a natural estimator of $\beta$ is simply

$$\hat{\beta}_{n,\text{tsls}} = \left( \frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i X_i' \right)^{-1} \left( \frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Y_i \right) \ . \tag{13.6}$$

This estimator of $\beta$ is called the *two-stage least squares* (TSLS) estimator of $\beta$. Looking closely at this equation, we see that it provides an interpretation of the TSLS estimator as an IV estimator with $\hat{\Pi}_n' Z_i$ playing the role of the instrument. To see this clearly, let $\hat{X}_i = \hat{\Pi}_n' Z_i$ and note that we can write the

TSLS estimator as:

$$\hat{\beta}_{n,\text{tsls}} = \left(\frac{1}{n}\sum_{1\le i\le n} \hat{X}_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{1\le i\le n} \hat{X}_i Y_i\right) .$$

which is simply the formula in (13.1) with $\hat{X}_i$ playing the role of the instrument. In other words, when there are more instruments than endogenous variables, the TSLS estimator takes a linear combination of the instruments. The fitted value $\hat{X}_i = \hat{\Pi}_n' Z_i$ is the OLS projection of $X_i$ on $Z_i$, so it is the linear combination of the available instruments that best predicts $X_i$ in the least-squares sense.

It is termed the *two-stage least squares* estimator because it may be obtained by running two separate LS regressions:

- **First regression**: regress (each component of) $X_i$ on $Z_i$ to obtain $\hat{X}_i = \hat{\Pi}_n' Z_i$;
- **Second regression**: regress $Y_i$ on $\hat{X}_i$ to obtain $\hat{\beta}_{n,\text{tsls}}$.

These two interpretations are algebraically equivalent because the first-stage residuals are orthogonal to the fitted values $\hat{X}_i$. Indeed, it follows from Problem 13.2 that we can alternatively write $\hat{\beta}_{n,\text{tsls}}$ as

$$\begin{aligned}\hat{\beta}_{n,\text{tsls}} &= \left(\frac{1}{n}\sum_{1\le i\le n} \hat{\Pi}_n' Z_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{1\le i\le n} \hat{\Pi}_n' Z_i Y_i\right) \\ &= \left(\frac{1}{n}\sum_{1\le i\le n} \hat{X}_i \hat{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{1\le i\le n} \hat{X}_i Y_i\right) ,\end{aligned}$$

which shows that $\hat{\beta}_{n,\text{tsls}}$ is a LS estimator from the regression of $Y_i$ on $\hat{X}_i$.

## 13.4 IV for Binary Endogenous Variables

Thus far, we have derived the IV and TSLS estimators in a general linear setting. We now return to the randomized experiment with imperfect compliance discussed at the beginning of the chapter and show how the IV formula simplifies when both the endogenous variable and the instrument are binary. In the opening section we derived the same formula directly from the homogeneous-effects model; here we re-derive it from the general IV estimand, confirming that the two routes deliver the same expression.

Let $X = (1, A)'$, where $A \in \{0, 1\}$ denotes treatment receipt, and let $Z =$

$(1, Z_1)'$, where $Z_1 \in \{0, 1\}$ denotes assignment or another binary instrument. The model takes the form

$$Y = \beta_0 + \beta_1 A + U \ ,$$

where $A$ is endogenous, meaning $E[AU] \neq 0$. Unlike many LS applications, this specification is always intended to capture a *causal* effect rather than simply describing associations.

As emphasized earlier, this model imposes a *homogeneous treatment effect*: every individual responds to treatment in exactly the same way, and this common effect is represented by $\beta_1$. We assume homogeneity here for analytical convenience and discuss its limitations in the later chapter on heterogeneous and endogenous treatments.

In this section, we first work with population quantities; the corresponding sample estimators are obtained by replacing population means with sample averages. It follows immediately from (13.2) that the slope coefficient $\beta_1$ is given by

$$\beta_1 = \frac{\mathrm{Cov}(Y, Z_1)}{\mathrm{Cov}(A, Z_1)} \ . \tag{13.7}$$

However, because both $A$ and $Z_1$ are binary random variables, this formula can be simplified further. Dividing both the numerator and denominator by $\mathrm{Var}[Z_1]$, we observe that: (a) the numerator $\mathrm{Cov}(Y, Z_1)/\mathrm{Var}[Z_1]$ is the least squares (LS) slope from regressing $Y$ on $(1, Z_1)$, and (b) the denominator $\mathrm{Cov}(A, Z_1)/\mathrm{Var}[Z_1]$ is the LS slope from regressing $A$ on $(1, Z_1)$.

In Chapter 5 (equation (5.1)), we showed that regressing a variable on a binary regressor and a constant yields the so-called *mean contrast*. Therefore,

$$\frac{\mathrm{Cov}(Y, Z_1)}{\mathrm{Var}[Z_1]} = E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0] \ , \tag{13.8}$$

$$\frac{\mathrm{Cov}(A, Z_1)}{\mathrm{Var}[Z_1]} = E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0] \ . \tag{13.9}$$

Hint 13.1 provides an alternative derivation of this result, applied specifically to the denominator.

Combining (13.7), (13.8), and (13.9), we obtain:

$$\beta_1 = \frac{\mathrm{Cov}(Y, Z_1)/\mathrm{Var}[Z_1]}{\mathrm{Cov}(A, Z_1)/\mathrm{Var}[Z_1]} = \frac{E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0]}{E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0]} \ . \tag{13.10}$$

The right-hand side of (13.10) is known as the *Wald estimand* [Wald, 1940]. Thus, when both $A$ and $Z_1$ are binary, the IV estimand and the Wald estimand coincide. In the imperfect-compliance example, the numerator is the effect of assignment on the outcome, and the denominator is the effect of assignment on treatment receipt. The Wald estimand scales the assignment effect by the compliance difference. Naturally, the sample analogs of these quantities (i.e., the estimators of $\beta_1$) also coincide.

This formula also clarifies what *instrument relevance* means in this context: we require that the denominator be nonzero. That is, $E[A \mid Z_1 = 1] \neq E[A \mid Z_1 = 0]$, which, using Hint 5.1, can be rewritten as

$$P\{A = 1 \mid Z_1 = 1\} \neq P\{A = 1 \mid Z_1 = 0\} \ .$$

In other words, the instrument must affect the probability of receiving treatment; the probability that an individual chooses $A = 1$ must differ across values of $Z_1$.

**Hint 13.1** *Consider a regression of $A$ on $(1, Z_1)$, where $Z_1$ is binary with $p_Z := P\{Z_1 = 1\}$. The population LS slope satisfies*

$$\begin{aligned}
\frac{\text{Cov}(A, Z_1)}{\text{Var}[Z_1]} &= \frac{E[A(Z_1 - p_Z)]}{p_Z(1-p_Z)} = \frac{E\left[E[A(Z_1 - p_Z) \mid Z_1]\right]}{p_Z(1-p_Z)} \\
&= \frac{E\left[Z_1 E[A|Z_1]\right] - p_Z E\left[E[A|Z_1]\right]}{p_Z(1-p_Z)} \\
&= \frac{E[A|Z_1 = 1]p_Z - p_Z\left[E[A|Z_1 = 1]p_Z + E[A|Z_1 = 0](1-p_Z)\right]}{p_Z(1-p_Z)} \\
&= \frac{E[A|Z_1 = 1]p_Z(1-p_Z) - p_Z(1-p_Z)E[A|Z_1 = 0]}{p_Z(1-p_Z)} \\
&= E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0] \ ,
\end{aligned}$$

*where the second equality uses the LIE, the third pulls $Z_1$-measurable terms outside the inner expectation, and the remaining steps use Hint 5.1 to evaluate expectations at the two possible values of $Z_1$.*

## 13.5 An Application: Angrist and Evans

Angrist and Evans [1998] is one of the well-known papers of modern labor economics that represents a good example of endogenous decisions and valid instruments. The paper looks at the effect of fertility decisions (having an extra child) on female labor force participation and earnings. A first instinct would be to regress labor force status on whether a woman has children, or on the number of children that she has. The correlation between the number of children and labor force participation tends to be strongly negative. But the decision to have children is plausibly correlated with expected income. Indeed, these are joint *choices* that lead to many possible endogeneity stories. Here's just one: high earning women may have fewer children due to higher opportunity cost.

To map the empirical strategy in Angrist and Evans to our notation, consider the following definitions. First, we denote by $Y$ the outcome of interest, which will be a measure of labor market participation or outcome for

TABLE 4—DIFFERENCES IN MEANS FOR DEMOGRAPHIC VARIABLES BY *SAME SEX* AND *TWINS*-2

| | Difference in means (standard error) | | |
|---|---|---|---|
| | By *Same sex* | | By *Twins-2* |
| Variable | 1980 PUMS | 1990 PUMS | 1980 PUMS |
| *Age* | −0.0147 | 0.0174 | 0.2505 |
| | (0.0112) | (0.0112) | (0.0607) |
| *Age at first birth* | 0.0162 | −0.0074 | 0.2233 |
| | (0.0094) | (0.0114) | (0.0510) |
| *Black* | 0.0003 | 0.0021 | 0.0300 |
| | (0.0010) | (0.0011) | (0.0056) |
| *White* | 0.0003 | −0.0006 | −0.0210 |
| | (0.0012) | (0.0013) | (0.0066) |
| *Other race* | −0.0006 | −0.0014 | −0.0090 |
| | (0.0005) | (0.0009) | (0.0041) |
| *Hispanic* | −0.0014 | −0.0007 | −0.0069 |
| | (0.0009) | (0.0010) | (0.0047) |
| *Years of education* | −0.0028 | 0.0100 | 0.0940 |
| | (0.0076) | (0.0074) | (0.0415) |

*Notes:* The samples are the same as in Table 2. Standard errors are reported in parentheses.

TABLE 13.3: Differences in means for demographic variables by same-sex and twins instrument.

the woman (or her husband). Angrist and Evans restrict the sample to only women (or couples) with 2 or more children, and so $A$ is an indicator for having more than 2 children (versus exactly 2). That is,

$$A := I\{\text{number of children} > 2\} \ .$$

The authors then consider two instruments for $A$. The first one is $Z_1 = 1$ if the first two children had the *same sex*. This is based on the idea that there is *preference to have a mix of boys and girls*, so that

$$P\{A = 1 \mid Z_1 = 1\} > P\{A = 1 \mid Z_1 = 0\} \ .$$

The instrument is also expected to be exogenous, as sex is "randomly" determined. The authors also consider another instrument, where $Z_1 = 1$ if the second birth was a twin. The twins instrument is used primarily for comparison as it had been used in previous papers. While one could tell a history of how the probability of $A = 1$ may vary with such an instrument, exogeneity of

the instrument is more difficult to defend as it is well-known that older women are more likely to have twins, on top of the fact that such an instrument affects both the number and spacing of children. Table 13.3 reproduces Table 4 in Angrist and Evans [1998] and shows that while same-sex is uncorrelated with a variety of observed confounders, the twins instrument is well-known to be correlated with age (so, education) and race.

Angrist and Evans compute the Wald estimators for a variety of outcomes, including labor income (which we denote by $Y$) below. If we let $N_1 = \{i \in \mathbf{N} : Z_{1,i} = 1\}$ and $N_0 = \{i \in \mathbf{N} : Z_{1,i} = 0\}$, then the numerator of the Wald estimator is given by

$$\frac{1}{|N_1|}\sum_{i=1}^{n} Y_i Z_{1,i} - \frac{1}{|N_0|}\sum_{i=1}^{n} Y_i(1 - Z_{1,i}) = -132.5 \quad (34.4) \ ,$$

with standard errors between parenthesis, while the denominator is

$$\frac{1}{|N_1|}\sum_{i=1}^{n} A_i Z_{1,i} - \frac{1}{|N_0|}\sum_{i=1}^{n} A_i(1 - Z_{1,i}) = 0.060 \quad (0.0016) \ .$$

The ratio of the two leads to

$$\hat{\beta}_1 = -2208.8(569.2) \ .$$

These numbers can be found in Table 5 of Angrist and Evans [1998]. Under homogeneous effects, this estimate is interpreted as the constant marginal effect of having an additional child on annual earnings. Under heterogeneous effects, the same calculation has a different interpretation, which we discuss in the later chapter on heterogeneous and endogenous treatments.

## 13.6 Key Concepts

- A randomized experiment with imperfect compliance provides a simple motivation for IV: assignment $Z$ is not the same as treatment receipt $A$, but it can generate exogenous variation in $A$. In the binary case, the IV estimand rescales the effect of assignment on the outcome by the effect of assignment on treatment receipt.
- The **instrumental variables (IV) estimator** solves sample analogs of the moment conditions $E[Z(Y - X'\beta)] = 0$. When the number of instruments equals the number of regressors ($\ell = k$), the IV estimator has a simple closed form.
- In the **over-identified case** ($\ell > k$), the **Two-Stage Least**

**Squares (TSLS)** estimator is used. TSLS proceeds by: (i) projecting $X$ onto $Z$ (first stage), and (ii) regressing $Y$ on the predicted values from the first stage (second stage).

- When both the endogenous regressor $A$ and the scalar instrument $Z_1$ are binary, the IV estimand reduces to the **Wald estimand**, given by the ratio of mean differences:

$$\beta_1 = \frac{E[Y \mid Z_1 = 1] - E[Y \mid Z_1 = 0]}{E[A \mid Z_1 = 1] - E[A \mid Z_1 = 0]} \ .$$

## 13.7 Concluding Remarks

The material today borrows from notes by Azeem Shaikh that he kindly shared with me. In general, the topic of linear IV and endogeneity is covered in most standard sources, including Hansen [2022] and Wooldridge [2010].

## 13.8 Problems

**Problem 13.1** *Show that if $k = \ell$ and $\hat{\Pi}_n$ is invertible, then the TSLS estimator of $\beta$ is exactly equal to the IV estimator of $\beta$.*

**Problem 13.2** *Consider the case with $\ell > k$. Recall that $X_i = \Pi' Z_i + V_i$ with $E[Z_i V_i'] = 0$. By the properties of least squares, we know that the residual vector, $\hat{V}_i$, from the regression of $X_i$ on $Z_i$ is uncorrelated with $Z_i$. Use this result to show that*

$$\frac{1}{n} \sum_{1 \le i \le n} \hat{\Pi}_n' Z_i X_i' = \frac{1}{n} \sum_{1 \le i \le n} \hat{X}_i \hat{X}_i' \ . \tag{13.11}$$

**Problem 13.3** *Consider the model $Y = \beta X + U$ where $X \in \mathbf{R}$. Note the lack of an intercept term. Suppose $X$ is endogenous and that $Z$ is an instrument satisfying $E[ZU] = 0$ and $E[ZX] \neq 0$. You have a sample of size $n$ from the distribution of $(Y, X, Z)$. Show that the estimator*

$$\tilde{\beta}_n = \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i X_i} \xrightarrow{P} \beta \quad \text{as } n \to \infty \ .$$

**Problem 13.4** *Let $(Y, X, Z, U, V)$ be a random vector where all variables take values in* $\mathbf{R}$*, and consider the causal model*

$$Y = \beta^* X \ ,$$

*where $\beta^*$ is* random. *Further assume that $\beta^* = \beta + U$ where $E[U] = E[XU] = 0$ and $\beta$ is a constant to be estimated. The random variable $Z$ is viewed as an instrument satisfying $Z \perp\!\!\!\perp U$. Assume we observe an i.i.d. sample of size $n$ from the distribution of $(Y, X, Z)$.*

1. *Suppose that you run a regression of $Y$ on $X$ (without a constant), is the LS estimator consistent for $\beta$?*
2. *Suppose you compute the IV estimator of a regression of $Y$ on $X$ (without a constant) using $Z$ as an instrument, is the IV estimator consistent for $\beta$?*

**Problem 13.5** *Let $A \in \{0, 1\}$ denote a treatment, $Y$ an outcome, and $Z \in \{0, 1\}$ a binary candidate instrument. Suppose the homogeneous-effects model*

$$Y = \beta_0 + \beta_1 A + U$$

*holds, with $E[U] = 0$.*

*(a) Using the Wald estimand in (13.10), define the* intent-to-treat *effect as* $\text{ITT} = E[Y \mid Z = 1] - E[Y \mid Z = 0]$ *and the first-stage compliance effect as $\pi = E[A \mid Z = 1] - E[A \mid Z = 0]$. Express $\beta_1$ in terms of* ITT *and $\pi$, and interpret the formula in words.*

*(b) Suppose there is perfect compliance, so $A = Z$. Show that $\pi = 1$ and that $\beta_1$ reduces to the simple difference in means $E[Y \mid Z = 1] - E[Y \mid Z = 0]$.*

*(c) Suppose compliance is one-sided, so $P\{A = 1 \mid Z = 0\} = 0$. Express $\beta_1$ in terms of* ITT *and $\pi$. Why might this case be relevant in practice?*

*(d) Suppose instead that assignment has no effect on treatment receipt, so $E[A \mid Z = 1] = E[A \mid Z = 0]$. Explain why the Wald estimand is not defined and connect this failure to instrument relevance.*

**Problem 13.6** *Consider an experiment in which units are randomly assigned to one of two arms. Let $Z = 1$ denote assignment to treatment and $Z = 0$ denote assignment to control. Let $A = 1$ denote actual treatment receipt and $A = 0$ otherwise. Suppose:*

- $P\{Z = 1\} = 1/2$;
- $P\{A = 1 \mid Z = 1\} = 0.80$;
- $P\{A = 1 \mid Z = 0\} = 0.05$;

- *the homogeneous-effects model* $Y = \beta_0 + \beta_1 A + U$ *holds with* $\mathrm{Cov}(Z, U) = 0$.

*(a) Compute the population first-stage coefficient* $\pi$ *in the regression* $A = \alpha + \pi Z + V$.

*(b) Show that the asymptotic bias of the LS estimator of* $\beta_1$ *from a regression of* $Y$ *on* $(1, A)$ *equals* $\mathrm{Cov}(A, U)/\mathrm{Var}(A)$. *Compute* $\mathrm{Var}(A)$ *under the assumptions above and state the bias as a function of* $\mathrm{Cov}(A, U)$.

*(c) Explain why* $Z$ *is a more credible instrument than* $A$ *is a credible regressor. What assumption justifies* $\mathrm{Cov}(Z, U) = 0$*, and why is the analogous assumption* $\mathrm{Cov}(A, U) = 0$ *less credible?*

*(d) Suppose a colleague proposes to avoid IV by restricting the analysis to compliers, meaning units who would take treatment if assigned to treatment and would not take treatment if assigned to control. Explain why this is not feasible using the observed data. How does the IV/Wald estimand use assignment* $Z$ *to address this problem?*

## Bibliography

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

N. Baum-Snow. Did highways cause suburbanization? *The quarterly journal of economics*, 122(2):775–805, 2007.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

A. Wald. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

# 14

# *Properties of the TSLS Estimator*

Let $(Y, X, Z, U)$ be a random vector where $Y$ and $U$ take values in $\mathbf{R}$, $X$ takes values in $\mathbf{R}^{k+1}$, and $Z$ takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in $Z$, and that the rank of $E[ZX']$ is $k+1$. Denote by $P$ the marginal distribution of $(Y, X, Z)$ and let $\{(Y_i, X_i, Z_i) : 1 \le i \le n\}$ be an i.i.d. sample from $P$.

Section 13.3 described how to estimate $\beta$ by two-stage least squares (TSLS). Here we review the large-sample properties of the resulting estimator $\hat{\beta}_n$, namely consistency and asymptotic normality.

It is useful to keep three questions separate. First, *identification* asks whether the population assumptions determine the parameter $\beta$. Second, *consistency* asks whether the estimator converges to that population parameter as the sample size grows. Third, *inference* uses the large-sample distribution of the estimator to quantify sampling uncertainty through standard errors, hypothesis tests, and confidence intervals. The assumptions above play all three roles: the moment and rank conditions identify $\beta$, while the sampling and moment assumptions justify the large-sample approximations used for estimation and inference.

## 14.1 Consistency of the TSLS Estimator

We first prove that the TSLS estimator, $\hat{\beta}_n$, is consistent for $\beta$ under the assumptions we have previously stated. That is, $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$.

To prove this result, first write down $\hat{\beta}_n$ as in the previous class,

$$\hat{\beta}_n = \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \le i \le n} Z_i Y_i \right) \right) \ .$$

That is, $\hat{\beta}_n$ is a (continuous) function of two averages and the estimated

LS coefficients contained in $\hat{\Pi}_n$. We therefore need to (a) analyze each term individually using the LLNs, and (b) combine the results using the CMT.

First, recall from our results on OLS that

$$\hat{\Pi}_n \xrightarrow{P} \Pi$$

as $n \to \infty$. This result, in turn, followed from an application of the LLN and the CMT. Next, note that the LLN implies that

$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \xrightarrow{P} E[ZZ']$$
$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i Y_i \xrightarrow{P} E[ZY]$$

as $n \to \infty$. The desired result therefore follows from the CMT,

$$\hat{\beta}_n \xrightarrow{P} (\Pi' E[ZZ']\Pi)^{-1} \Pi' E[ZY] = \beta \ .$$

The second equality uses $\Pi' E[ZZ']\Pi = \Pi' E[ZX']$ together with the identification result from the previous chapter.

The consistency argument therefore has two pieces. The statistical piece is that the sample averages and first-stage coefficient estimates converge to fixed population quantities. The identification piece is that the limiting population expression equals $\beta$.

## 14.2 Limiting Distribution of TSLS Estimator

We now derive the asymptotic distribution of the TSLS estimator. In addition to the assumptions we made earlier, assume that $\text{Var}[ZU] = E[ZZ'U^2] < \infty$. Then

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \quad \text{as } n \to \infty \ ,$$

where

$$\mathbb{V} = (\Pi' E[ZZ']\Pi)^{-1} \Pi' \text{Var}[ZU] \Pi (\Pi' E[ZZ']\Pi)^{-1} \ .$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \left( \hat{\Pi}_n' \left( \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} Z_i U_i \right) \right) \ .$$

As in the preceding section, we have that

$$\begin{aligned} \hat{\Pi}_n &\xrightarrow{P} \Pi \\ \frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' &\xrightarrow{P} E[ZZ'] \end{aligned}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} Z_i U_i \xrightarrow{d} N(0, \mathrm{Var}[ZU]) \ .$$

The desired result thus follows from the CMT.

The key object in this display is the empirical moment $\frac{1}{n}\sum_i Z_i U_i$. At the true parameter value, the population moment is zero: $E[ZU] = 0$. In a finite sample, however, the sample analog is not exactly zero. The large-sample behavior of $\hat{\beta}_n - \beta$ is therefore driven by the random fluctuation of this sample moment around zero. The other terms in the display, namely $\hat{\Pi}_n$ and $\frac{1}{n}\sum_i Z_i Z_i'$, converge to fixed matrices and act like constants in the limit. This is why the CLT enters through $\frac{1}{\sqrt{n}}\sum_i Z_i U_i$, while the CMT and Slutsky's theorem handle the matrix terms.

The variance formula also has a useful interpretation. The middle term, $\mathrm{Var}[ZU] = E[ZZ'U^2]$, captures the variability of the moment $ZU$. The matrices on the left and right translate this moment variability into coefficient variability. In the simple just-identified scalar case, this translation depends inversely on the strength of the first stage, which is why weak instruments lead to imprecise IV estimates.

### 14.2.1 Estimating the Asymptotic Variance

A natural estimator of $\mathbb{V}$ is given by

$$\begin{aligned} \hat{\mathbb{V}}_n \quad = \quad & \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \\ & \times \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \hat{U}_i^2 \right) \hat{\Pi}_n \\ & \times \left( \hat{\Pi}_n' \left( \frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} , \end{aligned}$$

where $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ . This estimator can be shown to be consistent, i.e.,

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \quad \text{as } n \to \infty \ .$$

However, proving such a result is beyond the scope of this class; the main idea is to apply the LLN componentwise to the matrix averages and then use the CMT. The term

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \hat{U}_i^2$$

is the sample analog of $E[ZZ'U^2]$. Because it uses the squared residual observation by observation, it allows the variance of $U$ to vary across observations. Thus, this is the IV analog of the heteroskedasticity-robust variance estimator used in LS.

The important aspect for us is how we can use these results to: (a) test hypotheses about components of $\beta$, and (b) construct confidence intervals for each component of $\beta$. We do this next.

## 14.3 Using the Asymptotic Distribution for Inference

We now discuss how to use the tools developed so far to test hypotheses and build confidence intervals for a specific component of $\beta$, call it $\beta_j$ for $j = 1, \ldots, k$ (we may also do the same for the constant term but this is rarely an object of interest). Statistical inference is the set of formal procedures by which we use the estimator $\hat{\beta}_{n,j}$ to draw conclusions about the unknown population parameter $\beta_j$. A *hypothesis test* starts with a null hypothesis $H_0$, which describes the benchmark claim to be assessed, and an alternative hypothesis $H_1$, which describes the departures from the null that we want power to detect. A Type I error occurs when we reject $H_0$ even though $H_0$ is true. The *level* of a test is the largest rejection probability allowed under the null; in this class we usually denote it by $\alpha$.

### 14.3.1 Standard Errors

Before testing hypotheses, we need to distinguish three objects. The number $\hat{\beta}_{n,j}$ is the estimate. The diagonal element $\hat{\mathbb{V}}_{n,[j+1,j+1]}$ estimates the asymptotic variance of $\sqrt{n}(\hat{\beta}_{n,j}-\beta_j)$, where the $j+1$ index accounts for the intercept. Dividing by $n$ gives an estimate of the variance of $\hat{\beta}_{n,j}$ itself. The *standard error* is the square root of this estimated variance:

$$\operatorname{se}(\hat{\beta}_{n,j}) = \sqrt{\hat{\mathbb{V}}_{n,[j+1,j+1]}/n} \; . \tag{14.1}$$

The standard error estimates the sampling variability of the estimator $\hat{\beta}_{n,j}$. If we repeatedly drew new samples of size $n$ from the same population and recomputed TSLS each time, the standard error estimates how much the resulting estimates would vary across samples.

Suppose we want to test whether a particular component of $\beta$, say $\beta_j$, is equal to some constant $c$. This is equivalent to testing the hypothesis:

$$H_0 : \beta_j = c \quad \text{vs} \quad H_1 : \beta_j \neq c \; .$$

Our previous results then show that

$$\frac{\sqrt{n}(\hat{\beta}_{n,j} - \beta_j)}{\sqrt{\hat{\mathbb{V}}_{n,[j+1,j+1]}}} = \frac{\hat{\beta}_{n,j} - \beta_j}{\sqrt{\hat{\mathbb{V}}_{n,[j+1,j+1]}/n}} = \frac{\hat{\beta}_{n,j} - \beta_j}{\text{se}(\hat{\beta}_{n,j})} \xrightarrow{d} N(0,1) \ . \tag{14.2}$$

### 14.3.2 Testing a Hypothesis About One Coefficient

We now discuss how to test hypotheses about a specific component of $\beta$, call it $\beta_j$ for $j = 1, \ldots, k$. The same logic can be applied to the intercept, but in applications the slope coefficients are usually the main objects of interest.

Suppose we want to test whether $\beta_j$ is equal to some known value $c$. The hypothesis testing problem is

$$H_0 : \beta_j = c \quad \text{vs.} \quad H_1 : \beta_j \neq c \ .$$

The null hypothesis $H_0$ is the benchmark claim. The alternative hypothesis $H_1$ describes the values of $\beta_j$ that we regard as departures from the null. In this case, the alternative is two-sided because values both above and below $c$ contradict the null.

The starting point is the asymptotic normality result derived above. Recall that

$$\frac{\hat{\beta}_{n,j} - \beta_j}{\text{se}(\hat{\beta}_{n,j})} \xrightarrow{d} N(0,1) \ .$$

Under the null hypothesis $H_0 : \beta_j = c$, the unknown value $\beta_j$ can be replaced by the null value $c$. Therefore, under $H_0$,

$$t_n(c) \equiv \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \xrightarrow{d} N(0,1) \ .$$

This is the usual *t-statistic* for testing $H_0 : \beta_j = c$.

For a two-sided alternative, evidence against the null comes from values of $\hat{\beta}_{n,j}$ that are far from $c$ in either direction. Thus, both very positive and very negative values of $t_n(c)$ provide evidence against $H_0$. For this reason, it is convenient to define the test statistic as

$$T_n = |t_n(c)| = \left| \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \right| \ .$$

Notice the convention: in this two-sided test, $T_n$ is the absolute value of the usual $t$-statistic. We define it this way so that *large values of* $T_n$ always provide evidence against the null.

We will use a rejection rule of the form

$$\text{Reject } H_0 \quad \text{if and only if} \quad T_n > z^* \ ,$$

where $z^*$ is a critical value to be chosen. The question is: how should we choose $z^*$?

To answer this, recall that a Type I error occurs when we reject $H_0$ even though $H_0$ is true. A Type II error occurs when we fail to reject $H_0$ even though $H_0$ is false. These two errors can be summarized as follows:

| | Reject $H_0$ | Do not reject $H_0$ |
|---|---|---|
| $H_0$ true | Type I error | Correct decision |
| $H_0$ false | Correct decision | Type II error |

In hypothesis testing, we choose a significance level $\alpha \in (0,1)$, usually $\alpha = 0.10$, 0.05, or 0.01. The significance level is the probability of a Type I error that we are willing to tolerate, at least approximately in large samples. Thus, we choose the critical value $z^*$ so that, under the null,

$$P\{T_n > z^*\} \approx \alpha \ .$$

Using the asymptotic normal approximation, under $H_0$ we have

$$t_n(c) \approx N(0,1) \quad \text{and therefore} \quad T_n = |t_n(c)| \approx |N(0,1)| \ .$$

Hence,

$$P\{T_n > z^*\} \approx P\{|N(0,1)| > z^*\} = 2\{1 - \Phi(z^*)\} \ .$$

To make this probability equal to $\alpha$, we choose $z^*$ to solve

$$2\{1 - \Phi(z^*)\} = \alpha \ .$$

Equivalently,

$$\Phi(z^*) = 1 - \frac{\alpha}{2} \ ,$$

so

$$z^* = z_{1-\frac{\alpha}{2}} \ ,$$

where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution.

Therefore, the two-sided test of

$$H_0 : \beta_j = c \quad \text{vs.} \quad H_1 : \beta_j \neq c$$

uses the rejection rule

$$\text{Reject } H_0 \quad \text{if and only if} \quad T_n = \left| \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \right| > z_{1-\frac{\alpha}{2}} \ .$$

For the most common significance levels, the critical values are approximately

| $\alpha$ | 0.10 | 0.05 | 0.01 |
|---|---|---|---|
| $z_{1-\alpha/2}$ | 1.64 | 1.96 | 2.57 |

Thus, at the 5% level, the familiar rule is to reject whenever the absolute value of the $t$-statistic is larger than 1.96.

**Example 14.1** Suppose $\hat{\beta}_{n,j} = 1.8$ and $\text{se}(\hat{\beta}_{n,j}) = 0.3$. We want to test

$$H_0 : \beta_j = 2 \quad \text{vs.} \quad H_1 : \beta_j \neq 2$$

at the 5% significance level. The usual $t$-statistic is

$$t_n(2) = \frac{1.8 - 2}{0.3} = -0.667 \ .$$

Because this is a two-sided test, our test statistic is the absolute value:

$$T_n = |t_n(2)| = 0.667 \ .$$

At the 5% level, the critical value is $z_{0.975} = 1.96$. Since

$$0.667 < 1.96 \ ,$$

we do not reject $H_0$ at the 5% level. The estimate is below 2, but it is not far enough from 2, relative to its standard error, to reject the null. ■

#### 14.3.2.1 One-sided alternatives.

The definition of $T_n$ depends on the alternative hypothesis. In the two-sided case, we used $T_n = |t_n(c)|$ because deviations in both directions count as evidence against the null. For one-sided alternatives, only one direction counts.

For example, suppose the hypotheses are

$$H_0 : \beta_j \leq c \quad \text{vs.} \quad H_1 : \beta_j > c \ .$$

Here, evidence against the null comes from estimates that are much larger than $c$. Therefore, we use

$$T_n = t_n(c) = \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \ .$$

Large positive values of $T_n$ provide evidence against $H_0$. The rejection rule is

$$\text{Reject } H_0 \quad \text{if and only if} \quad T_n > z_{1-\alpha} \ .$$

The critical value is now $z_{1-\alpha}$, not $z_{1-\alpha/2}$, because the rejection region is only in the right tail of the standard normal distribution.

On the other hand, suppose the hypotheses are

$$H_0 : \beta_j \geq c \quad \text{vs.} \quad H_1 : \beta_j < c \ .$$

Now evidence against the null comes from estimates that are much smaller than $c$. To preserve the convention that large values of $T_n$ lead to rejection, we define

$$T_n = -t_n(c) = -\frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \ .$$

The rejection rule is again

$$\text{Reject } H_0 \quad \text{if and only if} \quad T_n > z_{1-\alpha} \ .$$

The important lesson is that $T_n$ is always defined so that large values provide evidence against the null. For a two-sided test, this means using the absolute value of the usual $t$-statistic. For a right-sided test, this means using the usual $t$-statistic. For a left-sided test, this means using the negative of the usual $t$-statistic.

### 14.3.3 P-values

The rejection rule described above requires us to choose a significance level $\alpha$ in advance. For example, at the 5% level, we reject the two-sided null hypothesis

$$H_0 : \beta_j = c \quad \text{vs.} \quad H_1 : \beta_j \neq c$$

whenever

$$T_n = \left| \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \right| > z_{1-\alpha/2} \ .$$

Equivalently, once we compute the test statistic, we can ask the following question: for which significance levels would this value of the test statistic lead us to reject the null? This question leads to the notion of a p-value.

For a two-sided test, recall that the test statistic is

$$T_n = |t_n(c)| \ ,$$

where

$$t_n(c) = \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \ .$$

Under the null hypothesis, $t_n(c)$ is approximately standard normal. Therefore, under $H_0$,

$$T_n = |t_n(c)| \approx |N(0,1)| \ .$$

The two-sided asymptotic *p-value* is defined as

$$p_n = P\{|N(0,1)| \geq T_n\} = 2\{1 - \Phi(T_n)\} \ .$$

This is the probability, computed under the null hypothesis, of observing a test statistic at least as large as the one observed.

The p-value can also be interpreted as the smallest significance level at which we would reject the null hypothesis. To see this, recall that the two-sided rejection rule at level $\alpha$ is

$$T_n > z_{1-\alpha/2} \ .$$

This is equivalent to

$$2\{1 - \Phi(T_n)\} < \alpha \ ,$$

which is equivalent to

$$p_n < \alpha \ .$$

Thus, instead of comparing $T_n$ to a critical value, we can compare the p-value to the significance level:

$$\text{Reject } H_0 \quad \text{if and only if} \quad p_n < \alpha \ .$$

The p-value is useful because it summarizes the strength of evidence against the null in a scale that is easy to compare across common significance levels. For example, if $p_n = 0.03$, then we reject the null at the 5% level but not at the 1% level. If $p_n = 0.20$, then we do not reject the null at the usual 10%, 5%, or 1% levels.

It is important, however, to interpret p-values correctly. A p-value is not the probability that the null hypothesis is true. The null hypothesis is treated as fixed in the calculation. The p-value is the probability, under the null, of obtaining a test statistic at least as extreme as the one observed. In other words, the p-value answers the question:

If $H_0$ were true, how surprising would this value of the test statistic be?

Small p-values indicate that the observed statistic would be unlikely under the null, and therefore provide evidence against $H_0$. Large p-values indicate that the observed statistic is not unusual under the null, and therefore do not provide strong evidence against $H_0$.

The p-values reported by most regression packages correspond to the default null hypothesis

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \ .$$

Thus, when reading regression output, the reported p-value for a coefficient usually answers the question: if the true coefficient were zero, how likely would it be to observe a $t$-statistic at least this large in absolute value?

**Example 14.2** Continue with the previous example. Suppose $\hat{\beta}_{n,j} = 1.8$ and $\text{se}(\hat{\beta}_{n,j}) = 0.3$, and suppose we want to test

$$H_0 : \beta_j = 2 \quad \text{vs.} \quad H_1 : \beta_j \neq 2 \ .$$

We computed the usual $t$-statistic:

$$t_n(2) = \frac{1.8 - 2}{0.3} = -0.667 \ .$$

Because this is a two-sided test, the test statistic is

$$T_n = |t_n(2)| = 0.667 \ .$$

The two-sided p-value is therefore

$$p_n = 2\{1 - \Phi(0.667)\} \ .$$

Using the standard normal distribution,

$$\Phi(0.667) \approx 0.748 \ ,$$

so

$$p_n \approx 2(1 - 0.748) = 0.504 \ .$$

Thus, the p-value is approximately 0.50.

This means that, if the null hypothesis $\beta_j = 2$ were true, observing a test statistic at least as large as 0.667 in absolute value would not be unusual. In fact, it would happen with probability about 0.50 under the null. Therefore, this p-value does not provide strong evidence against $H_0$.

Equivalently, because $p_n \approx 0.50$, we would not reject $H_0$ at the 10%, 5%, or 1% significance levels. This agrees with the conclusion from the critical-value approach: since $0.667 < 1.96$, we do not reject at the 5% level. ■

**Remark 14.1 (Common mistake)** It would be incorrect to say that the p-value 0.50 means there is a 50% probability that $H_0$ is true. The p-value does not assign probabilities to hypotheses. It assumes the null is true and then asks how unusual the observed test statistic would be under that assumption. ■

### 14.3.4 Confidence Intervals

We now use the hypothesis test developed above to construct a confidence interval for $\beta_j$. The key idea is the *duality* between tests and confidence intervals. Instead of testing only one null value, say

$$H_0 : \beta_j = c \ ,$$

we can imagine testing many different null values of $c$. Some values of $c$ would be rejected by the data, while others would not be rejected. The confidence interval is the set of null values that would *not* be rejected.

Consider the two-sided test

$$H_0 : \beta_j = c \quad \text{vs.} \quad H_1 : \beta_j \neq c$$

at significance level $\alpha$. From the previous section, the test rejects whenever

$$\left| \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \right| > z_{1-\frac{\alpha}{2}} \ .$$

Therefore, the test does *not* reject whenever

$$\left| \frac{\hat{\beta}_{n,j} - c}{\text{se}(\hat{\beta}_{n,j})} \right| \leq z_{1-\frac{\alpha}{2}} \ .$$

Equivalently,

$$|\hat{\beta}_{n,j} - c| \leq z_{1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{n,j}) \ .$$

Solving this inequality for $c$ gives

$$\hat{\beta}_{n,j} - z_{1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{n,j}) \leq c \leq \hat{\beta}_{n,j} + z_{1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{n,j}) \ .$$

Thus, the set of values of $c$ that would not be rejected by the two-sided test is

$$C_n = \left[\hat{\beta}_{n,j} - z_{1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{n,j}),\ \hat{\beta}_{n,j} + z_{1-\frac{\alpha}{2}} \text{se}(\hat{\beta}_{n,j})\right] \ .$$

This is the usual $(1-\alpha)$ confidence interval for $\beta_j$.

This construction shows why confidence intervals and two-sided hypothesis tests are closely related. A value $c$ belongs to the $(1-\alpha)$ confidence interval if and only if the two-sided test of

$$H_0 : \beta_j = c$$

does not reject at level $\alpha$. For example, a 95% confidence interval contains exactly the null values that would not be rejected by a two-sided test at the 5% level.

The interval has the asymptotic coverage property

$$P\{\beta_j \in C_n\} \to 1 - \alpha \quad \text{as } n \to \infty \ .$$

This is an asymptotic statement. In finite samples, the coverage is only approximately $1-\alpha$. The approximation becomes more accurate as the sample size grows, provided the assumptions behind the asymptotic normal approximation are reasonable.

It is important to interpret confidence intervals correctly. A 95% confidence interval is not, in the frequentist interpretation used here, a statement that there is a 95% probability that the fixed parameter lies in this particular realized interval. The parameter $\beta_j$ is fixed, and the interval is random because it is constructed from the data. The interpretation is that if we repeatedly drew samples and constructed intervals using the same procedure, approximately 95% of those intervals would contain the true value.

**Example 14.3** Continue with the previous example. Suppose

$$\hat{\beta}_{n,j} = 1.8 \quad \text{and} \quad \text{se}(\hat{\beta}_{n,j}) = 0.3 \ .$$

The 95% confidence interval is

$$[1.8 - 1.96(0.3),\ 1.8 + 1.96(0.3)] = [1.21,\ 2.39] \ .$$

This interval contains 2. Therefore, the two-sided test of

$$H_0 : \beta_j = 2$$

does not reject at the 5% level. This agrees with the test statistic computed above:

$$t_n(2) = \frac{1.8 - 2}{0.3} = -0.667 \ ,$$

whose absolute value is smaller than 1.96.

On the other hand, the interval does not contain, for example, 0. Therefore, the two-sided test of

$$H_0 : \beta_j = 0$$

would reject at the 5% level. In this sense, the confidence interval provides a compact summary of many hypothesis tests at once. ■

### 14.3.5 Empirical Illustration

Consider the empirical application in Baum-Snow [2007], who studied the growth of suburbs in U.S. cities. See Section 13.2.1 for details. In that application, we run the IV command in R and obtain the following:

```
Call:
ivreg(formula = dlcpop2 ~ draccld + rcarea + Dlincomeh + dlpop |
    rays_planc + rcarea + Dlincomeh + dlpop, data = dw_filtered)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.58804    0.24562  -2.394    0.018 *
draccld     -0.12322    0.02890  -4.264 3.77e-05 ***
rcarea       0.11308    0.01973   5.730 6.33e-08 ***
Dlincomeh    0.04768    0.36166   0.132    0.895
dlpop        0.42436    0.06186   6.860 2.31e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code Snippet 14.1: IV results

Here, the first column reports the estimates for each $\beta_j$, the second column reports the standard error (as defined above), the third column reports the t value, and the final column reports a p-value for the null hypothesis $H_0 : \beta_j = 0$. For example, the coefficient on `draccld` is $-0.12322$ with standard error 0.02890, so the reported t-statistic is $-0.12322/0.02890 \approx -4.264$. The p-value $3.77 \times 10^{-5}$ says that the coefficient would be rejected against zero even at very small significance levels. An approximate 95% confidence interval is

$$-0.12322 \pm 1.96(0.02890) = [-0.180, \ -0.067] \ .$$

In the Baum-Snow application, this interval is more informative than the p-value alone because it gives a range of economically plausible magnitudes. Under the maintained IV assumptions, the estimate suggests that one additional radial highway caused a decline in central-city population growth. Accounting for sampling uncertainty, values between about $-0.18$ and $-0.07$ are plausible at the 95% level. The p-value says the estimate is statistically

distinguishable from zero; the confidence interval helps us assess the range of economically meaningful magnitudes.

The default standard errors reported by `ivreg` assume homoskedasticity. In cross-sectional applications this assumption is often difficult to defend, so one should typically report heteroskedasticity-robust standard errors. In R, these can be computed by combining `ivreg` with, for example, `coeftest` and `vcovHC` from the `lmtest` and `sandwich` packages. Up to finite-sample degrees-of-freedom adjustments, these robust standard errors correspond to the estimator $\hat{\mathbb{V}}_n$ defined above, which allows the conditional variance of $U$ to vary with the instruments.

## 14.4 Key Concepts

- The TSLS estimator is consistent and has an asymptotically normal distribution with variance matrix $\mathbb{V}$ that can be consistently estimated.
- The standard error of $\hat{\beta}_{n,j}$ estimates the standard deviation of the estimator and is the key input for both hypothesis tests and confidence intervals.
- Hypothesis testing for a component $\beta_j$ is based on $t$-statistics formed from $\hat{\beta}_{n,j}$, the null value, and the standard error.
- P-values report the smallest significance level at which the null hypothesis would be rejected, given the observed test statistic. They are not probabilities that the null is true.
- $(1 - \alpha)$ confidence intervals for $\beta_j$ are constructed using standard normal critical values and the standard error.
- Weak instruments make TSLS estimates imprecise and can make the usual normal approximation unreliable, so applied IV work should report and interpret first-stage evidence.

## 14.5 Concluding Remarks

The material today borrows from notes by Azeem Shaikh that he kindly shared with me. In general, the topic of linear IV and endogeneity is covered in most standard sources, including Hansen [2022] and Wooldridge [2010].

## 14.6 Problems

**Problem 14.1** *Consider the just-identified case with a single endogenous regressor and a single instrument, so* $X = (1, X_1)'$ *and* $Z = (1, Z_1)'$*. The model is*

$$Y = \beta_0 + \beta_1 X_1 + U$$

*with* $E[U] = E[Z_1 U] = 0$*. Assume* $\mathrm{Cov}(Z_1, X_1) \neq 0$ *and that all relevant second moments are finite.*

*(a) Show that the IV estimator of* $\beta_1$ *can be written as*

$$\hat{\beta}_{1,\mathrm{iv}} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \bar{Z}_{1,n})U_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_{1i} - \bar{Z}_{1,n})X_{1i}} \ .$$

*(b) Apply the LLN, CLT, and CMT to derive the asymptotic distribution of* $\sqrt{n}(\hat{\beta}_{1,\mathrm{iv}} - \beta_1)$*. Express the asymptotic variance in terms of* $\mathrm{Cov}(Z_1, X_1)$ *and* $\mathrm{Var}((Z_1 - E[Z_1])U)$*.*

*(c) Suppose* $\mathrm{Cov}(Z_1, X_1)$ *is small. What does this imply for the asymptotic variance and for the precision of the IV estimator?*

**Problem 14.2** *Suppose that for a given coefficient* $\beta_1$ *in a TSLS regression you have* $\hat{\beta}_{n,1} = 0.45$ *and* $\mathrm{se}(\hat{\beta}_{n,1}) = 0.20$*.*

*(a) Test* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *at the 5% level. Report the t-statistic and the conclusion.*

*(b) Compute the two-sided p-value. State, in one sentence, what the p-value means.*

*(c) Construct the 95% and 99% confidence intervals for* $\beta_1$*.*

*(d) Suppose economic theory predicts* $\beta_1 > 0$*. State the appropriate null and alternative, the appropriate rejection rule at the 5% level, and conduct the one-sided test.*

*(e) Mark each of the following statements as true or false and justify briefly:*

*(i) If the 95% confidence interval for $\beta_1$ excludes zero, then the two-sided test of $H_0 : \beta_1 = 0$ rejects at the 5% level.*

*(ii) If a 95% confidence interval and a 99% confidence interval both contain a value $c$, then the test of $H_0 : \beta_1 = c$ does not reject at the 1% level.*

*(iii) If the p-value for $H_0 : \beta_1 = 0$ is 0.04 and the p-value for $H_0 : \beta_1 = 0.10$ is 0.18, then the 95% confidence interval contains 0.10 but does not contain 0.*

*(iv) If the p-value is 0.04, there is a 4% probability that $H_0$ is true.*

**Problem 14.3** *You estimate a TSLS regression and obtain $\hat{\beta}_1 = -0.80$ with standard error 0.35. The first-stage coefficient on the excluded instrument is 0.12 with standard error 0.04.*

*(a) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the 5% level. Report the t-statistic and the conclusion.*

*(b) Construct a 95% confidence interval for $\beta_1$.*

*(c) Interpret the confidence interval in words.*

*(d) Explain why the first-stage estimate matters for interpreting the IV results.*

*(e) A classmate says, "The p-value is 0.02, so there is a 2% probability that the null is true." Explain what is wrong with this statement.*

## Bibliography

N. Baum-Snow. Did highways cause suburbanization? *The quarterly journal of economics*, 122(2):775–805, 2007.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

# 15

# *Heterogeneous and Endogenous Treatments*

Up until this point we have focused our attention on a linear model with homogeneous effects. In such models,

$$Y = X'\beta + U \ ,$$

and the effect of a change in $X$ (say, from $X = x$ to $X = x'$) is the *same* for everybody, and captured by the constant $\beta$. Despite requiring strong assumptions, TSLS remains popular. One reason stems from the fact that this estimand has a well-understood interpretation under (unobserved) heterogeneity (i.e., cases where the effect of a change in $X$ on $Y$ would be expected to be different for different people). The easiest way to allow for such heterogeneity, as we have done in previous classes, is to allow for $\beta$ to be *random*. When $\beta$ is random, we may absorb $U$ into the intercept and simply write

$$Y = X'\beta \ .$$

In this random-coefficients notation, the residual variation is built into the random intercept and slope. Note that this means that when we work with a random sample where variables are indexed by $i$, we would write $Y_i = X_i'\beta_i$, which makes it explicit that every individual has a unique effect $\beta_i$.

## 15.1 Wald Estimand, Heterogeneity, and LATE

We now study the properties of TSLS, and the Wald estimand in particular, in the presence of heterogeneous effects of the variables $X$ on $Y$. In order to provide the cleanest possible exposition, assume that $k = 1$ and let $X = (1, A)'$ with $A$ being a binary random variable taking values in $\{0, 1\}$. In this notation,

$$Y = \beta_0 + \beta_1 A \ .$$

In this case, we interpret $\beta_0$ as $Y(0)$ and $\beta_1$ as $Y(1) - Y(0)$, where $Y(1)$ and $Y(0)$ are *potential* or *counterfactual outcomes*. Using this notation, we may rewrite the equation as

$$Y = AY(1) + (1 - A)Y(0) \ .$$

The potential outcome $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 0; the potential outcome $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) $A$ were 1. Following the same terminology we have been using in previous lectures, the variable $A$ is called the *treatment*, $Y(1) - Y(0)$ is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is referred to as the *average treatment effect*.

If $A$ were randomly assigned (e.g., by the flip of a coin, as in a randomized controlled trial), then

$$(Y(0), Y(1)) \perp\!\!\!\perp A \ .$$

In this case, under mild assumptions, the slope coefficient from OLS regression of $Y$ on a constant and $A$ yields a consistent estimate of the average treatment effect; see Section 5.2.

We generally expect $A$ to depend on $(Y(1), Y(0))$. For example, in the application in Angrist and Evans [1998] we would expect that the probability of having another child may be decreasing in $Y(0)$.

In this case, OLS will not yield a consistent estimate of the average treatment effect. To proceed further, we therefore assume, as usual, that there is an instrument $Z$ that also takes values in $\{0, 1\}$. We may thus consider the slope coefficient from TSLS regression of $Y$ on $A$ with $Z$ as an instrument. The estimand in this case is

$$\frac{\text{Cov}[Y, Z]}{\text{Cov}[A, Z]} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[A \mid Z = 1] - E[A \mid Z = 0]} \ ,$$

where the equality follows by multiplying and dividing by $\text{Var}[Z]$ and using the binary-regressor formula $\text{Cov}[Y, Z]/\text{Var}[Z] = E[Y \mid Z = 1] - E[Y \mid Z = 0]$, with the analogous expression for $A$. Our goal is to express this quantity in terms of the treatment effect $Y(1) - Y(0)$ somehow. To this end, analogously to our equation for $Y$ above, it is useful to also introduce a similar equation for $A$:

$$\begin{aligned} A &= ZA(1) + (1 - Z)A(0) \\ &= A(0) + (A(1) - A(0))Z \\ &= \pi_0 + \pi_1 Z \ , \end{aligned}$$

where $\pi_0 = A(0)$, $\pi_1 = A(1) - A(0)$, and $A(1)$ and $A(0)$ are *potential* or *counterfactual treatments* (rather than outcomes). We impose the following versions of instrument exogeneity and instrument relevance, respectively:

$$(Y(1), Y(0), A(1), A(0)) \perp\!\!\!\perp Z$$

and

$$P\{A(1) \neq A(0)\} = P\{\pi_1 \neq 0\} > 0 \ .$$

Note that the first part of the assumption basically states that $Z$ is as good

as randomly assigned. In addition, note that we are implicitly assuming that $Z$ does not affect $Y$ directly, i.e., potential outcomes take the form $Y(a)$ as opposed to $Y(a, z)$. This is the exclusion restriction in this setting.[1] In the linear model with constant effects, the exclusion restriction is expressed by the omission of the instruments from the causal equation of interest. Exogeneity is then imposed separately by requiring that $E[ZU] = 0$.

We further assume the following *monotonicity* (or perhaps better called *uniform monotonicity*) condition:

$$P\{A(1) \geq A(0)\} = P\{\pi_1 \geq 0\} = 1 \ .$$

The monotonicity assumption states that while the instrument may have no effect on some people, all those who are affected are affected *in the same way.*

Under these assumptions, note that

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] &= E[Y(1)A(1) + Y(0)(1 - A(1)) \mid Z = 1] \\ &\quad - E[Y(1)A(0) + Y(0)(1 - A(0)) \mid Z = 0] \\ &= E[Y(1)A(1) + Y(0)(1 - A(1))] \\ &\quad - E[Y(1)A(0) + Y(0)(1 - A(0))] \\ &= E[(Y(1) - Y(0))(A(1) - A(0))] \\ &= E[Y(1) - Y(0)|A(1) > A(0)]P\{A(1) > A(0)\} \ , \end{aligned}$$

where the first equality follows from the equations for $Y$ and $A$, the second equality follows from instrument exogeneity, and the third equality follows from algebraic rearrangement. The fourth equality follows because, under monotonicity, $A(1) - A(0) = I\{A(1) > A(0)\}$. Furthermore,

$$E[A \mid Z = 1] - E[A \mid Z = 0] = E[A(1) - A(0)] = P\{A(1) > A(0)\} \ .$$

Hence, the Wald estimand equals

$$E[Y(1) - Y(0)|A(1) > A(0)] \ ,$$

which is termed the *local average treatment effect* (LATE). It is the average treatment effect among the subpopulation of people for whom a change in the value of the instrument switched them from being non-treated to treated. We often refer to such subpopulation as *compliers.*

A few remarks are in order. First, this result depends crucially on the monotonicity assumption. Second, this quantity may or may not be of interest. Third, a consequence of this calculation is that in a world with heterogeneity "different instruments estimate different parameters." Finally, this result also depends on the simplicity of the model. When covariates are present, the entire calculation breaks down as we will illustrate below.

[1] In the linear-model setting, "excluded instrument" means that the instrument is omitted from the causal equation of interest. In the potential-outcomes setting, exclusion is the stronger statement that $Z$ affects $Y$ only through $A$.

### 15.1.1 The Importance of Monotonicity

It is instructive to understand the power of monotonicity to obtain LATEs. Without monotonicity, $A(1) - A(0)$ can take the values $-1$, 0, and 1, so the expectation splits across compliers and defiers. In particular, we would have

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] = {} & E[Y(1) - Y(0) | A(1) > A(0)] P\{A(1) > A(0)\} \\ & - E[Y(1) - Y(0) | A(1) < A(0)] P\{A(1) < A(0)\} \ . \end{aligned}$$

We might therefore have a situation where treatment effects are positive for everyone (i.e., $Y(1) - Y(0) > 0$) yet the reduced form is zero because effects on compliers are canceled out by effects on *defiers*, i.e., those individuals for whom the instrument pushes them out of treatment ($A(1) = 0$ and $A(0) = 1$). This doesn't come up in a constant effect model where $\beta = Y(1) - Y(0)$ is constant, as in such case

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] &= \beta\{P\{A(1) > A(0)\} - P\{A(1) < A(0)\}\} \\ &= \beta E[A(1) - A(0)] \ , \end{aligned}$$

and so a zero reduced-form effect means either the first stage is zero or $\beta = 0$.

For a simple numerical illustration, suppose 60 percent of the population are compliers and 40 percent are defiers. Suppose also that the treatment effect is positive for both groups: it equals 2 for compliers and 3 for defiers. The first stage is positive, since $0.6 - 0.4 = 0.2$, but the reduced form is proportional to

$$2 \times P\{A(1) > A(0)\} - 3 \times P\{A(1) < A(0)\} = 2(0.6) - 3(0.4) = 0 \ .$$

Thus, the instrument can have a zero effect on the outcome even though treatment is beneficial for everyone. This is the cancellation that monotonicity rules out.

The monotonicity assumption is difficult to defend in *many* empirical settings. Yet, it is massively popular and researchers often perceive the assumption to be mild. Why? It turns out that monotonicity does hold in randomized controlled experiments with one-sided compliance. That is, consider the context of randomized trials, where treatment assignment is independent of potential outcomes by design. The fact that the treatment is randomly assigned does not mean that every unit that is assigned to treatment actually takes the treatment. In medical trials, units may get a certain new medication for a disease but may decide not to take it at home. In this case one could interpret the treatment assignment as an "offer of treatment" $Z$ (the instrument), and the actual treatment $A$ as the variable that determines whether the subject actually had the intended treatment. This is the case in experiments where participation is voluntary among those randomly assigned to receive treatment. At the same time, it is often the case that no one in the control group has access to the experimental intervention. In other words, $A(0) = 0$ while $A(1) \in \{0, 1\}$. It immediately follows that $A(1) \geq A(0)$ a.s., and monotonicity

automatically holds. The compliers are exactly the units with $A(1) = 1$, i.e., those who would take the treatment if offered, and this set need not be representative of everyone offered treatment. Hence, a comparison between those actually treated ($A = 1$) and the control ($A = 0$) group is misleading.

In the setting we just described, two alternatives are frequently used. The first one is a comparison between those who were *offered* treatment ($Z = 1$) and the control ($Z = 0$) group. This comparison is based on randomly assigned $Z$ and identifies a parameter known as the *intention to treat* (ITT) effect:

$$E[Y \mid Z = 1] - E[Y \mid Z = 0] \ .$$

The ITT effect measures the causal effect of being offered treatment, not necessarily the causal effect of receiving treatment. The second alternative is to use the randomly assigned offer $Z$ as an instrumental variable for actual treatment received $A$, which solves the sort of compliance problem previously discussed.

In this one-sided noncompliance case, LATE returns the effect of *treatment on the treated*, i.e., $E[Y(1) - Y(0)|A = 1]$. To see why, note that $A(0) = 0$ implies that the compliers are exactly the units with $A(1) = 1$. Also, $A = A(1)Z$, so the event $\{A = 1\}$ is the same as $\{A(1) = 1, Z = 1\}$. By random assignment of $Z$, conditioning on $Z = 1$ does not change the distribution of potential outcomes, and therefore

$$E[Y(1) - Y(0) \mid A = 1] = E[Y(1) - Y(0) \mid A(1) = 1] = \text{LATE} \ .$$

## 15.2 An Application: Angrist and Evans revisited

Let's go back to the application we discussed in Section 13.5. Angrist and Evans look at the effect of fertility decisions (having an extra child) on female labor force participation and earnings. To recap the notation, we denote by $Y$ the outcome of interest, which will be a measure of labor market participation or outcome for the woman (or her husband). Angrist and Evans restrict the sample to only women (or couples) with 2 or more children, and so $A$ is an indicator for having more than 2 children (versus exactly 2). That is,

$$A := I\{\text{number of children} > 2\} \ .$$

The authors then consider two instruments for $A$, but the main one is $Z = 1$ if the first two children had the *same sex*. Part of the reason the authors prefer this instrument over the twins instrument is that, as we have discussed earlier, the twins instrument is less likely to be exogenous. But what about monotonicity? Monotonicity requires that $P\{A(1) \geq A(0)\} = 1$. This is equivalent to saying that there are no "defiers", i.e., individuals such that

| | 1980 PUMS | | |
|---|---|---|---|
| | | Wald estimate using as covariate: | |
| Variable | Mean difference by *Same sex* | *More than 2 children* | *Number of children* |
| *More than 2 children* | 0.0600 (0.0016) | — | — |
| *Number of children* | 0.0765 (0.0026) | — | — |
| *Worked for pay* | −0.0080 (0.0016) | −0.133 (0.026) | −0.104 (0.021) |
| *Weeks worked* | −0.3826 (0.0709) | −6.38 (1.17) | −5.00 (0.92) |
| *Hours/week* | −0.3110 (0.0602) | −5.18 (1.00) | −4.07 (0.78) |
| *Labor income* | −132.5 (34.4) | −2208.8 (569.2) | −1732.4 (446.3) |
| ln(*Family income*) | −0.0018 (0.0041) | −0.029 (0.068) | −0.023 (0.054) |

TABLE 15.1: Table 5 in AE98. Wald Estimates for five different outcomes and 2 instruments.

$\{A(1) = 0, A(0) = 1\}$. In this context, it means that there are no individuals who really want to have two girls (or two boys). Note that the intuition that people tend to have a *preference for a mix of boys and girls* does not imply that there are no people who *want to have two kids of the same sex*. Monotonicity, however, requires such an assumption and therefore restricts preference heterogeneity in unattractive ways (some families may want two boys or girls).

Table 15.1 reports Wald estimates, including the one we presented in Section 13.5, whereas Table 15.2 reports the coefficient $\beta_{2\text{sls}}$ in the following regression,

$$Y = \beta_0 + \beta_{2\text{sls}} A + \gamma' W + U \tag{15.1}$$

using $Z$ as an excluded instrument for $A$, and where $W$ is a vector of demographic covariates that includes age, age at first birth, plus indicators for the sex of the first and second child, Black, Hispanic, and other race.

There are multiple numbers to interpret in these Tables. A first observation is that OLS estimates tend to be quite different from the IV counterparts, which is consistent with endogeneity (or selection). A second observation is that the Wald estimates, which do not control for covariates, tend to be further from the OLS estimates than the estimates of $\beta_{2\text{sls}}$. However, we have not yet discussed the interpretation of $\beta_{2\text{sls}}$. Finally, Table 15.1 reports Wald estimates when the variable $A$ is re-defined to be the "number of children" (which is no longer binary), which introduces yet another deviation from the Wald es-

TABLE 7—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1980 CENSUS DATA

| | All women | | | Married women | | | Husbands of married women | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Estimation method | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Instrument for *More than 2 children* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* | — | *Same sex* | *Two boys, Two girls* |
| Dependent variable: | | | | | | | | | |
| *Worked for pay* | −0.176 (0.002) | −0.120 (0.025) | −0.113 (0.025) [0.013] | −0.167 (0.002) | −0.120 (0.028) | −0.113 (0.028) [0.013] | −0.008 (0.001) | 0.004 (0.009) | 0.001 (0.008) [0.013] |
| *Weeks worked* | −8.97 (0.07) | −5.66 (1.11) | −5.37 (1.10) [0.017] | −8.05 (0.09) | −5.40 (1.20) | −5.16 (1.20) [0.071] | −0.82 (0.04) | 0.59 (0.60) | 0.45 (0.59) [0.030] |
| *Hours/week* | −6.66 (0.06) | −4.59 (0.95) | −4.37 (0.94) [0.030] | −6.02 (0.08) | −4.83 (1.02) | −4.61 (1.01) [0.049] | 0.25 (0.05) | 0.56 (0.70) | 0.50 (0.69) [0.71] |
| *Labor income* | −3768.2 (35.4) | −1960.5 (541.5) | −1870.4 (538.5) [0.126] | −3165.7 (42.0) | −1344.8 (569.2) | −1321.2 (565.9) [0.703] | −1505.5 (103.5) | −1248.1 (1397.8) | −1382.3 (1388.9) (0.549) |
| ln(*Family income*) | −0.126 (0.004) | −0.038 (0.064) | −0.045 (0.064) [0.319] | −0.132 (0.004) | −0.051 (0.056) | −0.053 (0.056) [0.743] | — | — | — |
| ln(*Non-wife income*) | — | — | — | −0.053 (0.005) | 0.023 (0.066) | 0.016 (0.066) [0.297] | — | — | — |

TABLE 15.2: Table 7 in AE98. Reports estimates of $\beta_{\text{2sls}}$ in eq.(15.1).

timand (that requires a binary-binary situation). We discuss the implications of some of these generalizations in the next section.

## 15.3 When TSLS is Not LATE

The previous section shows that under certain assumptions—binary instrument, binary treatment, no covariates, monotonicity, and exclusion—TSLS recovers a clear causal object: the local average treatment effect (LATE). However, this LATE interpretation does *not* always carry over to more general settings. Here we highlight a few important cases where TSLS no longer equals LATE.

### 15.3.1 Multivalued Instruments

Binary instruments are common, but so are multivalued instruments. Suppose that $Z$ takes $L+1$ values $\{z_0, z_1, \ldots, z_L\}$ while the treatment $A$ remains binary. Exogeneity now requires

$$Z \perp\!\!\!\perp (A(z_0), A(z_1), \ldots, A(z_L), Y(0), Y(1)) \ ,$$

and a natural monotonicity assumption is

$$P\{A(z_0) \leq A(z_1) \leq \cdots \leq A(z_L)\} = 1 \ .$$

This means that the values of the instrument can be ordered in terms of how strongly they encourage treatment, and that the ordering is the same for everyone. For example, if $Z$ is a price or subsidy, different values of $Z$ can induce different groups of people to take treatment.

Multiple instrument values therefore produce multiple complier groups. The group induced into treatment when $Z$ changes from $z_0$ to $z_1$ need not be the same as the group induced when $Z$ changes from $z_1$ to $z_2$. A linear IV estimand can still be written as a weighted average of these different local effects, but the weights depend on how the multivalued instrument is coded and used in the regression. This is why using $Z$ itself as a single scalar instrument can deliver a different estimand from using indicators such as

$$I\{Z = z_1\}, I\{Z = z_2\}, \ldots, I\{Z = z_L\}$$

as instruments. The broad lesson is that with multivalued instruments, the LATE interpretation becomes less transparent and researchers must be explicit about which variation in $Z$ is being used.

### 15.3.2 Covariates

When covariates are included, it matters how they enter the IV regression. If the regression includes exogenous covariates $W$ linearly, such as

$$Y = \beta_0 + \beta_{\text{tsls}} A + \gamma' W + U \ ,$$

and we estimate $\beta_{\text{tsls}}$ using $Z$ as an instrument for $A$, then $\beta_{\text{tsls}}$ is not generally a weighted average of conditional LATEs. The difficulty is that the regression is using variation in both $Z$ and $W$, and the resulting coefficient may depend on potential outcome levels, not only on treatment effects.

There are special cases in which a LATE-type interpretation can be recovered. For example, when $W$ is discrete and the specification is saturated in $W$ (that is, it includes indicators for each value of $W$), the coefficient on $A$ can be written as a weighted average of conditional LATEs. But the weights are determined by the covariance between the instrument and treatment within covariate cells, so the interpretation is still more complicated than in the binary-binary case without covariates.

### 15.3.3 Nonbinary Treatments

The classic LATE interpretation assumes binary $A$. If $A$ is continuous or multivalued, running a linear regression on $A$ generally does not lead to a coefficient that can be expressed as a weighted average of LATEs. In the Angrist–Evans regression with covariates and alternative definitions of fertility, $\beta_{\text{2sls}}$ therefore does not in general equal a clean LATE. Applied work often interprets such coefficients as LATE-like objects, but the additional structure needed for that interpretation should be stated explicitly.

These limitations imply that we must be cautious about interpreting the coefficient on $A$ as a LATE, or as a weighted average of conditional LATEs. Whenever monotonicity is questionable, or when the design departs from the binary-binary setup, researchers should clearly justify whether a LATE-type interpretation remains meaningful.

## 15.4 Key Concepts

- Under exogeneity, exclusion, instrument relevance, and monotonicity, TSLS identifies the *local average treatment effect* (LATE): the average treatment effect for *compliers*—those whose treatment status is affected by the instrument.
- The LATE interpretation is specific to the binary-binary case without covariates. In more general cases (e.g., covariates, multivalued instruments, or nonbinary treatments), TSLS typically does *not* recover a LATE.
- Monotonicity is critical for the LATE interpretation. Without it, the Wald estimand may reflect a net effect of compliers and defiers and become difficult to interpret.
- In randomized experiments with one-sided noncompliance, monotonicity often holds by design, and TSLS using assignment as an instrument recovers the effect of treatment on compliers, which in this setting is also the effect of treatment on the treated.

## 15.5 Concluding Remarks

The material today borrows from several useful sources, most notably the lecture notes kindly shared by Alex Torgovitsky, Azeem Shaikh, and material in Angrist and Pischke [2008].

## 15.6 Problems

**Problem 15.1** *Consider a randomized experiment where everybody that gets assigned to treatment receives the treatment (like a vaccine), but people that do not get randomly chosen to get the treatment may obtain it outside the context of the experiment (say, they could obtain the vaccine somewhere else). Would monotonicity hold in this setting ?*

**Problem 15.2** *Consider the binary treatment, binary instrument case under the instrument exogeneity assumption and monotonicity assumption. Show that the following quantities are identified for any $y$ (can be expressed as functions of the distribution of $(Y, A, Z)$):*

- $P\{Y(1) \leq y \mid A(1) = A(0) = 1\}$.
- $P\{Y(0) \leq y \mid A(1) = A(0) = 0\}$.
- $P\{Y(1) \leq y \mid A(1) > A(0)\}$.
- $P\{Y(0) \leq y \mid A(1) > A(0)\}$.

*Provide an intuitive explanation of these identification results.*

**Problem 15.3** *Consider the binary-treatment, binary-instrument LATE setup with $A(0) = 0$ for all units, i.e., units assigned to control cannot access the treatment. Maintain instrument exogeneity,*

$$(Y(1), Y(0), A(1), A(0)) \perp\!\!\!\perp Z \ ,$$

*and suppose $P\{A(1) = 1\} > 0$.*

*(a) Show that the set of compliers $\{A(1) > A(0)\}$ coincides with $\{A(1) = 1\}$.*

*(b) Show that the realized treatment satisfies $A = A(1)Z$.*

*(c) Show that $\{A = 1\} = \{A(1) = 1, Z = 1\}$.*

*(d) Use the previous parts and instrument exogeneity to show that*

$$E[Y(1) - Y(0) \mid A = 1] = E[Y(1) - Y(0) \mid A(1) = 1] = \text{LATE} \ .$$

*(e) Briefly explain why this identification result fails when $A(0) \neq 0$ for some units.*

**Problem 15.4** *Consider a population with three types of units, defined by their potential treatments $(A(0), A(1))$:*

- *Compliers:* $(A(0), A(1)) = (0, 1)$*, with population fraction* $\pi_c$ *and constant treatment effect* $\tau_c$*.*
- *Defiers:* $(A(0), A(1)) = (1, 0)$*, with population fraction* $\pi_d$ *and constant treatment effect* $\tau_d$*.*
- *Always-takers:* $(A(0), A(1)) = (1, 1)$*, with population fraction* $\pi_{at}$*.*

*Assume there are no never-takers, so* $\pi_c + \pi_d + \pi_{at} = 1$*, and assume* $Z$ *is independent of all potential outcomes and potential treatments.*

*(a) Compute the first-stage effect* $E[A \mid Z = 1] - E[A \mid Z = 0]$ *as a function of* $(\pi_c, \pi_d, \pi_{at})$*.*

*(b) Compute the reduced-form effect* $E[Y \mid Z = 1] - E[Y \mid Z = 0]$ *as a function of* $(\pi_c, \pi_d, \tau_c, \tau_d)$*.*

*(c) Find values of* $(\pi_c, \pi_d, \pi_{at}, \tau_c, \tau_d)$ *such that the first-stage effect is positive,* $\tau_c > 0$*,* $\tau_d > 0$*, but the reduced form is zero.*

*(d) Show that if* $\tau_c = \tau_d = \tau > 0$*, then a positive first-stage effect forces a positive reduced form. Explain why this makes monotonicity central in the heterogeneous-effects model but not in the constant-effects model.*

## Bibliography

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2008.