Causal inference in econometrics

Leonard Goff

This version: April 12, 2023

Guide to using these notes

These notes feature two kinds of box, to help organize the material:

Gray boxes sometimes offer section summaries.

White boxes indicate material that is optional, and understanding this material is not required for the course or exam.

Sections that have an asterisk * at the end of their title can be skipped in their entirety (for students in ECON 497 or ECON715): understanding this material is not required for the course or exam. These sections are mostly there for your interest and reference. Sections with a dagger † at the end of their title can be skipped for students in ECON497 but are required for students in ECON715.

I use the convention that := denotes equalities that are definitions.

Chapter 1

Causality and the experimental ideal

Most interesting questions in social science concern causality. We aren't just interested in observing what happens in the social world, but understanding *how* and *why* they happen as they do. And we're usually interested in what changes to policy or behavior could lead to changes that we might deem desirable.

These types of questions concern causality. The meaning of the term "causal" is a long-standing philosophical question; see Lewis (1973) for a fairly modern treatment that will accord with our approach in this class. We will take a very simple perspective: A causes B if B *would* be different if A were different. For example, on a day in which rain was forecast and I took my umbrella to school, we might say that the rain forecast caused me to bring my umbrella, if I *wouldn't have* taken the umbrella, absent the forecast for rain. We of course can't directly observe what would have happened if the forecast had been different; we call this a *counterfactual*.

1.1 Potential outcomes and treatment effects

The potential outcomes framework offers an elegant and tractable way to talk about counterfactuals, in the language of random variables (Rubin, 1974). This connects questions of causality to questions of statistics, which we have been developing tools to study.

As a running example, consider the question of the effect of obtaining a college degree on a worker's earnings. Suppose we have data in the form of an $i.i.d$ sample of (D_i, Y_i), where $D_i \in \{0, 1\}$ indicates whether individual i completed a college degree, and Y_i indicates the workers average hourly earnings at age 30. We call D_i our *treatment* variable, and Y_i our *outcome* variable. We're interested in the causal effect of the treatment variable on the outcome. This is a setting in which we have a *binary* treatment. We'll start here because it's the simplest setting to develop the concept of causality. In Section 1.7 I'll discuss how these ideas generalize beyond a binary treatment.

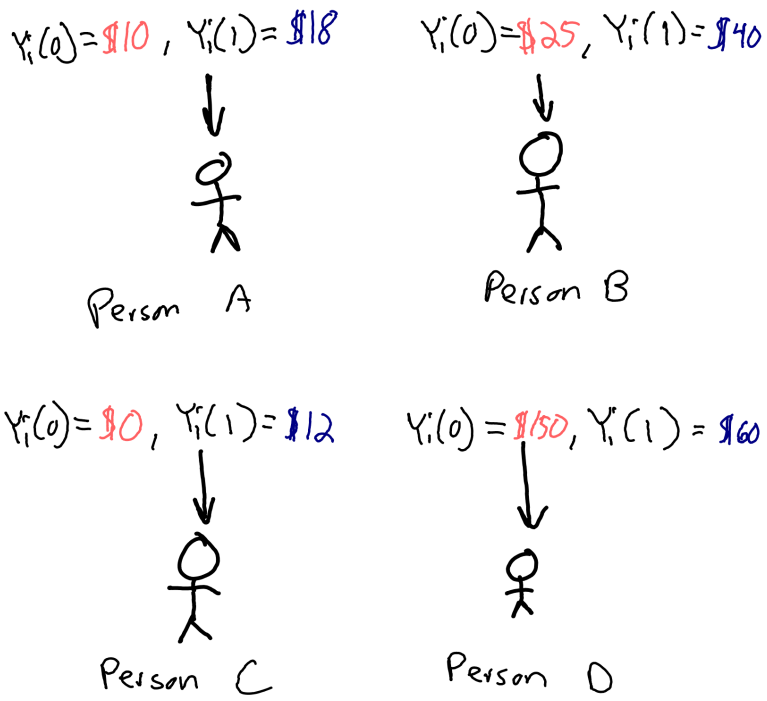
Definition 1.1. *An individual's* ***potential outcomes*** *are: $Y_i(1), Y_i(0)$, where $Y_i(1)$ is the outcome they would receive if they received the treatment, and the outcome $Y_i(0)$ they would receive if they did not.*

In the returns-to-college example, $Y_i(0)$ is the earnings i would have if they didn't go to college, and $Y_i(1)$ is the earnings that i would have if they did go to college. The key thing to keep in mind in the definition of counterfactuals is that we assume each individual i has a well-defined value both of $Y_i(0)$ *and* of $Y_i(1)$. Regardless of whether i went to college or not, there is an answer to the question of how much they would earn if they did go to college, and how much they would earn if they did not.

> Note that some authors use the notation Y_{1i} and Y_{0i}, or Y_i^1 and Y_i^0, for potential outcomes. I prefer the notation $Y_i(1)$ and $Y_i(0)$ in general because it makes the extension to a non-binary treatment variable look nicer. For example, if x is years of schooling, we can define potential outcomes $Y_i(x)$, e.g. earnings of individual i as a function of their years of schooling. See Section 1.7 for further discussion of non-binary treatments.

Consider for example a population composed of four individuals, pictured below. Person A would earn \$10 an hour if they didn't graduate college, but if they did go to college they would get a higher-paying job that paid them \$18 an hour at age 30. Person B is a higher earner, and would earn \$25 an hour without a college degree, and would earn \$40 and hour with one. Person C would choose to leave

the labor force and earn \$0 without a degree, but with a college degree would find a job that pays \$12 an hour. Notice that for all three of these individuals, $Y_i(1) > Y_i(0)$: the causal effect of college on their earnings is positive. But this not need be the case: suppose person D would found a successful company if they didn't go to college, earning them \$150 an hour by age 30, but if they did go to college they would have missed a chance opportunity to start the company and earned \$60 as an employee somewhere else.



Definition 1.2. *An individual's* ***treatment effect*** *is defined as:* $\Delta_i = Y_i(1) - Y_i(0)$*, the difference between their treated and untreated potential outcomes.*

In the above example, the treatment effects Δ_i are \$8 an hour for Person A, \$15 an hour for Person B, \$12 an hour for Person C, and $\$ - 90$ an hour for Person D. On average, treatment effects or positive—although Person D's individual treatment effect is negative. This example is thus a case of *heterogeneous* treatment effects.

Definition 1.3. *The phrase* ***"homogenous treatment effects"*** *describes a situation in which* $\Delta_i = Y_i(1) - Y_i(0)$ *is the same for all individuals* i

Definition 1.4. *The phrase* ***"heterogenous treatment effects"*** *describes a situation in which* $\Delta_i = Y_i(1) - Y_i(0)$ *differs across individuals* i

When treatment effects are heterogeneous, a useful summary of them is provided by the *average treatment effect*(ATE):

$$ATE = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{1.1}$$

The meaning of this quantity is discussed further in Section 1.3.

The important "leap of faith" that you need to take with potential outcomes is to believe that there exists a value $Y_i(0)$ *and* $Y_i(1)$ for each individual, regardless of whether they actually went to college. If i does graduate college (i.e $D_i = 1$), then their actual earnings Y_i, will be $Y_i = Y_i(1)$. Similarly, if they don't go to college, then their earnings will be $D_i = 0$. Another way of writing this is that, for each i:

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0) \tag{1.2}$$

Notice that since $D_i \in \{0, 1\}$, there is always one of the above terms that is equal to zero, and the other term gives us the appropriate potential outcome. In the above example, suppose that Persons

A and D do go to college and graduate, while B and C do not. Then if we measure the earnings and college-graduation status of each of the four individuals, our data will be $\{(Y_i, D_i)\}_{i=1,2,3,4} = \{(\$18, 1), (\$25, 0), (\$0, 0), (\$60, 0)\}$.

> An assumption implicit in Eq. (1.2) is that each individual's potential outcomes does not depend on whether *other* individuals go to college. This is known as the *stable unit treatment value assumption*, or SUTVA. This is not always a harmless assumption, as it rules out spillover effects.

1.2 The fundamental problem of causal inference

What can we say about *treatment effects*, given this data? Consider for example individual D, who in reality missed their opportunity to start the business and earn \$150 an hour. This is a *counterfactual*, something that would have happened if the world were different. Since we can't observe what would have happened, we'll never be able to answer the question of what person D's value of Δ_i is, empirically.

Definition 1.5. *The **fundamental problem of causal inference** is that for a given i, we only observe one of the two potential outcomes: either $Y_i(1)$ if $D_i = 1$, or $Y_i(0)$ if $D_i = 0$. In other words, we only observe i's **realized value** $Y_i = Y_i(D_i)$, and not their other potential outcome.*

The fundamental problem of causal inference means that we have a an "identification" problem. The concept of *identification* comes from statistics and is at the core of econometrics. Here I'll give a fairly informal definition:

Definition 1.6. *Given a set of assumptions about a population, we say that a quantity is **identified** when there is a unique value of that quantity that is compatible with the distribution of observable variables in that population.*

The idea of the distribution of a quantity in a population is reviewed in Appendix A. When discussing identification, it is common to refer to the quantity for which identification is being discussed as a "parameter of interest". See Appendix C for further discussion of identification.

Suppose for the moment that our population has a single individual i and our parameter of interest is their treatment effect $\Delta_i = Y_i(1) - Y_i(0)$. Our observable variables are Y_i and D_i. Suppose that i did graduate from college, so $D_i = 1$. Therefore the observed outcome Y_i is equal to $Y_i(1)$. By contrast, we do not observe $Y_i(0)$. Therefore, we must conclude that the quantity Δ_i is not identified. Individual treatment effects Δ_i are always unidentified, due to the fundamental problem of causal inference.

Nevertheless, we'll see that we can still sometimes make statements about average treatment effects, by using *other* students who didn't go to college as a comparison group. A key result that we will see below, is that the ATE is identified when D_i is randomly assigned.

1.3 Potential outcomes as random variables

Recall the definition of the average treatment effect given in Eq. (1.1)

$$ATE = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$$

In applying the expectation operator to Δ_i, we are here invoking the idea of a probability distribution over treatment effects (see Appendix A.4 for the definition of expectation). You can think of this distribution as the one that arises from drawing an individual at random from the population and observing their value of Δ_i. When the relevant population is a finite collection I of N individuals, the expectation defined in this way coincides with a simple arithmetic mean of Δ_i over all i in I:

$$ATE = \frac{1}{N} \sum_{i \in I} \Delta_i$$

It is thus natural to refer to $ATE = \mathbb{E}[\Delta_i]$ as the "population mean" of Δ_i (see Appendix B.1 for a discussion of random variables defined by sampling from a population).

We thus view $Y_i(1)$ and $Y_i(0)$ as *random variables*. To view them as defined in a common probability space, we need to make the "leap of faith" described above: each individual i has a value both of $Y_i(1)$ and a value of $Y_i(0)$ (if they did not, we could not define Δ_i or associate any meaning to its expectation).

Since the expectation operator is linear, we can rewrite the average treatment effect as

$$ATE = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

showing that the ATE is equal to the difference between the mean of $Y_i(1)$ and the mean of $Y_i(0)$ in the population. Note that the expectation of $Y_i(0)$ depends only on its *marginal* distribution: we don't need to know e.g. whether $Y_i(0)$ and $Y_i(1)$ are correlated with one another to compute $\mathbb{E}[Y_i(0)]$. Similarly, $\mathbb{E}[Y_i(1)]$ depends only on the marginal distribution of $Y_i(1)$, and thus the ATE depends only on the marginal distributions of $Y_i(0)$ and $Y_i(1)$. However, average treatment effect is not the *only* thing we can say about causality given access to marginal distributions of potential outcomes: see Section 1.8 for discussion.

The random variables $Y_i(1)$ and $Y_i(0)$ do have a joint distribution in the population, and it's unlikely that they would be statistically independent of one another. In particular, it's natural to expect individuals with higher $Y_i(1)$ to also have higher $Y_i(0)$. Because of the fundamental problem of causal inference though, there is not a whole lot we can say about the joint distribution of potential outcomes by looking at observable data. Fortunately, averages of treatment effects like the ATE typically only depend upon the marginal distributions, so we can have a hope of identifying these kinds of parameters.

> Note that although we can't observe the joint distribution of $Y_i(1)$ and $Y_i(0)$ directly, we can make assumptions about it and in some cases test these assumptions. For instance, if we assume that treatment effects are homogenous, this implies a perfect positive relationship between $Y_i(0)$ and $Y_i(1)$, since then $Y_i(1) = Y_i(0) + \Delta$. Although we cannot check whether treatment effect homogeneity holds directly (because we never see both potential outcomes for the same individual), homogenous treatment effects does make a prediction about the marginal distributions of $Y_i(0)$ and $Y_i(1)$ (in particular, that the CDF of $Y_i(0)$ is a horizontal shift of the CDF of $Y_i(1)$. Thus, there do exist statistical tests for treatment effect heterogeneity (see e.g. Heckman et al. (1997) for a discussion).

Note that the realized value of Y_i and treatment status D_i are also random variables. The goal of causal inference is to use the joint distribution of Y_i and D_i, which is observable, to learn something about the distribution of treatment effects Δ_i.

1.4 A naive comparison of means suffers from selection bias

A natural instinct is to compare the average value of the outcome variable among the "treatment group" $D_i = 1$ and "control group" $D_i = 0$. Let θ_{DM} denote this difference in means:

$$\theta_{DM} := \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

Suppose that we can observe that $\mathbb{E}[Y_i|D_i = 1] \geq \mathbb{E}[Y_i|D_i = 0]$. Can we conclude from our data that going to college causes ones earnings at age 30 to be higher?

> *Review:* A quantity like $\mathbb{E}[Y_i|D_i = 1]$ can be interpreted as the average value of Y_i among all individuals i in a population for which $D_i = 1$. See Appendix Section A.5 for a formal definition of the *conditional expectation*.

We know from Equation (1.2) that for any individual for whom $D_i = 1$, our observed Y_i is $Y_i = Y_i(1)$. Similarly for any individual who doesn't go to college, $Y_i = Y_i(0)$. Thus, we can rewrite the estimand of our difference-in-means estimator as:

$$\theta_{DM} = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \tag{1.3}$$

Notice that the first term in Eq. (1.3) conditions on the event $D_i = 1$, and the second term conditions on $D_i = 0$. This means that the difference-in-means estimand compares two different groups, which might

not be comparable to one another. For example, students who go to college might have higher "ability" (e.g. as measured by test scores) than students who do not. This might be why wages are higher among college graduates, rather than it being from a causal effect of college on earnings.

Suppose for the moment that the second term in Eq. (1.3) had also conditioned on the event $D_i = 1$, rather than on $D_i = 0$. If this were the case, then we could use linearity of the expectation to rewrite θ_{DM} as being equal to $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$, the average treatment effect Δ_i among students who do go to college. We call this the *average treatment effect on the treated*, or ATT. The ATT is a causal parameter, because it compares the values of $Y_i(1)$ and $Y_i(0)$, on average, for the *same group*.

Back to the general setting. Note that by adding and subtracting $\mathbb{E}[Y_i(0)|D_i = 1]$ to equation (1.3), we can write:

$$\theta_{DM} = \underbrace{\{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]\}}_{ATT} + \underbrace{\{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]\}}_{\text{selection bias}} \tag{1.4}$$

The parameter ATT is not identified, unless the selection bias term $\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]$ is equal to zero (or more generally, has a known value). This term represents a measure of non-comparability between the students who go to college and the students who do not, in terms of their counterfactual earnings $Y_i(0)$.

For example, students who obtain a college degree may be more likely to come from family backgrounds in which their parent(s) had time and resources to help the student accumulate skills that are valued by the labor market. As a result, these students would have earned more on average, even if they didn't go to college and hence $\mathbb{E}[Y_i(0)|D_i = 1] > \mathbb{E}[Y_i(0)|D_i = 0]$. Many other stories also lead to a positive correlation between D_i and $Y_i(0)$: students whose parents are well-connected may be more likely to go to college, and earn more even if they didn't go to college, and any genetic traits that are associated with higher earnings are likely to also increase college attendance.

1.5 Randomization eliminates selection bias

A sufficient condition for the selection bias term to be zero is that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$. Often a condition like this is refered to as $Y_i(0)$ being *mean-independent* of D_i. One case in which this will hold is when D_i is assigned completely randomly, as in a randomized controlled trial.

Definition 1.7. ***Random assignment*** *says that* $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$

Random assignment says that treatment D_i is unrelated to potential outcomes, in a statistical sense. For example, individuals who would earn more even without a college degree are no more or less likely to go to college.

The random assignment assumption is stronger than we need to kill the selection bias term in Equation (1.4). All we need for that is $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$. This is implied by random assignment, because $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$ implies that $Y_i(0) \perp\!\!\!\perp D_i$, which in turn implies that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$.

Exercise: show that for any random variable V_i, if $V_i \perp\!\!\!\perp D_i$ and D_i is binary, then $\mathbb{E}[V_i|D_i = 1] = \mathbb{E}[V_i|D_i = 0] = \mathbb{E}[V_i]$.

Review: what is statistical independence?

Definition 1.8. *We say that random variables A and B are independent if $F_{AB}(a, b) = F_A(a) \cdot F_B(b)$ for all a and b.*

This definition extends naturally to cases where A or B is a random vector, rather than a scalar random variable. In the case of 1.7, we can let $A = (Y_i(1), Y_i(0))$ and $B = D_i$.

When A and B are independent, we denote this fact as $A \perp\!\!\!\perp B$. When they are not, we say $A \not\perp\!\!\!\perp B$. See Appendix A.3.3 for more details.

When the selection-bias term in Equation (1.4) is equal to zero, the ATT is identified. There is only one value of ATT compatible with the population distribution of observables, since (Y_i, D_i) is observed

and $ATT = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 1]$. However, under random-assignment we can actually say more. Not only is the ATT identified, but so is the *average treatment effect*:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)] = P(D_i = 1) \cdot ATT + (1 - P(D_i = 1)) \cdot ATU$$

where $ATU := \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 0]$ is the average treatment effect on the untreated, and we've used the law of iterated expectations to decompose ATE into the ATT and ATU. Since the random-assignment assumption says that treated potential outcomes $Y_i(1)$ are also independent of treatment D_i, we have not only that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$, but also that $\mathbb{E}[Y_i(1)|D_i = 1] = \mathbb{E}[Y_i(1)|D_i = 0]$, and thus the ATT, ATU, and ATE are all equal to one another.

In non-experimental settings, one may be able to identify a parameter like the ATT without being able to identify the ATE. An example of this is the difference-in-differences research design, which (in its basic, most common form) only yields identification of the ATT and not the ATU or ATE.

> *Note:* Even the above argument that $ATT = ATU = ATE = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$ only ever makes use of $Y_i(0)$ being independent of D_i, and $Y_i(1)$ being independent of D_i. This is still weaker than the assumption made above, that $Y_i(0)$ and $Y_i(1)$ are *jointly* independent of D_i. In practice, it's usually hard to come up for an argument for why only the marginal distributions of $Y_i(1)$ and $Y_i(0)$ would be independent of D_i, and not their joint distribution, which is why I've written it the way I have.

> *Note:* Definition 1.7 corresponds to a randomized controlled trial with *perfect compliance*. In many real-world trials, the only thing that can be randomized is whether an individual is *assigned* to receive treatment. But subjects may still choose whether to actually receive treatment. In these cases, one can use the method of *instrumental variables* to estimate causal effects, which you'll see later in the course.

1.6 Simple linear regression with a binary treatment

We've seen in the last section that when we have random assignment the difference in means θ_{DM} uncovers the average treatment effect (ATE):

$$\underbrace{\mathbb{E}[Y_i|D_i = 1] = \mathbb{E}[Y_i|D_i = 0]}_{\theta_{DM}} = \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{ATE}$$

In the next chapter, we will see that regression provides a tool to extend this result to a setting in which we use control variables to isolate random variation in D_i, an approach known as *selection-on-observables*.

However, we can use linear regression even with random assignment, which represents the simplest and most idealized case. To see this, consider the regression model:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \tag{1.5}$$

Recall (see Appendix **??**) that the OLS parameter β_1 is given by the formula

$$\beta_1 = \frac{Cov(Y_i, D_i)}{Var(D_i)}$$

This can be used to show that β_1 in Eq. (1.5) is identical to the difference in means θ_{DM} introduced in the preceeding sections. The proof is left as an exercise.

Exercise: Show that with a binary D, we have that $\frac{Cov(Y_i,D_i)}{Var(D_i)} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

This in turn implies that β_1 from a simple linear regression of Y on D uncovers the ATE, when we have random assignment of D.

In the special case of homogenous treatment effects ($\Delta_i = \Delta$ for all i), we can see the equivalence between β_1 and the ATE directly in terms of Eq. (1.5). Note that in this case $ATE = \Delta$, so we wish to show that $\beta_1 = \Delta$.

$$\begin{aligned} Y_i &= Y_i(D_i) = Y_i(0) + \Delta_i \cdot D_i \\ &= Y_i(0) + \Delta \cdot D_i \\ &= \underbrace{\mathbb{E}[Y_i(0)]}_{\beta_0} + \underbrace{\Delta}_{\beta_1} \cdot D_i + \underbrace{Y_i(0) - \mathbb{E}[Y_i(0)]}_{\epsilon_i} \\ &= \beta_0 + \beta_1 D_i + \epsilon_i \end{aligned}$$

where we simply define the regression residual ϵ_i for individual i to be the difference between $Y_i(0)$ and it's average in the population $\mathbb{E}[Y_i(0)]$. Note that it is important that treatment effects be homogenous for us to replace Δ_i with Δ in the above. Nevertheless, a simple linear regression of Y on D contiues to identify the ATE even if treatment effects are heterogenous.

In this simple setting, the assumption of homogenous treatment effects is exactly equivalent to the "structural" interpretation of linear regression that you may have learned as an undergraduate, that we interperet Eq. (1.5) as a story about how the world works: e.g. wages equal β_0, plus an effect of college β_1, plus an idiosyncratic error term. This amounts to assuming that potential outcomes take the form $Y_i(d) = \beta_0 + \beta_1 \cdot d + \epsilon_i$ for $d \in \{0, 1\}$ and by random assignment $\mathbb{E}[\epsilon_i | D_i = d]$ does not depend on d (without loss of generality we can now take $\mathbb{E}[\epsilon_i | D_i] = 0$ since a non-zero value for this expectation could be absorbed into β_0). Then, we have that the realized value of Y_i

$$Y_i = Y_i(D_i) = \beta_0 + \beta_1 \cdot D_i + \epsilon_i,$$

recovering (1.5).

1.7 Causality beyond a binary treatment*

In this chapter we've focused on a *binary* treatment, which takes just two values: $D_i = 1$ ("treatment"), and $D_i = 0$ ("control"). However, we're often interested in the causal effect of a treatment variable that takes on many values. For example, what is the effect of *years* of schooling on earnings, rather than just the effect of completing any college degree?

Setting up the notation for multivalued treatment variables is pretty straightforward. We can define our potential outcomes $Y_i(d)$ in the same way as before, where now d index all of the values that D_i might take. Here are some examples:

- Let d be the number of years of schooling student i completes, and $Y_i(d)$ be their earnings at age 30.
- Let d be the price of some good, and let the function $Y_i(d)$ be the demand function for that good in market i.
- Let d be the high school that student i attends, and let $Y_i(d)$ be an indicator for whether they were accepted to UCalgary, e.g. $d \in \{\text{school A}, \text{school B}, \text{school C}, \textit{etc.}\}$.
- In a randomized experiment about the effect of social media on mental health, subjects i are assigned to three different treatments:

$$d \in \{\text{no social media}, \text{Facebook only}, \text{Twitter only}, \text{Facebook and Twitter}\}$$

Regardless of the setting, we can still define random assignment and selection-on-observables exactly as we did before, we just now have to phrase it in terms of the vector of all $Y_i(d)$ for all treatment values d instead of just the two potential outcomes $(Y_i(1), Y_i(0)$.

However, with more than two values of treatment, there are now many different ways to think about treatment effects. For example, in the first example above, we can think about the effect of finishing grade 12 as:

$$Y_i(12) - Y_i(11),$$

while the effect of completing high-school versus dropping out after grade 10 is:

$$Y_i(12) - Y_i(10)$$

The overall average causal effect of the last year of schooling that each student actually completes would be

$$\mathbb{E}[Y_i(D_i) - Y_i(D_i - 1)]$$

In the first two examples above, the values of treatment D_i have a natural order to them. In the third and fourth examples, treatment is categorical, and there may not be a natural such order. With an unordered treatment, like in the last example, we might pick one comparison category and consider treatment effects with respect to it, e.g. separately estimating $\mathbb{E}[Y_i(\text{Facebook only}) - Y_i(\text{no social media})]$, $\mathbb{E}[Y_i(\text{Twitter only}) - Y_i(\text{no social media})]$ and $\mathbb{E}[Y_i(\text{Facebook and Twitter}) - Y_i(\text{no social media})]$.

1.8 Moving beyond average treatment effects*

Although our discussion here has been focused on parameters that *average* over treatment effects $\Delta_i = Y_i(1) - Y_i(0)$, this isn't the only type of causal question that we can answer with random-assignment.

Consider a binary treatment D_i and random assignment: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i$. Note that we can apply any function $g(\cdot)$ to the potential outcomes, without destroying independence, i.e. $(g(Y_i(0)), g(Y_i(1))) \perp\!\!\!\perp D_i$. Why is this useful? Consider the function $g(t) = \mathbb{1}(t \leq y)$ for some value y. Given that random-assignment implies that the random variable $\mathbb{1}(Y_i(1) \leq y)$ is independent of D_i, we have that

$$\underbrace{\mathbb{E}[\mathbb{1}(Y_i \leq y)|D_i = 1]}_{F_{Y|D=1}(y)} = \mathbb{E}[\mathbb{1}(Y_i(1) \leq y)|D_i = 1] = \underbrace{\mathbb{E}[\mathbb{1}(Y_i(1) \leq y)]}_{F_{Y(1)}(y)}$$

The term on the left is the conditional CDF of Y_i given $D_i = 1$, which can be computed from the data. The term on the right is the (unconditional) CDF of the treated potential outcome $Y_i(1)$. This expression shows that we can identify the CDF of $Y_i(1)$ at any point y. Collecting over all y, we can thus compute the entire distribution of $Y_i(1)$.

By the same logic, we can also identify the entire distribution of $Y_i(0)$, using $F_{Y|D=1}(y) = \mathbb{E}[\mathbb{1}(Y_i \leq y)|D_i = 1]$. That means that we can use random-assignment to uncover the effect of treatment on the entire *distribution* of outcomes. This lets us answer a new set of causal questions. For instance: what is the difference between the median value of $Y_i(1)$ and the median value of $Y_i(0)$? This is an example of a so-called *quantile-treatment effect*.

A natural question that you might hope to answer is: how many individuals in my population have a negative treatment effect $Y_i(1) < Y_i(0)$, versus a positive one? This is a harder type of question, because it depends on the joint distribution of potential outcomes. By contrast, random assignment (and similarly selection-on-observables, or quasi-experimental approaches), only let us identify each of the *marginal* distributions of $Y_i(0)$ and $Y_i(1)$, due to the fundamental problem of causal inference.

The situation is not completely hopeless: the marginal distributions of $Y_i(1)$ and $Y_i(0)$ do put some restrictions on the distribution of treatment effects. For instance, it can be shown that a lower bound on the proportion "harmed" by treatment $P(\Delta_i \leq 0)$ is the supremum of $F_{Y(1)}(y) - F_{Y(0)}(y)$ over all values of y (see e.g. Fan and Park, 2010 for details). We can also make additional assumptions that allow us to say more about the distribution of treatment effects. For example, the strong assumption of *rank-invariance* allows us to trace out the entire CDF of Δ_i, and in principle estimate the treatment effect for any given individual (see e.g. Heckman et al., 1997).

Chapter 2

Selection on observables

Outside of an actual experimental setting, the random-assignment assumption is very strong. Typically, economic agents "select into" treatment, meaning they choose for themselves whether or not $D_i = 1$ or $D_i = 0$. There are usually a variety of reasons why the circumstances and preferences that lead to a choice of taking treatment can be expected to be correlated with potential outcomes.

2.1 The selection-on-observables assumption

Suppose we observe a vector of covariates X_i, along with Y_i and D_i. Then, the following assumption is often considered to be weaker than assuming fully random-assignment:

Definition 2.1. ***Selection-on-observables****, also referred to as* ***unconfoundedness****, says that*

$$\{(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i\}|X_i$$

Selection-on-observables makes the same assumption as random assignment, but we assume it holds conditional on each value X_i. It is thus often also called a *conditional independence assumption*.

Review: what is *conditional* statistical independence?

Definition 2.2 (conditional independence). *We say that A and B are independent conditional on C, denoted $(A \perp\!\!\!\perp B)|C$, if for any values a, b, c of A, B, C: $F_{AB|C=c}(a, b) = F_{A|C=c}(a) \cdot F_{B|C=c}(b)$.*

In this definition, A, B and C can all be random vectors, each having multiple components. See Appendix A.6.2.2 for more details.

Analagous to the random-assignment case, conditional independence will be useful for us because it allows us to remove conditioning on a random variable inside of an expectation. In this case, the relevant property is that if $(A \perp\!\!\!\perp B)|C$, then for any b and c:

$$\mathbb{E}[A|B = b, C = c] = \mathbb{E}[A|C = c]$$

A nice example of a setting in which selection-on-observables is very credible is provided by Washington (2008), who looks at the effect of having daughters on the feminist sympathies of legislators in the U.S. House of Representatives. While the sex (assigned at birth) of a given child is random, whether or not a legislator has a daughter is correlated with political views because conservatives tend to have more childen. Let $Y_i(0)$ indicate feminist sympathies if legislator i does not have a daugher (measured by a score, see paper), and $Y_i(1)$ their feminist sympathies if they do. While $D_i \not\perp\!\!\!\perp Y_i(0)$, selection-on-observables is very plausible if we condition on number of children i has overall, X_i. For another clever and compelling example of using selection-on-observables, I recommend looking at Dale and Krueger (2002).

How does the selection-on-observables assumption help us? Note that if it holds then

$$\begin{aligned}\mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0] &= \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, D_i = 0] \\ &= \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x] \\ &= \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] := ATE(x) \qquad (2.1)\end{aligned}$$

Thus, the average treatment effect, conditional on X is identified by a version of the difference in means estimand that conditions on any given value x of X_i. Let us denote this parameter as $ATE(x)$. Equation 2.1 shows that under selection-on-observables, it is identified. Since we also observe the marginal distribution of X_i, we can then recover for example the overall average treatment effect by averaging over values of the control variables:

$$ATE = \int \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] \cdot dF(x) = \int ATE(x) \cdot dF(x)$$

which follows by the law of iterated expectations.

There are three main approaches to making use of the selection-on-observables assumption in this way: *inverse-propensity score weighting*, *matching*, and *regression*. In this class, we'll focus on the third of these, regression, but I briefly introduce the other two at the end of this section. The three approaches can be thought of as essentially three different strategies to construct an estimator for $ATE(x)$, but are all fundamentally based off of the identification result (2.1).

Recall that the difference-in-means θ_{DM} from Section 1.4 is equal to the ATE under random assignment. However, one can not simply estimate θ_{DM} to get the ATE under selection-on-observables. Conditioning on X_i is necessary, which the three methods above all accomplish in various ways. The excersize below shows that θ_{DM} instead estimates the average treatment on the treated ATT, plus another term that depends on the correlation between D_i and X_i.

Exercise: Show that θ_{DM} does not condition on X_i does generally estimate the ATE under selection-on-observables.

Solution: Suppose for simplicity that X_i is discrete. Then, by LIE:

$$\begin{aligned}\theta_{DM} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\ &= \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\ &= \sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1)|X_i = x, D_i = 1] - \sum_x P(X_i = x|D_i = 0) \cdot \mathbb{E}[Y_i(0)|X_i = x, D_i = 0] \\ &= \sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1)|X_i = x] - \sum_x P(X_i = x|D_i = 0) \cdot \mathbb{E}[Y_i(0)|X_i = x] \\ &= \underbrace{\sum_x P(X_i = x|D_i = 1) \cdot \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]}_{\text{=ATT, by law of iterated expectations}} + \sum_x \{P(X_i = x|D_i = 1) - P(X_i = x|D_i = 0)\} \cdot \mathbb{E}[Y_i(0)|X_i = x]\end{aligned}$$

Question: Under selection on observables, is it true that $\mathbb{E}[Y_i|X_i = x, D_i = d] = \mathbb{E}[Y_i|X_i = x]$?
Answer: No! While it is true that $\mathbb{E}[Y_i(d)|X_i = x, D_i = d] = \mathbb{E}[Y_i(d)|X_i = x]$ for any $d \in \{0, 1\}$, the same cannot be said about $\mathbb{E}[Y_i|X_i = x, D_i = d]$ in general. The reason is that we have *not* assumed that $\{Y_i \perp\!\!\!\perp D_i\}|X_i$, but rather $\{(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i\}|X_i$. The former of these would mean that realized outcomes are conditionally indepedent of treatment. The latter says that *potential* outcomes are.

Note: When using the selection-on-observables assumption, it is important that the variables in the vector X_i are *unaffected* by treatment. That is, if we introduced potential outcomes $X_i(0)$ and $X_i(1)$, we would have $X_i(0) = X_i(1)$ for all i. To make sure of this, researchers typically consider variables X_i that are measured earlier in time than treatment D_i is assigned. When this condition fails, causal inference can fail even when the selection-on-observables assumption holds, via a problem often referred to as "bad-control".

Why might selection-on-observables be more reasonable than random assignment, in general? The basic idea is that if we observe a rich enough set of X_i, we might be able to control for confounding factors that lead to selection bias. For example, in the returns-to-college example,

we might include in the vector X_i whether ot not i's parents graduated from college, their socio-economic status, and i's test scores in high school.

Imagine that we observed literally *everything* that matters for determining the outcome Y_i, in addition to treatment. In this case, we could write potential outcomes as

$$Y_i(0) = Y(0, X_i) \quad \text{and} \quad Y_i(0) = Y(0, X_i),$$

where the function $Y(d, x)$ is common to everybody: once we know d and x we can say exactly what is going to happen to you. Then selection-on-observables would be satisfied automatically, since if we condition on $X_i = x$, then $Y_i(d) = Y(d, x)$ for either $d \in \{0, 1\}$. Notice that $Y(d, x)$ doesn't depend on i: it is no longer random once we've fixed X_i. It is hence is uncorrelated with D_i, since degenerate random variables are statistcally independent of everything! This can be seen as mimicking the logic of a carefully controlled experiment in the natural sciences, in which we make sure "everything else" that matters X_i is held fixed, while varying D_i between 0 and 1.

A similar logic would apply if X_i includes everything that determines D_i: e.g. $D_i = d(X_i)$ for some function d. Then we'd also get selection-on-observables for free. In practice, apart from vey specific settings, we'll never observe everything that determines outcomes Y_i, or selection into treatment D_i. However, if we can control for most of obvious threats to eliminating selection bias, we might be willing to think that our X_i get us most of the way there. However, adding many X_i can also do more harm than good, because of the possibility of bad controls (see above) and the curse of dimensionality hampering estimation (see Appendix C.3.1.2).

2.2 How to use selection-on-observables

2.2.1 Matching*

The approach of *matching* is probably the most intuitive application of selection-on-observables assumption. It simply attempts to find, for each treated unit ($D_i = 1$), an untreated unit ($D_i = 0$) *with the same value* of X_i. For this it is of course necessary that for each value of X_i, there are both treated and untreated units, a condition oftern referred to as *overlap* or *common support*: $0 < P(D_i = 1|X_i = x) < 1$ *for all* x.

Suppose for the moment that X_i is a discrete variable, so that it's possible to find pairs of observations that have identical X_i. In the most basic version of matching (*one-to-one, exact* matching), we would for each treated unit i find a control unit i' such that $X_i = X_{i'}$. We drop any control units that are not matched, and then calculate the difference in means between treatment and control in this modified sample. When X is a vector with many components, finding pairs such that $X_i = X_{i'}$ can become difficult. If X includes any components that are continuously distributed, it becomes impossible. In these cases we'd need to settle for finding an i' such that $X_i \approx X_{i'}$.

However, a clever application of the selection-on-observables assumption (Rosenbaum and Rubin, 1983) allows us to simplify the problem considerably, leading to the most popular implementation of matching: called *propensity-score matching*. They observe that selection-on-observables implies that for any $p \in (0, 1)$:

$$\mathbb{E}[Y_i|D_i = 1, \mathcal{P}(X_i) = p] - \mathbb{E}[Y_i|D_i = 0, \mathcal{P}(X_i) = p] = \mathbb{E}[Y_i(1) - Y_i(0)|P(X_i) = p]$$

where $\mathcal{P}(x) = P(D_i = 1|X_i = x)$ is called the *propensity score function*. This expression says that conditioning on values of the *propensity score* rather than on X_i itself is sufficient to estimate causal effects. This is useful because while X_i may have many components, the propensity score is always a scalar. Thus, we simply need to estimate the function $\mathcal{P}(x)$, and then match units i and i' such that $\mathcal{P}(X_i) \approx \mathcal{P}(X_{i'})$, rather than finding a good way to compare X on all dimensions.

A good exercise in the law of iterated expectations to verify the above expression. Begin by

noticing that

$$\mathbb{E}[Y_i|D_i = d, \mathcal{P}(X_i) = p] = \mathbb{E}\left\{\mathbb{E}[Y_i|D_i = d, X_i]|\, D_i = d, \mathcal{P}(X_i) = p\right\}$$

or any $d \in \{0, 1\}$. This expression follows from the law of iterated expectations, where the outer average is over the distribution of X_i such that $P(X_i) = p$. If for example X_i is continuously distributed, then this outer average is an integral over the conditional density $f_{X|\mathcal{P}(X)=p}(x) = \frac{f(x)}{\int_{x:\mathcal{P}(x)=p} f(x)dx}$ if $\mathcal{P}(x) = p$, and is equal to zero otherwise. Note that we've been able to remove the conditioning on $\mathcal{P}(X_i) = p$ in the inner expectation, because once X_i is fixed, $\mathcal{P}(X_i)$ is as well.

Using the standard selection-on-observables argument, we know that for any x, $\mathbb{E}[Y_i|D_i = d, X_i = x] = \mathbb{E}[Y_i(d)|X_i = x]$. Thus:

$$\mathbb{E}[Y_i|D_i = d, \mathcal{P}(X_i) = p] = \mathbb{E}\left\{\mathbb{E}[Y_i(d)|X_i]|\, D_i = d, \mathcal{P}(X_i) = p\right\}$$

We now use the fact that $\{X_i \perp\!\!\!\perp D_i\}|\mathcal{P}(X_i)$ to remove the conditioning on $D_i = d$ on the outer expectation above. Once we've done this, we can apply the law of iterated expectations again (this time in the reverse direction), this is equal to $\mathbb{E}[Y_i(d)|\mathcal{P}(X_i) = p]$ and we're done!

To see that $\{X_i \perp\!\!\!\perp D_i\}|\mathcal{P}(X_i) = p$, observe that $P(D_i = 1|X_i = x, \mathcal{P}(X_i) = p) = P(D_i = 1|X_i = x) = \mathcal{P}(x)$. The event $X_i = x, \mathcal{P}(X_i) = p$ is only possible if the value x is such that $\mathcal{P}(x) = p$ (otherwise the conditional probability $P(D_i = 1|X_i = x, \mathcal{P}(X_i) = p) = P(D_i = 1|X_i = x)$ is undefined, and we don't need to worry about this x). Therefore, we've shown that $P(D_i = 1|X_i = x, \mathcal{P}(X_i) = p) = \mathcal{P}(x) = p$ for any such x, which does not depend on the precise value of x. This is equivalent to saying X_i and D_i are independent, conditional on $\mathcal{P}(X_i) = p$.

2.2.2 Inverse propensity score weighting*

Under selection-on-observables, one can also show that:

$$ATE = \mathbb{E}\left[\frac{D_i \cdot Y_i}{\mathcal{P}(X_i)} - \frac{(1 - D_i) \cdot Y_i}{1 - \mathcal{P}(X_i)}\right]$$

Estimating the ATE through the above expression is known as inverse propensity score weighting. Note that this first requires estimating the propensity score function $\mathcal{P}(x) = P(D_i = 1|X_i = x)$ for all values of x. Then one can form a sample estimator of the expectation above, using $\mathcal{P}(X_i)$ for each observation. Inverse propensity score weighting is a lot like propensity score matching, but doesn't have a step in which we need to actually pair up treatment/control observations by matching their values of the propensity score. You can think of the pairing as happening automatically "under-the-hood".

Why then does propensity score weighting work? You guessed it, we're going to use the law of iterated expectations to show that it does. Note that by the law of iterated expectations and selection-on-

observables:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i=1|X_i)} - \frac{(1-D_i)\cdot Y_i}{1-P(D_i=1|X_i)}\right] \\
&\quad = \mathbb{E}\left\{\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i=1|X_i)} - \frac{(1-D_i)\cdot Y_i}{1-P(D_i=1|X_i)}\middle| X_i\right]\right\} \\
&\quad = \int \left\{\mathbb{E}\left[\frac{D_i \cdot Y_i}{P(D_i=1|X_i)} - \frac{(1-D_i)\cdot Y_i}{1-P(D_i=1|X_i)}\middle| X_i = x\right]\right\} \cdot dF_X(x) \\
&\quad = \int \left\{\frac{\mathbb{E}[D_i \cdot Y_i|X_i=x]}{P(D_i=1|X_i=x)} - \frac{\mathbb{E}[(1-D_i)\cdot Y_i|X_i=x]}{1-P(D_i=1|X_i=x)}\right\} \cdot dF_X(x) \\
&\quad = \int \left\{\frac{\mathbb{E}[D_i \cdot Y_i(1)|X_i=x]}{P(D_i=1|X_i=x)} - \frac{\mathbb{E}[(1-D_i)\cdot Y_i(0)|X_i=x]}{1-P(D_i=1|X_i=x)}\right\} \cdot dF_X(x) \\
&\quad = \int \left\{\frac{\cancel{P(D_i=1|X_i=x)}\cdot \mathbb{E}[Y_i(1)|D_i=1,X_i=x]}{\cancel{P(D_i=1|X_i=x)}}\right. \\
&\qquad\qquad \left. - \frac{\cancel{(1-P(D_i=1|X_i=x))}\cdot \mathbb{E}[Y_i(0)|D_i=0,X_i=x]}{\cancel{1-P(D_i=1|X_i=x)}}\right\} \cdot dF_X(x) \\
&\quad = \int \{\mathbb{E}[Y_i(1)|X_i=x] - \mathbb{E}[Y_i(0)|X_i=x]\} \cdot dF_X(x) \\
&\quad = \mathbb{E}[Y_i(1) - Y_i(0)] = ATE
\end{aligned}
$$

2.2.3 Linear regression

In practice, the most common technique that implements a selection-on-observables identification strategy is *linear regression*. In it's most basic form, linear regression is just a simple way of estimating conditional expectation functions, when they are linear functions of the things being conditioned on.

Recall that under the selection-on-observables assumption, and with a binary treatment variable, the average treatment effect conditional on $X = x$ can be calculated as:

$$ATE(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i|X_i = x, D_i = 1] - \mathbb{E}[Y_i|X_i = x, D_i = 0]$$

This requires having a way to estimate conditional expectations of the form $\mathbb{E}[Y_i|X_i = x, D_i = d]$ for $d = 0$ and $d = 1$, given a sample of data. How should we do this?

If X is a discrete random variable, there is a pretty straightforward way we could do this. With i.i.d. data, a consistent estimator is simply the mean among the sub-sample of data for which $D_i = d$ and $X_i = x$:

$$\underbrace{\frac{1}{\#\text{ of observations } i \text{ for which } X_i = x \text{ and } D_i = d}\sum_{i:X_i=x\&D_i=d} Y_i}_{\hat{\mathbb{E}}[Y|X=x,D=d]} \xrightarrow{p} \mathbb{E}[Y|X = x, D = d]$$

For a view of the concept of a consistent estimator, see Appendix C.3.1.1.

But remember that for the selection-on-observables assumption, we want X to be an extensive-enough set of control variables to eliminate selection bias. So how should we proceed if $X = (X_1, X_2, \ldots X_k)$ is a vector of several random variables, some of which might also be continuously distributed?

This is actually a hard problem, in general. Recall that $\mathbb{E}[Y_i|X_i = x, D_i = d]$ is a function of x and d, which (in the notation of A.5.3) we might call m:

$$\mathbb{E}[Y_i|X_i = x, D_i = d] = m(d, x_1, x_2, \ldots x_k)$$

where $x_1, x_2, \ldots x_k$ are the components of the vector x. Provided that (Y, D, X) are all observed in a random sample, the function m is *identified*. That is, for fixed values (x, d) there is only one value of $m(d, x_1, x_2, \ldots x_k)$ compatible with the joint distribution of our observables. Once we know the function m, we can calculate treatment effects easily since

$$ATE(x) = m(1, x_1, \ldots x_k) - m(0, x_1, \ldots x_k)$$

However, estimation is another thing. Given our finite sample, how do we uncover the function $m(d, x_1, x_2, \ldots x_k)$? This turns out to be particularly straightforward when the function m is *linear*, that is:

$$m(d, x_1, x_2, \ldots x_k) = \beta_0 + \beta_D d + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2.2}$$

for some set of coefficients $(\beta_D, \beta_0, \beta_1, \ldots \beta_k)$. In this case note that $ATE(x) = m(1,x) - m(0,x) = \beta_D$ for all x. Since this difference yields the same fixed number β_D regardless of x, the conditional-on-X ATE is the same as the overall average treatment effect, so $ATE = \beta_D$.

Assuming that the conditional expectation be linear in both d and x as in Eq. (2.2) avoids the need to determine the functional form of function m from the data, which is probably why using linear regression to "control" for X_i is so common in practice. This bypasses the so-called "curse of dimensionality" which makes estimation difficult when we have several X (see Sec C.3.1.2). However, linearity of the CEF is often a strong assumption, and a good practice is to consider the possibilities of nonlinearities and interactions among the X_i and between D_i and X_i (see box below).

Angrist (1998) shows that if $\mathbb{E}[D_i|X_i]$ is linear in the X_i, then estimating the regression equation $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{1i} + \beta_2 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$ uncovers a weighted average of the $ATE(x)$ across covariate cells, even if 2.2 does not hold. This is perhaps surprising, because the regression equation being estimated does not include any interactions between D_i and X_i to model heterogeneity in treatment effects by X_i. A version of this result extends to a single ordered treatment like years of schooling (Angrist and Krueger, 1999). However, it does not work in general for multi-valued or multiple treatments (see Goldsmith-Pinkham et al. 2021). We return to these issues in Section 3.5.

Chapter 3

Review of linear regression

This section represents an abridged version of Chapter 7 from my notes *Probability and Statistics for Econoemtrics*, which you can find on my website.

3.1 Review of statistical concepts

This section briefly reviews some of the statistical concepts that we'll be using as we review linear regression. This is an abridged version of the material presented in Appendices B and C, you might consult them for further detail.

3.1.1 Samples and estimators

We'll use the terms *dataset* or *sample* to refer to a collection of characteristics $X_i = (X_{1i}, X_{2i}, \dots X_{ki})$ for each of n observational units (such as individuals) i. These observations can be arranged into an $n \times k$ matrix $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} (X_{11}, X_{21}, \dots X_{k1}) \\ (X_{12}, X_{22}, \dots X_{k2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{kn}) \end{pmatrix}$$

Notation: Note that the entries of the sample matrix $\mathbf{X}$ are denoted X_{ji}, where i index rows (individual observations) and j index columns (variables/characteristics). This is backwards from the way we often denote entries M_{ij} of a matrix $\mathbf{M}$, where the row i comes before the column j.

We will think of our dataset $\mathbf{X}$ as the realization of a collection of random vectors $\{X_1, X_2, \dots X_n\}$. The typical view is that randomness of $\mathbf{X}$ comes from the fact that we could have drawn a different set of individuals from the population, in which case we would have seen a different dataset $\mathbf{X}$. To understand this randomness, let us assume that individuals are sampled at random from a population:

Definition 3.1. *A collection of random vectors $\{X_1, X_2, \dots X_n\}$ are called independent and identically distributed (i.i.d.) if $X_i \perp\!\!\!\perp X_j$ for $i \neq j$ and each X_i has the same marginal distribution as the others.*

The *i.i.d.* model is typically used to describe *simple random sampling*. Simple random sampling occurs when individuals are selected at random from some underlying population I, and a set of variables $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$ are recorded for each sampled individual i. Imagine for example a telephone survey, in which enumerators have a long list I of potential individuals to contact. They use a random number generator to choose an i at random from this list, contact them, and record responses to a set of k questions. This process is then repeated n times.

When X_i for $i = 1 \dots n$ denotes a collection of *i.i.d* random vectors, we'll refer to the distribution F that describes the marginal distribution of each X_i as the *population distribution*. The population distribution is the distribution we get when we randomly select any individual from the population.

Data is not always generated by simple random sampling, but when it is, we can imagine $\mathbf{X}$ as being formed by randomly choosing rows from a much larger matrix that records X_i for all individuals in the

population, depicted in Figure B.1.

> An alternative view of randomness in data used for causal inference is that it comes from the assignment of who gets treated and who does not, rather than from who is selected in the sample. This *design-based* view of uncertainty can be contrasted with the typical *sampling-based* uncertainty. This leads in some cases to different types of statistical tests to perform inference on treatment effects.

Another piece of terminology will be useful as we discuss samples and their population counterparts:

Definition 3.2. *A **statistic** or **estimator** is any function of the sample* $\mathbf{X} = (X_1', X_2', \ldots, X_n')'$.

A generic estimator or statistic will apply some function $g(\mathbf{X}) = g(X_1, X_2, \ldots X_n)$ to the collection of random vectors that constitute the sample. An example is the so-called *sample mean* $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$, which simply adds together X_i for across the sample and divides by the number of observations n. $\bar{X}_n$ is an example of a statistic. Since each of the X_i is a random variable/vector, it follows that $\bar{X}_n$ is itself a random variable/vector. This is true of statistics in general: they are random.

The reason that we also refer to statistics as "estimators" is that statistics often attempt to estimate a population quantity of some kind from data. For example, we might use $\bar{X}_n$ as an estimate of μ. Note that $\bar{X}_n$ is random, while μ is just a fixed number.

Notation: Often estimators are depicted with a "hat" on them, e.g. $\hat{\theta} = g(\mathbf{X})$. We'll use this notation to denote a generic estimator.

3.1.2 Convergence in probability and the law of large numbers

Consider an *i.i.d.* sample $\{X_1, \ldots X_n\}$ of some random variable X_i. The *sample average* of X_i in our data simply takes the arithmetic mean across these n observations:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

The law of large numbers (LLN) states the deep and useful fact that for very large n, it becomes very unlikely that $\bar{X}_n$ is very far from $\mu = \mathbf{E}[X_i]$, the "population mean" of X_i.

Theorem 1 (law of large numbers). *If X_i are i.i.d random variables and $E[X_i]$ is finite, then for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Note: The LLN is stated above for a random variable, but the result generalizes easily to random vectors. In that case, $\lim_{n\to\infty} P(||\bar{\mathbf{X}}_n - \mu||_2 > \epsilon) = 0$ where $||\cdot||_2$ denotes the Euclidean norm, i.e.: $|\bar{\mathbf{X}}_n - \mu| = (|\bar{\mathbf{X}}_n - \mu|)'(|\bar{\mathbf{X}}_n - \mu|)$, where $\bar{\mathbf{X}}_n$ is a vector of sample means for each component of X_i, and similarly for μ.

The law of large numbers is an example of *convergence in probability*:

Definition 3.3. *We say that Z_n converges in probability to Z if for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P(||Z_n - Z|| > \epsilon) = 0$$

In this definition, Z_n can be a random variable/vector. When Z_n is a random variable, then the notation $||Z_n - Z||$ jut refers to the absolute value of the difference: $|Z_n - Z|$. When Z_n is a vector, we can take $||Z_n - Z||$ to be the Euclidean norm of the difference.

The law of large numbers says that $\bar{X}_n \xrightarrow{n} \mu$, the sample mean converges in probability to the "population mean", or expectation, of X_i.

Exercise: This problem gives an example of a sequence that converges in probability to another random variable, rather than to a constant. Let $Z_n = Z + \bar{X}_n$, where Z is a random variable and $\bar{X}_n$ is the

sample mean of i.i.d. random variables X_i having zero mean and finite variance. Suppose furthermore that Z and $\bar{X}_n$ are independent. Show that $plim(Z_n) = Z$.

3.1.3 Convergence in distribution and the central limit theorem †

Our second notion of convergence of a sequence of random vectors is *convergence in distribution*. Consider first a sequence of scalar random variables:

Definition 3.4. *We say that a random variable Z_n converges in distribution to Z if, for any z such that the CDF $F_Z(z) = P(Z \leq z)$ of Z is continuous at z:*

$$\lim_{n\to\infty} P(Z_n \leq z) = F_Z(z)$$

Notation: When Z_n converges in distribution to Z, we write this as $Z_n \xrightarrow{d} Z$. As with convergence in probability, Z can be a random vector or a constant.

Convergence in distribution essentially says that the CDF of Z_n point-wise converges to the CDF of Z. By "point-wise", we mean that this occurs for each value z. When $Z_n \xrightarrow{d} Z$, we often refer to Z as the "large-sample" or "asymptotic" distribution of Z_n.

Note: The requirement that we only consider z where $F_Z(z)$ is continuous is a technical condition, which we can often ignore because we'll be thinking about continuously distributed Z.

Note: The definition given above for convergence in distribution takes Z_n to be a random (scalar) variable to emphasize the idea, but the concept extends naturally to sequences of random vectors.

Note: Convergence in distribution is a weaker concept of convergence than convergence in probability: one can show that if $Z_n \xrightarrow{p} Z$ then $Z_n \xrightarrow{d} Z$, but the reverse is not true (except when Z is a constant).

The central limit theorem (CLT) is the most important application of the concept of convergence in distribution. The CLT tells us that if we construct from the sample mean $\bar{X}_n$ the a random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)$, then the sequence Z_n converges in distribution to that of a normal random variable.

Theorem 2 (central limit theorem). *If X_i are i.i.d random vectors and $\mathbb{E}[X_i' X_i] < \infty$, then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where $\Sigma = Var(X_i)$, $\mu = \mathbb{E}[X_i]$, and $\mathbf{0}$ is a vector of zeros for each component of X_i.

The central limit theorem is quite remarkable. It says that *whatever* the distribution of X_i is, the limiting distribution of $\bar{X}_n$ (recentered by μ and rescaled by $\sqrt{n}$) will be a normal distribution. This striking result will pave the way for us to perform inference on the expectation of a random variable, without knowing its full distribution.

The practical value of the CLT is that it delivers an approximation to the distribution of $\bar{X}_n$. For large n, we know that $\sqrt{n}(\bar{X}_n - \mu)$ has approximately the distribution $N(0, \Sigma)$. Using properties of the normal distribution, we can re-arrange this to say that $\bar{X}_n \sim N(\mu, \Sigma/n)$, approximately. To get a good guess of the distribution of $\bar{X}_n$, we only need to have estimates of μ and Σ, which is much easier than estimating the full CDF of X_i from data.

Several important properties of convergence in probability and convergence in distribution that will be used in analyzing linear regression, including the continuous mapping theorem and the delta method, are described in Section B.6.

3.1.4 The continuous mapping theorem †

The last piece of statistical theory that we'll need to understand the OLS linear regression estimator is the *continuous mapping theorem* which let's us apply the LLN or the CLT to pieces of an expression, and then combine them to say something about the asymptotic behavior of the whole.

Theorem 3 (continuous mapping theorem). *Consider a sequence Z_n of random vectors and a continuous function h. Then:*

- *if $Z_n \xrightarrow{p} Z$, then $h(Z_n) \xrightarrow{p} h(Z)$*
- *if $Z_n \xrightarrow{d} Z$, then $h(Z_n) \xrightarrow{d} h(Z)$*

Formally, what the CMT states is that the notions of convergence in probability and convergence in distribution are preserved when we apply a continuous function to each random vector in a sequence Z_n.

Examples: Suppose $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$ with c a constant. Then:

- $Z_n + Y_n \xrightarrow{d} Z + c$
- $Z_n \cdot Y_n \xrightarrow{d} cZ$
- $Z_n / Y_n \xrightarrow{d} Z/c$ if $c \neq 0$.

These expressions are referred to collectively as *Slutsky's Theorem*, but they are really just applications of the CMT. See Appendix B.6 for details.

3.2 The linear regression model

Note on notation: In this section we'll simplify notation by dropping i subscripts when discussing population quantities. We'll add them back in Section 3.3 when we get to estimation. Remember that with *i.i.d.* data, it doesn't matter whether we include the i indices or not, because the distribution of variables in each observation i is the same as the population distribution.

Given a random variable Y and a random vector X, the *linear-regression model* says that

$$Y = X'\beta + \epsilon \tag{3.1}$$

where

$$\mathbb{E}[\epsilon|X] = 0 \tag{3.2}$$

We'll refer to the vector β appearing in Eq. (3.1) the *coefficient vector* from a regression of Y on X (as a reminder of notation: $\beta'X = \sum_j \beta_j \cdot X_j$). The term ϵ is often called an *error term* or *residual*.[1]

The linear regression model holds for some β if and only if the conditional expectation function of Y on X is a linear function of X, that is:

$$\mathbb{E}[Y|X] = X'\beta \tag{3.3}$$

In almost all cases in which we use the linear regression model, one of the components of X is taken to be non-random and simply equal to one. It thus contributes a constant to the function $X'\beta$, for example:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_k \cdot X_k + \epsilon \tag{3.4}$$

where here we have started the numbering at 0, so that β has $k+1$ components. In this notation X also has $k+1$ components: $X = (1, X_1, \ldots X_k)'$. However, to keep notation compact, we'll often ignore the distinction between a constant and random elements in X.

Accordingly, if we let k be the total number of components in $X = (X_1, X_2, X_3 \ldots X_k)'$ (including any constant term), then notice that Eq. (3.2) implies the following k equations:

$$\mathbb{E}[X\epsilon] = \begin{pmatrix} \mathbb{E}[X_1 \cdot \epsilon] \\ \mathbb{E}[X_2 \cdot \epsilon] \\ \vdots \\ \mathbb{E}[X_k \cdot \epsilon] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_1 \cdot (Y - X'\beta)] \\ \mathbb{E}[X_2 \cdot (Y - X'\beta)] \\ \vdots \\ \mathbb{E}[X_k \cdot (Y - X'\beta)] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.5}$$

[1] The Hansen textbook reserves the term "residual" for an *estimated* value of ϵ that arises in the context of the ordinary least squares estimator. I'll refer to ϵ above as a residual, and what Hansen calls a residual a "fitted residual" in Sec. 3.3.

To see that $\mathbb{E}[\epsilon|X] = 0$ implies $\mathbb{E}[\epsilon \cdot X_j] = 0$ for any $j = 1 \dots k$, use the law of iterated expectations:

$$\mathbb{E}[\epsilon \cdot X_j] = \mathbb{E}\left\{\mathbb{E}[\epsilon \cdot X_j | X]\right\} = \mathbb{E}\left\{\mathbb{E}[\epsilon|X] \cdot X_j\right\} = \mathbb{E}\left\{0 \cdot X_j\right\} = 0$$

It's probably a good idea to stare at this and make sure it makes sense. Conditional on any value $X = x$, the component X_j has some fixed value x_j. Thus, we can pull it out of the inner expectation, so that $\mathbb{E}[\epsilon \cdot X_j | X = x] = \mathbb{E}[\epsilon|X = x] \cdot x_j$. Then we take the outer expectation (curly braces) over values x.

3.2.1 Regression as least squares

We can also write the linear regression vector in a second way: it minimizes the population mean-squared error between Y and a linear function of the components of X:

$$\beta = \underset{\gamma \in \mathbb{R}^k}{\operatorname{argmin}} \ \mathbb{E}[(Y - X'\gamma)^2] \tag{3.6}$$

This says that the value of the β appearing in Eq. (3.1) is exactly the one that minimizes the expectation of the squared difference between Y and the "regression line" $X'\beta$ implied by β and X. We'll establish Eq. (3.6) in Section 3.2.2.

Note that we are not constraining the values that γ can take in this minimization problem, rather we have an unconstrained minimization in which we search over *all* $\gamma \in \mathbb{R}^k$. That means that to minimize the mean squared error, it must satisfy the following k first-order-conditions (FOCs), one for each of its components β_j for $j = 1 \dots k$:

$$\frac{\partial \mathbb{E}[(Y - X'\beta)^2]}{\partial \beta_j} = \mathbb{E}[2(Y - X'\beta) \cdot X_j] = 0 \tag{3.7}$$

where we've used that $X'\beta = \sum_{j=1}^k X_j \cdot \beta_j$. This is equivalent to $\mathbb{E}[X_j \cdot \epsilon] = 0$, if we define $\epsilon = Y - X'\beta$. This leads exactly to the linear regression model of Equations 3.1 and 3.5.

Thus we've seen that the minimizer of the mean squared error between Y and a linear function of X must be equal to the regression coefficient vector β. The box at the end of Section 3.2.2 shows that this also goes in the other direction: the β defined by Equations 3.1 and 3.5 must be the β that solves (3.6).

Note: I've assumed in the above that $\mathbb{E}[(Y - X'\gamma)^2]$ is differentiable with respect to γ and that we can interchange the derivative and the expectation (this requires regularity conditions that allow us to appeal to the dominated convergence theorem, but we don't need to worry about these technicalities here).

3.2.2 The population regression vector β

Since the restrictions (3.5) implied by the linear regression model provide a system of k equations in the k unknowns $\beta_1 \dots \beta_k$, it generally has a unique solution. A general expression for this solution is (see box below):

$$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[X \cdot Y] \tag{3.8}$$

From this expression it is clear that to define β we need the inverse matrix $\mathbb{E}[XX']^{-1}$ to exist, meaning that the matrix $\mathbb{E}[XX']$ is *invertible*. A convenient characterization of when $\mathbb{E}[X'X]$ will be invertible is given by the following proposition:

Proposition 3.1. *The matrix $\mathbb{E}[XX']$ has an inverse $\mathbb{E}[XX']^{-1}$, if and only if for all $\gamma \in \mathbb{R}^k$:*

$$P(X'\gamma \neq 0) > 0$$

Proposition 3.1 says that there exists no value γ that makes $X'\gamma$ equal to the zero vector, with probability one (remember that X here is a random vector). When there is such a γ, we say that there is *perfect multicollinearity* among our regressions $X = (X_1, X_2, \dots X_k)$.

Definition. *We say that there is **perfect multicollinearity** among our regressors (in the population) if there exists some $\gamma \in \mathbb{R}^k$ such that $P(X'\gamma = 0) = 1$.*

Example: Suppose that our regression includes a constant $X_1 = 1$, a binary variable indicating that a given individual is married: $X_2 = married$, and a second binary variable X_3 that indicates that a given individual is not married. Then, since $X = (1, married, 1 - married)'$, we have that $X'(-1, 1, 1) = 0$ for all realizations of X. Thus, we have perfect multicollinearity: $X'\gamma = 0$ regardless of the value of *married* and hence with probability one.

Review: using matrix inverses to solve a system of linear equations

Suppose we have a system of k equations in k variables

$$\begin{aligned} a_{11} \cdot x_1 + a_{21} \cdot x_2 + \cdots + a_{k1} \cdot x_k &= b_1 \\ a_{12} \cdot x_1 + a_{22} \cdot x_2 + \cdots + a_{k2} \cdot x_k &= b_1 \\ &\vdots \\ a_{1n} \cdot x_1 + a_{2n} \cdot x_2 + \cdots + a_{kk} \cdot x_k &= b_k \end{aligned} \tag{3.9}$$

We seek a solution $\mathbf{x} = (x_1, x_2, \ldots x_n)$ that satisfies all of the above equations. Let us gather all of the coefficients in to a $k \times k$ matrix and call it $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{n1} \\ a_{21} & a_{22} & \ldots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \ldots & a_{nk} \end{pmatrix}$$

Our system of Equations (3.9) says, in vector notation, that $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = (b_1, b_2, \ldots b_k)'$ is a vector composed of the values appearing on the RHS in Eq. (3.9).

If the matrix $\mathbf{A}$ is *invertible*, this means that there exists a unique matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_k$, where $\mathbf{I}_k$ is the $k \times k$ *identity matrix*. It has entries of one along the diagonal and zeros everywhere else:

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

Note that the identity matrix $\mathbf{I}_k$ has the property that $\mathbf{I}_k\boldsymbol{\lambda} = \boldsymbol{\lambda}$ for any vector $\boldsymbol{\lambda} \in \mathbb{R}^n$.

Thus, if we start with the equation $\mathbf{Ax} = \mathbf{b}$ and multiply both sides by $\mathbf{A}^{-1}$, we get that

$$\mathbf{A}^{-1}(\mathbf{Ax}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{x} = \mathbf{I}_k\mathbf{x} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Thus, we've shown that $\mathbf{x}$ must be equal to $\mathbf{A}^{-1}\mathbf{b}$. This value definately satisfies (3.9), which we can verify by:

$$\mathbf{A}(\mathbf{A}^{-1}\mathbf{b}) = (\mathbf{A}\mathbf{A}^{-1})\mathbf{b} = \mathbf{I}_k\mathbf{b} = \mathbf{b}$$

Also, it is the *only* value of $\mathbf{x}$ that satisfies the system (3.9). The solution exists and is unique, provided that $\mathbf{A}^{-1}$ exists.

Furthermore, one can show that the $\mathbf{x}$ solving $\mathbf{Ax} = \mathbf{b}$ is unique *only if* $\mathbf{A}$ is invertible. $\mathbf{A}$ is invertible if and only if there exists no $\boldsymbol{\lambda} \in \mathbb{R}^k$ that differs from the zero vector (i.e. it is not all zeros), for which $\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}$ (here $\mathbf{0}_k$ is a vector composed of k zeros). Thus if $\mathbf{A}$ is not invertible, there is such a vector $\boldsymbol{\lambda}$. Suppose we have one solution $\mathbf{x}$ to $\mathbf{Ax} = \mathbf{b}$. Then $\mathbf{x} + \alpha\boldsymbol{\lambda}$ is another solution, for any value of α, because $\mathbf{A}(\mathbf{x} + \alpha\boldsymbol{\lambda}) = \mathbf{Ax} + \alpha\mathbf{A}\boldsymbol{\lambda} = \mathbf{b} + \mathbf{0}_k = \mathbf{b}ss$.

Linear regression vs. linear projection When people talk about "running a regression", the quantity they are estimating is (3.8), whether or not the conditional expectation function $\mathbb{E}[Y|X]$ is

actually linear in X as the linear regression model assumes. Thus, rather than Eqs. (3.1) and (3.2) we could have gotten away with introducing β with a so-called *linear projection model*, which just says that

$$Y = X'\beta + \epsilon \qquad \text{where} \qquad \mathbb{E}[\epsilon \cdot X_j] = 0 \text{ for all } j = 1 \dots k \tag{3.10}$$

Whether one starts from Eq. (3.3) or from (3.10), we're talking about the same β. We'll call this β, which has the explicit formula (3.8), the *coefficient vector* or the *linear regression vector*.

3.2.3 Regression in terms of covariances

We know the general formula (3.8) for the vector β, which involves inverting the matrix $\mathbb{E}[XX']^{-1}$ and multiplying it by the vector $\mathbb{E}[XY]$. While the matrix formula holds generally, it turns out that we can still write expressions for the individual components of β in terms of covariances and variances, which is helpful in understanding the mechanics of how regression works.

Simple linear regression

When we just have a single regressor and a constant, we call this *simple linear regression*:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon \tag{3.11}$$

where X is a scalar. Note that this is really a $k = 2$ instance of regression, in which one regressor is a constant and the other is a random variable. In this case the familiar expressions $\beta_1 = \frac{Cov(X,Y)}{Var(X)}$ and $\beta_0 = \mathbb{E}[Y] - \beta_1 \cdot \mathbb{E}[X]$ can be derived from 3.8, using the formula for the inverse of a 2×2 matrix (this is a fun exercise!).

Multiple regressors and a constant

This principle generalizes to the general setting in which we have a regression equation with a constant and k additional regressors $X_1, X_2, \dots X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots \beta_k \cdot X_k + \epsilon \tag{3.12}$$

Note first that since one of our regressors is a constant, the system of equations (3.5) implies that $\mathbb{E}[\epsilon] = 0$. Then the remainder of the equations in (3.5) can be read as saying that each X_j is *uncorrelated* with the error, since $Cov(X_j, \epsilon) = \cancel{\mathbb{E}[X_j \cdot Y]} - \cancel{\mathbb{E}[X_j]} \cdot \mathbb{E}[Y] = 0$.

Proposition 3.2 ("regression anatomy" formula). *The coefficient on X_j in regression (3.12) is*

$$\beta_j = Cov(\tilde{X}_j, Y)/Var(\tilde{X}_j),$$

where $\tilde{X}_j$ is the residual from a regression of X_j on all of the other regressors and a constant.

The text *Mostly Harmless Econometrics* refers to Proposition 3.2 as the "regression anatomy" formula because it allows us to translate the complicated expression for the full vector $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ into a simpler expression for each of the components β_j.

Note: We'll see when we get to estimation in Section 3.3 that Proposition 3.2 has a sample analog, referred to as the *Frisch-Waugh-Lovell* theorem. Proposition 3.2 constitutes a "population version" of this very useful result.

Note: A Corollary to Proposition 3.2 is that we can also write β_j as $Cov(\tilde{X}_j, \tilde{Y}_j)/Var(\tilde{X}_j)$, where we define $\tilde{Y}_j$ to be the residual from a regression of Y on all the regressors aside from X_j, and a constant. This follows because the difference between $Y - \tilde{Y}_j$ is uncorrelated with $\tilde{X}_j$.

3.3 The ordinary least squares (OLS) estimator†

Now let's turn to *estimation* in the linear regression model. The standard estimator for β in the linear regression model is referred to as the *ordinary least squares* (OLS) estimator $\hat{\beta}_{OLS}$. Since this is the only estimator for β that we'll consider, we'll just write it as $\hat{\beta}$, to avoid writing OLS over and over again.

To define the OLS estimator $\hat{\beta}$ we suppose that we have a sample $(Y_i, X_{1i}, X_{2i}, \dots X_{ki})$ of Y and some set of regressors X_1 to X_k. Let n be the number of observations in our sample. *Note:* we will later assume that our sample is $i.i.d$, but we don't need to use that fact right now.

A simple way to define the OLS estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots \hat{\beta}_k)$ is as the minimizer of the sample analog of the least squares minimization, in which we replace the population expectation with the sample mean:

$$\hat{\beta} = \underset{\gamma \in \mathbb{R}^k}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\gamma)^2 \tag{3.13}$$

Given the OLS estimator $\hat{\beta}$, let us make the following definitions:

- The *fitted value* $\hat{Y}_i$ for observation i is $\hat{Y}_i = X_i'\hat{\beta} = \sum_{j=1}^{k} \hat{\beta}_j \cdot X_{ji}$
- The *fitted residual* for observation i is $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Note that for each i, we have that $Y_i = \hat{Y}_i + \hat{\epsilon}_i$ (by definition)

Equation (3.13) explains the origin of the name "ordinary least squares", as $\hat{\beta}$ is defined as the value of γ that minimizes the sample sum of squares.

What is the solution to the minimization problem (3.13)? Taking the first-order-condition with respect to each γ_j, we obtain the following system of equations:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_{1i} \cdot (Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{1i} \cdot \hat{\epsilon}_i = 0 \\
\frac{1}{n}\sum_{i=1}^{n} X_{2i} \cdot (Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{2i} \cdot \hat{\epsilon}_i = 0 \\
&\vdots \\
\frac{1}{n}\sum_{i=1}^{n} X_{ki} \cdot (Y_i - X_i'\hat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} X_{ki} \cdot \hat{\epsilon}_i = 0
\end{aligned} \tag{3.14}$$

which can be summarized by the matrix equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}) = \mathbf{0}$$

$\mathbf{0}$ is a vector of k zeros. This is exactly analogous to Eq. (3.5), except that we have replaced the population expectations $\mathbb{E}$ with sample averages $\frac{1}{n}\sum_i$. Rearranging the above:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \tag{3.15}$$

where recall that $X_i = (X_{1i}, X_{2i}, \dots X_k)'$ is a vector and Y_i is a scalar for each i. Since X_i is $k \times 1$ and X_i' is $1 \times k$, $X_i X_i'$ is a $k \times k$ matrix. In Equation 3.15 we've used that by the distributive property of matrix multiplication, we can sum over the observations i *and then* multiply by β, which is equivalent to multiplying and then summing the $k \times 1$ vector $X_i X_i'\hat{\beta}$ over observations.

We can obtain a more compact notation for Equation 3.15 by introducing an $n \times k$ matrix $\mathbf{X}$, that records all of our observations of all of the regressors:

$$\mathbf{X} := \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \underbrace{\left.\begin{pmatrix} (X_{11}, X_{21}, \dots X_{k1}) \\ (X_{12}, X_{22}, \dots X_{k2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{kn}) \end{pmatrix}\right\} n \text{ rows}}_{k \text{ columns}}$$

The matrix $\mathbf{X}$ is often called the *design matrix*.

Similiarly, we define a $k \times 1$ vector of our observations of Y:

$$\mathbf{Y} := \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

In this notation, we can rewrite the matrix $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)$ as $\frac{1}{n}\mathbf{X}'\mathbf{X}$. We can then write 3.15 in the compact form:

$$(\mathbf{X}'\mathbf{X})\hat{\beta} = \mathbf{X}'\mathbf{Y} \tag{3.16}$$

where we've multiplied both sides by n.

For Equation 3.16 to have a unique solution, we need for the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ to be invertible (see the box in Section 3.2.2 for a review of solving a sytem of linear equations). The following proposition provides a characterization of when this will be the true:

Proposition 3.3. *Provided that $n > k$, the matrix $\mathbf{X}'\mathbf{X}$ is invertible if none of the columns of $\mathbf{X}$ can be written as linear combinations of the other. That is: $X'\gamma \neq \mathbf{0}$ for all $\gamma \in \mathbb{R}^k$.*

This condition can be referred to as no perfect multicollinearity *in the sample*. When it holds, we obtain an explicit expression for the OLS estimator $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.17}$$

More matrix notation:
Following the notation we've developed to define the OLS estimator, we can also define $k \times 1$ vectors of the fitted values $\hat{Y}_i$, the fitted residuals $\hat{\epsilon}_i$, and the population residuals ϵ_i:

$$\hat{\mathbf{e}} := \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \qquad \hat{\mathbf{Y}} := \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} \qquad \boldsymbol{\epsilon} := \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

While $\hat{\mathbf{e}}$ and $\hat{\mathbf{Y}}$ are built with estimates from the data, note that $\boldsymbol{\epsilon}$ is not observable. However, under the assumption that the regression model $Y_i = X_i'\beta + \epsilon_i$ holds for each i, we have that

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \tag{3.18}$$

Note that we can also write

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{e}} \tag{3.19}$$

where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$.

3.3.1 The Frisch-Waugh-Lovell theorem*

Suppose we're interested in just *part* of the vector $\hat{\beta}$. That is, we separate our regressors $X_1 \dots X_k$ into two groups, let's say $X_1 \dots X_j$ and $X_{j+1} \dots X_k$, for some j (this is without loss of generality since

we could always re-order the indexing of the regressors). Our object of interest will be $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, where we introduce the notation:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_{\mathbf{1}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_j \end{pmatrix} \\ \begin{pmatrix} \hat{\beta}_{j+1} \\ \hat{\beta}_{j+2} \\ \dots \\ \hat{\beta}_k \end{pmatrix} \end{pmatrix}$$

Analogously, define the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ as

$$\mathbf{X}_1 := \underbrace{\left.\begin{pmatrix} (X_{11}, X_{21}, \dots X_{j1}) \\ (X_{12}, X_{22}, \dots X_{j2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{jn}) \end{pmatrix}\right\} n \text{ rows}}_{j \text{ columns}} \quad \text{and} \quad \mathbf{X}_2 := \underbrace{\left.\begin{pmatrix} (X_{j+1,1}, X_{j+2,1}, \dots X_{k1}) \\ (X_{j+1,2}, X_{j+1,2}, \dots X_{k2}) \\ \vdots \\ (X_{j+1,n}, X_{j+2,n}, \dots X_{kn}) \end{pmatrix}\right\} n \text{ rows}}_{(k-j) \text{ columns}}$$

where $\mathbf{X}_1$ is a matrix of observations of the regressors $X_1 \dots X_j$ and $\mathbf{X}_2$ is a matrix of observations of the regressors $X_{j+1} \dots X_k$.

Define $n \times n$ projector matrices $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{P}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$, and corresponding annihilator matrices $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ and $\mathbf{M}_2 = \mathbf{I}_n - \mathbf{P}_2$. Note that by the same logic as Equation (**??**), the matrix $\mathbf{M}_1$ annihilates $\mathbf{X}_1$ (that is, $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$, where $\mathbf{0}$, is a set of j zeroes), and similarly $\mathbf{M}_2\mathbf{X}_2 = \mathbf{0}$, where now $\mathbf{0}$, is a set of $k - j$ zeroes

The matrix $\mathbf{P}_1$ projects vectors in $\mathbb{R}^n$ into the subspace spanned by the columns of $\mathbf{X}_1$, which are the first j columns of $\mathbf{X}$. The matrix $\mathbf{M}_1$ projects vectors in $\mathbb{R}^n$ into the subspace orthogonal to the columns of $\mathbf{X}_1$. Similarly,$\mathbf{P}_2$ projects vectors in $\mathbb{R}^n$ into the subspace spanned by the columns of $\mathbf{X}_2$, which are the last $(k - j)$ columns of $\mathbf{X}$.

With the matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ in hand, we can now give an explicit formula for $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$, known famously as the *Frisch-Waugh-Lovell theorem*:

Proposition 3.4 (Frisch-Waugh-Lovell theorem).

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}$$

and

$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}$$

Now let's see how the Frisch-Waugh-Lovell theorem relates to the "regression anatomy" result Proposition 3.2. Since $\mathbf{M}_2$ is idempotent, we can write

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = (\mathbf{X}_1'\mathbf{M}_2\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{Y}$$

where $\tilde{\mathbf{X}}_1 := \mathbf{M}_2\mathbf{X}_1$, and we've used that $\mathbf{M}_2$ is a symmetric matrix: $\mathbf{M}_2' = \mathbf{M}_2$. The $n \times k$ matrix $\tilde{\mathbf{X}}_1 := \mathbf{M}_2\mathbf{X}_1$ collects the residuals from a series of j regressions: for each $\ell = 1 \dots j$, column ℓ of $\tilde{\mathbf{X}}_1$ is composed of the residuals from a regression of X_ℓ on $X_{j+1} \dots X_k$.

An analogous formula applies for $\hat{\boldsymbol{\beta}}_{\mathbf{2}}$, where $\tilde{\mathbf{X}}_2$ collects the residuals from regressions of each X_ℓ on $X_1 \dots\dots X_j$, for $\ell = j + 1 \dots k$. In the special case in which $\mathbf{X}_2$ has a single column (e.g. we're interested only in $\hat{\beta}_k$, and we include a constant in the regression (e.g. $X_1 = 1$), then we get exactly a sample version of Proposition 3.2.

The Frisch-Waugh-Lovell theorem allows us to obtain an expression for each slope coefficient

estimate $\hat{\beta}j$ in terms of sample variances and covariances. In particular:

$$\hat{\beta}_j = \frac{\widehat{Cov}(\hat{\epsilon}^j, Y)}{\widehat{Var}(\hat{\epsilon}^j)} \tag{3.20}$$

where we let $\hat{\epsilon}^j$ denote the fitted residuals from a regression of X_j on all the other regressors and a constant. This is an analog of the population residual $\tilde{X}_j$ from this same regression. Equation (3.20) provides a "sample analog" to the regression anatomy formula: Proposition 3.2.
The operators $\widehat{Cov}$ and $\widehat{Var}$ appearing in Eq. (3.20) are defined as follows. Let $\mathbf{A} = (A_1, A_2, \dots A_n)$ $\mathbf{B} = (B_1, B_2, \dots B_n)$ be $n \times 1$ vectors composed of observations of a random variable A_i and B_i, respectively. Let $\bar{A}_n = \frac{1}{n}\sum_{i=1}^n A_i$ be the sample mean of A_i, and similarly for $\bar{B}_n$. Then, we define:

$$\widehat{Cov}(A, B) = \left(\frac{1}{n}\sum_{i=1}^n A_i \cdot B_i\right) - \bar{A}_n \cdot \bar{B}_n$$

and

$$\widehat{Var}(A) = \widehat{Cov}(A, B) = \left(\frac{1}{n}\sum_{i=1}^n A_i^2\right) - \left(\bar{A}_n\right)^2$$

As a special case of Eq. (3.20), we have that in simple linear regression

$$\hat{\beta}_1 = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}$$

where in this case $\hat{\epsilon}_i^j$ is simply equal to X_i, the i^{th} observation of our single regressor X. We can work out the estimate of the constant β_0 from the fact that the fitted residual $\hat{\epsilon}_i$ satisfies $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i) = 0$. $\hat{\beta}_0$ is thus $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_0 \cdot \bar{X}_n$.

3.3.2 A review of notation

Let's review the notation that we've introduced in this section, because it can be confusing.

- We began with a random variable Y and a random vector X, which are related by $Y = X'\beta + \epsilon$ in "the population". The random vector X can be written $X = (X_1, X_2, \dots X_k)'$, where each X_j is a different *regressor*. No i subscripts are necessary here.
- Then we draw a random *sample*, where observations are indexed by $i = 1 \dots n$. Y_i is a random variable reflecting the value of Y in the i^{th} observation, and X_i assembles the value of all regressors for observation i into a random vector: $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$.
- When discussing the OLS estimator, it is convenient to assemble information across all of the observations, leading to the $n \times 1$ vector $\mathbf{Y}$ and the $n \times k$ matrix $\mathbf{X}$.

Consider the following toy dataset, where $n = 4$ and $k = 3$. This reflects a realization of the random matrix $\mathbf{X}$ and the random vector $\mathbf{Y}$:

i	X1	X2	X3	Y
1	1	4	0	23
2	1	3	1	54
3	1	2	1	21
4	1	6	0	77

The 4×3 matrix framed by a large red box is $\mathbf{X}$ in our sample. The smaller green box inside indicates X_3 laid out as a row vector: the values of each of the three regressors in the third observation. The blue

skinny rectangle indicates the $n \times 1$ vector $\mathbf{Y}$. Note that our first "regressor" `X1` is simply one for each observation, and contributes a constant to our regression. Regressor `X3` is a binary or "dummy" random variable: taking values of only zero or one for all obseravtions.

3.4 Statistical properties of the OLS estimator†

In this section we'll see that the OLS estimator $\hat{\beta}$ has many of the desirable properties introduced in Section C.3. It is consistent for the true population regression coefficient vector β, and has an asymptotically normal distribution. Knowing this will allow us to test hypotheses about the regression vector β. We also show in this esction that OLS is an unbiased estimator of β, and is an efficient estimator in a precise sense.

Recall from Section C.3 that when considering the performance of an estimator, we want to compare it to the population parameter of interest, in this case β. How can we do this? Well, we know from Equation (3.17) that $\hat{\beta}$ is a function of our observations of the outcome $\mathbf{Y}$, and our observations of the regressor $\mathbf{X}$. So we need some way to relate these observations to the population parameter of interest β. Our model of Y_i does exactly that. Recall that Equation 3.1 describes how our observations of Y_i can be written in terms of the coefficients β. Equation (3.18) provides an equivalent statement of this in vector notation. Studying the statistical properties of $\hat{\beta}$ thus begins with the following crucial step: substitute our equation for $\mathbf{Y}$ (Eq. 3.18) into the definition of the estimator (Eq. 3.17):

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \boldsymbol{\epsilon}) \\ &= \cancel{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \end{aligned} \tag{3.21}$$

Eq. (3.21) is really quite remarkable: it says that regardless of whatever sample we ended up estimating $\hat{\beta}$ from, it is exactly equal to the true population parameter β, plus second term that depends on the vector of residuals $\boldsymbol{\epsilon}$ and the sample design matrix $\mathbf{X}$.

We'll now proceed in two steps. First, we'll study the distribution of $\hat{\beta}$ when our sample design matrix is held fixed. This allows us to establish that conditional on $\mathbf{X}$, the estimator $\hat{\beta}$ is unbiased and efficient. Then, we'll consider the properties of $\hat{\beta}$ as n gets very large.

Keep in mind what we're doing in this section: we're asking what the distribution of our estimator $\hat{\beta}$ is, given that the data in our sample was a random draw from an underlying population. This will allow us to think about questions like: how likely would we be to get an estimate $\hat{\beta}$ that is far from β, given that the sample we use to compute $\hat{\beta}$ is random (and thus could have been different that the one we actually see)?

3.4.1 Asymptotic properties of $\hat{\beta}$

Some statistical properties of the OLS estimator $\hat{\beta}$ hold in a finite sample, conditional on the sample $\mathbf{X}$ that is drawn. These properties, *unbiasedness* and *efficiency*, are described in Chapter 7 of *Statistics for Econometrics*.

Here we instead focus on properties of the OLS estimator as the sample size n gets very large. We'll first show that $\hat{\beta}$ is a consistent estimator for β, and then that its sampling distribution is asymptotically normal. For these results, we don't need for the linear regression model with $\mathbb{E}[\epsilon|X] = 0$ to hold. The large sample properties hold for the linear projection coefficient β even if the CEF of Y on X is not linear. As before, we assume that we have an independent and identically distributed sample:

Assumption 1 (linear projection model and *i.i.d* sampling). *(Y_i, X_i) is an i.i.d. sample from the model: $Y = X'\beta + \epsilon$ with $\mathbb{E}[\epsilon \cdot X] = \mathbf{0}$.*

To make claims that involve convergence in probability and convergence in distribution, we will consider a sequence of estimators $\hat{\beta}$, indexed by the sample size n. For each $n = 1, \ldots, \infty$ along the sequence, we assume that 1 holds. As a reminder (cf. Chapter B), in reality sample sizes never actually "grow" to infinity. In practice, we always have an actual sample that has some actual finite size n. The idea of an asymptotic sequence exists only to provide an *approximation* to the sampling distribution of $\hat{\beta}$ given our fixed n, which we will take to be accurate when the sample size is big enough.

3.4.1.1 Consistency

We'll first see that given the asymptotic sequence described above, $\hat{\beta} \xrightarrow{p} \beta$. That is, $\hat{\beta}$ is a consistent estimator of β.

Subtracting β from each side of Equation 3.21:

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$$

where in the second equality we've used that the factor of $\frac{1}{n}$ inside the matrix inverse cancels the one on $\mathbf{X}'\boldsymbol{\epsilon}$. Now let's consider this latter quantity alone. Expanding out the matrix product:

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i,$$

i.e. it is equal to the sample average of the random variable $X_i\epsilon_i$. To see the above, note that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$ is a $k \times 1$ vector, whose j^{th} element is equal to the inner product between ϵ and the j^{th} row of $\mathbf{X}'$. The j^{th} row of $\mathbf{X}'$ is equal to the j^{th} column of $\mathbf{X}$, which is comprised of the n observations of regressor X_j.

Thus, by the law of large numbers, we have that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{p} \mathbb{E}[X_i\epsilon_i]$, provided that $\mathbb{E}[X_i\epsilon_i] < \infty$. By the linear projection model (Assumption 1), $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, where $\mathbf{0}$ is a vector of k zeroes.

Similarly, we have by the law of large numbers that

$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} X_iX_i' \xrightarrow{p} \mathbb{E}[X_iX_i'],$$

In Chapter B we only considered the LLN for random vectors, not random *matrices* like X_iX_i'. But since you can always rewrite an $n \times m$ matrix as a vector with $n \cdot m$ elements, the LLN for vectors applies so long as each element of the matrix $\mathbb{E}[X_iX_i']$ is finite. In the box below, I state a set of assumptions, "regularity conditions", that ensure we can use the law of large numbers here, and that all expectations that appear in this section exist.

Given that $\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbb{E}[X_iX_i']$, the continuous mapping theorem implies that

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \xrightarrow{p} \mathbb{E}[X_iX_i']^{-1}$$

That's because for a general invertible matrix $\mathbf{M}$, the matrix inverse function $\mathbf{M}^{-1}$ is a continuous function of each of the elements of $\mathbf{M}$.

Finally, by the continuous mapping theorem, we have that

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} = \underbrace{\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}}_{\xrightarrow{p}\mathbb{E}[X_iX_i']^{-1}} \underbrace{\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}}_{\xrightarrow{p}\mathbf{0}} \xrightarrow{p} \mathbb{E}[X_iX_i']^{-1}\mathbf{0} = \mathbf{0}$$

Thus we have proved that $\hat{\beta} \xrightarrow{p} \beta$.

Proposition 3.5. *OLS is consistent for β given Assumption 1 and the regularity conditions 2 below.*

Assumption 2 (regularity conditions for consistency). *Suppose that:*

1. $\mathbb{E}[Y_i^2]$ *is finite*
2. $\mathbb{E}[||X_i||^2]$ *is finite*
3. *We have no perfect multicollinearity in the population: that is, $\mathbb{E}[X_iX_i']$ is positive definite.*

3.4.1.2 Asymptotic normality*

Now let's use the central limit theorem to derive the asymptotic distribution of the OLS estimator. Let us pick up from the expression $\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$. Knowing that the central limit theorem will involve a factor of $\sqrt{n}$, let's rewrite this as

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \cdot \sqrt{n}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right)$$

Recall that $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, where $\mathbf{0}$ is a vector of k zeroes, and that $\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}$ is the sample mean of the random vector $X_i \cdot \epsilon_i$. Using the notation of Chapter B, let's denote this as $\overline{(X\epsilon)}_n := \frac{1}{n}\sum_{i=1}^n X_i \cdot \epsilon_i$. Then we can write the above as:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \cdot \sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right)$$

The rightmost factor in the above expression has exactly the form that we need to apply the CLT, in particular:

$$\sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right) \xrightarrow{d} N(\mathbf{0}, Var(X_i\epsilon_i)),$$

Note that since $\mathbb{E}[X_i\epsilon_i] = \mathbf{0}$, we can write the variance as

$$Var(X_i\epsilon_i) = \mathbb{E}[(X_i\epsilon_i)(X_i\epsilon_i)'] = \mathbb{E}[\epsilon_i^2 X_i X_i'] \tag{3.22}$$

Now we use the Slutsky theorem

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)}_{\xrightarrow{p}\mathbb{E}[X_iX_i']}^{-1} \cdot \underbrace{\sqrt{n}\left(\overline{(X\epsilon)}_n - \mathbb{E}[X_i\epsilon_i]\right)}_{\xrightarrow{d}\mathbb{E}[\epsilon_i^2 X_i X_i']} \xrightarrow{d} \mathbb{E}[X_iX_i']^{-1}N(\mathbf{0}, \mathbb{E}[\epsilon_i^2 X_i X_i'])) \tag{3.23}$$

That is, $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a random vector whose distribution is that of the matrix $\mathbb{E}[X_iX_i']^{-1}$ times a normal vector with mean zero (for each component) and variance-covariance matrix $\mathbb{E}[\epsilon_i^2 X_i X_i']$.

The RHS of (3.23) is thus equal to a linear combination of normal random vectors. A property of the normal distribution is the following. Let $X \sim N(\mu, \mathbf{\Sigma})$ be a k-component random vector. Then for any $k \times k$ matrix $\mathbf{A}$:

$$\mathbf{A}'X \sim N(\mathbf{A}'\mu, \mathbf{A}'\mathbf{\Sigma}\mathbf{A}))$$

(this can be seen as an example of the delta method, applied to a vector-valued function h). Thus, we can write (3.23) as

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}) \tag{3.24}$$

where $\mathbf{V} := \mathbb{E}[X_iX_i']^{-1}\mathbb{E}[\epsilon_i^2 X_i X_i']\mathbb{E}[X_iX_i']^{-1}$. We refer to $\mathbf{V}$ as the *asymptotic variance* of the OLS estimator.

A sufficient condition for us to be able to apply the CLT (see Section B.5) is that $\mathbb{E}[(X_i\epsilon_i)'(X_i\epsilon_i)] = \mathbb{E}[\epsilon_i^2 X_i'X_i]$ be finite. This requires finite *fourth* moments of the data, rather than the finite second moments assumed to prove consistency of OLS. To see why, note that for any j and ℓ:

$$\mathbb{E}[\epsilon_i^2 X_{ji} \cdot X_{\ell i}] = \mathbb{E}[(Y_i - X_i'\beta)^2 X_{ji} \cdot X_{\ell i}] = \mathbb{E}[Y_i^2 \cdot X_{ji}X_{\ell i}] - 2\mathbb{E}[X_i'\beta \cdot Y_i \cdot X_{ji} \cdot X_{\ell i}] + \mathbb{E}[\beta' X_i X_i' \beta \cdot X_{ji} \cdot X_{\ell i}]$$

which can be written out as a sum over expectations that each involve the product of four random variables. To keep all such terms finite, Hansen assumes the following:

Assumption 3 (regularity conditions for asymptotic normality). *Suppose that:*

1. $\mathbb{E}[Y_i^4]$ *is finite*
2. $\mathbb{E}[||X_i||^4]$ *is finite*
3. *We have no perfect multicollinearity in the population: that is,* $\mathbb{E}[X_iX_i']$ *is positive definite.*

3.4.1.3 Estimating the asymptotic variance*

Equation 3.24 is not immediately useful, unless we know the asymptotic variance matrix $\mathbf{V}$. Since we don't know $\hat{\mathbf{V}}$ before seeing the data, we will estimate it! In this section we see that we can construct a consistent estimator $\hat{\mathbf{V}}$ such that $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$. Doing this will open the door to hypothesis testing, which we'll consider in the next section.

Before seeing how hypothesis testing will work, let's consider how to construct the estimator $\hat{\beta}$ for the asymptotic variance of OLS. Note that $\mathbf{V} = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[\epsilon_i^2 X_i X_i']\mathbb{E}[X_i X_i']^{-1}$ has a "sandwich" form: it puts the matrix $\mathbb{E}[\epsilon_i^2 X_i X_i']$ (the meat)[2], between two instances of the matrix $\mathbb{E}[X_i X_i']^{-1}$ (the bread). By the continuous mapping theorem, we can construct an estimator $\mathbf{V}$ by making a sandwich out of consistent estimators for the meat and for the bread.

We've already seen that $\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}$ is a consistent estimator for the bread: $\mathbb{E}[X_i X_i']^{-1}$. An estimator for the meat $\mathbb{E}[\epsilon_i^2 X_i X_i']$ is not quite as obvious. It's sample analog $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 X_i X_i'$ would definitely work, but the true residuals ϵ_i are not observed. However, we can use the *fitted* residuals $\hat{\epsilon}_i$, which are a function of the observed data, instead. We can write this in matrix form as:

$$\hat{\mathbf{\Omega}} := \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 X_i X_i'$$

One can verify that $\hat{\Omega} \xrightarrow{p} \mathbb{E}[\epsilon_i^2 X_i X_i']$. Thus, we can form a consistent variance estimator as

$$\hat{\mathbf{V}}_{HC0} := \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{3.25}$$

Eq. (3.25) is referred to as the "HC0" estimator of $\mathbf{V}$, where HC stands for *heterokedasticity consistent*. This name comes from the fact that $\hat{\mathbf{V}}_{HC0}$ does not require the assumption of homoskedasticity (that $Var(\epsilon_i|X_i) = \sigma^2$ for all i) to be a consistent estimator of $\mathbf{V}$.

When you run a command like `regress y x, robust` in Stata, the default covariance estimator is the so-called "HC1" estimator of $\mathbf{V}$:

$$\hat{\mathbf{V}}_{HC1} := \frac{n}{n-k}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \hat{\mathbf{\Omega}} \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \tag{3.26}$$

Note that the additional factor $\frac{n}{n-k}$ will make very little difference when n is large compared with k, and will make no difference in the asymptotic limit, since $\frac{n}{n-k} \to 1$ as $n \to \infty$. Applying this rescaling however can be helpful when n is small. It's easiest to understand the justification in the case of homoskedasticity, which is left as an exercise (see box below).

Note: there are further estimators floating around, with names HC2, HC3, and HC4. These apply further modifications to $\hat{\mathbf{V}}_{HC0}$ (see the Hansen text for details). Other variance estimators exist for certain violations of the *i.i.d* sampling assumption, including *cluster-robust* variance estimators for clustered sampling and autocorrelation-consistent estimators for serially correlated panel data.

3.4.2 Inference on the regression vector β*

Given a consistent estimator of $\mathbf{V}$ like the HC0 or the HC1 estimator, we can transform the quantity $\sqrt{n}(\hat{\beta}-\beta)$ into one whose limiting distribution is well-understood, and contains no unknown parameters. This proves to be a much more useful result than Equation (3.24), because it allows us to test hypotheses about the population regression vector β. In particular, if we pre-multiply $\sqrt{n}(\hat{\beta}-\beta)$ by the matrix $\hat{\mathbf{V}}^{-1/2}$ (see box below for the definition of $\hat{\mathbf{V}}^{-1/2}$):

Proposition 3.6. *Given Assumption 1, the regularity conditions 3, and a $\hat{\mathbf{V}}$ such that $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$:*

$$\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} N(\mathbf{0}, \mathbb{I}_k)$$

where $\mathbb{I}_k$ is the $k \times k$ identity matrix.

[2]Ideally plant-based meat :)

The distribution $N(\mathbf{0}, \mathbb{I}_k)$ is that of k standard normal random variables, each of which is independent of the others (the variance-covariance matrix $\mathbb{I}_k$ has an entry of zero for each off-diagonal element). The power of Proposition 3.6 lies in the fact that the distribution appearing on the RHS, $N(\mathbf{0}, \mathbb{I}_k)$, contains no unknown quantities. We know exactly the probability that it associates to any event. Thus, for large n, we have a very good approximation to the distribution of $\sqrt{n}\hat{\mathbf{V}}^{-1/2}(\hat{\beta} - \beta)$. This provides the foundation for us to quantify uncertainty in our estimates $\hat{\beta}$ and test hypotheses about the regression vector β.

As a simple example of how the logic of Proposition 3.6 is useful, let's consider a simple setting, which turns out to be the most common one in practice: we are interested in the true value of a single regression coefficient, say β_j, in a regression that contains k regressors. A detailed discussion of hypothesis testing in the linear regression model is omitted.

Note that we can write β_j as $\mathbf{e}_j'\beta$, where $\mathbf{e}_j = (0, \dots, 1, \dots 0)'$ is a k-component vector that puts a one in position j, and zeros everywhere else. Similarly, $\mathbf{e}_j'\hat{\beta}$ picks out the single component $\hat{\beta}_j$ from the OLS estimator. It then follows from Equation (3.24) and the Delta method that

$$\sqrt{n}(\hat{\beta}_j - \beta_j) = \mathbf{e}_j'\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{d} \mathbf{e}_j' N(\mathbf{0}, \mathbf{V}) = N(\mathbf{e}_j'\mathbf{0}, \mathbf{e}_j'\mathbf{V}\mathbf{e}_j) = N(0, V_{jj})$$

where $V_{jj} = \mathbf{e}_j'\mathbf{V}\mathbf{e}_j$ is the j^{th} element along the diagonal of the matrix $\mathbf{V}$.

This implies, analogously to Proposition 3.6, that

$$\sqrt{n} \cdot \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathbf{V}}_{jj}}} \xrightarrow{d} N(0, 1) \tag{3.27}$$

where $\hat{\mathbf{V}}_{jj}$ is the j^{th} element along the diagonal of the matrix $\hat{\mathbf{V}}$, which is a consistent estimator of V_{jj}. Note that we could have written the LHS of Eq. (3.27) as $\sqrt{n}\hat{\mathbf{V}}_{jj}^{-1/2}(\hat{\beta}_j - \beta_j)$ as in Proposition 3.6, but since $\hat{\mathbf{V}}_{jj}$ is a scalar we may take its conventional square root and divide by it.

We define the *standard error* for the estimate $\hat{\beta}_j$ to be $se(\hat{\beta}_j) := \sqrt{\hat{\mathbf{V}}_{jj}/n}$. Note that the standard error is a quantity that is computed from the data, given $\hat{\mathbf{V}}$ (it is an *estimate*, rather than a population quantity). By Eq. (3.27), we know that the quantity $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j)$ converges in distribution to a standard normal.

This allows us to test hypotheses about the value of β_j, using our estimate $\hat{\beta}_j$ and $se(\hat{\beta}_j)$. Consider the null hypothesis: $\mathbf{H}_0 : \beta_j = \beta_0$ for some value β_0 (e.g. zero). Define the *T-statistic* for this hypothesis to be

$$T(\beta_0) = \frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)}$$

If $\mathbf{H}_0$ is true, then we know that $T(\beta_0) \xrightarrow{d} N(0, 1)$. Recall from Section C.4.1 that the *size* of a hypothesis test is the maximum probability of rejecting the null hypothesis, when the null hypothesis is in fact true. We can form a test with size α in the following way:

$$\text{reject } \mathbf{H_0} \text{ iff } |T(\beta_0)| > c$$

where c is a value such that the probability of a standard normal random variable having a magnitude of at least c is less than α. To do this in a way that maximizes power, we choose c to be exactly the $1 - \alpha/2$ quantile of the standard normal distribution: $c = \Phi^{-1}(1 - \alpha/2)$.

3.5 Back to regression and causality

With our review of linear regression complete, let us now return to our motivation for studying it: as a tool to estimate causal effects under selection-on-observables.

Recall from Section 2.2.3 that if we have a linear regression model (equivalently, a linear CEF as in Eq. 2.2)

$$Y_i = \beta_0 + \beta_D d + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon_i \tag{3.28}$$

then the coefficient β_D yields the average treatment effect of D on Y, assuming selection-on-observables holds. Given the results of the last section, we know how to estimate β_D as well as perform statistical inference on the parameter β_D, to understand the role of random chance in the value of our estimator. Thus, if the CEF is linear, we can implement statistical tests for the value of the average treatment effect, since $\beta_D = ATE$.

As mentioned in Section 2.2.3, assuming the linear form of Equation 3.28 is quite a strong assumption. In principle, one can always get around this problem by making use of more flexible regression equations, or nonparametric regression techniques. We'll talk a little about these later in the course, but the main drawback of nonparametric techniques is that they tend to require much more data to have much statistical precision, compared with OLS. The remainder of this section reviews a specific situation in which even though Equation may be (3.28), we can show that β_D remains a weighted average of $ATE(x)$ over different values of the covariates X. This situation occurs when the treatment is binary and itself has a CEF that is linear in the control variables.

3.5.1 Binary-treatment regressions that are "saturated" in controls*

Consider a regression of Y on a binary treatment variable D and X:

$$Y = \beta_D D + \beta_X' X + \epsilon \tag{3.29}$$

where the vector X is a set of indicator variables for an underlying categorical variable G. By this, I mean that $X = (\mathbb{1}(G = 1), \mathbb{1}(G = 2), \ldots, \mathbb{1}(G = N_g))'$, where $P(G \in \{1, 2, \ldots N_G\}) = 1$. We'll return to this kind of regression later, which is sometimes referred to as "saturated" in controls. It turns out that the coefficient on D in this regression can be written as:

$$\beta_D = \frac{\mathbb{E}[\{\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]\} \cdot Var(D_i|X_i)]}{\mathbb{E}[Var(D_i|X_i)]}$$

If we assume selection-on-observables, then we know that the term in brackets is equal to $ATE(x) = E[Y_i(1) - Y_i(0)|X_i = x]$. Then, we have that $\beta_D = \sum_{j=1}^{N_G} w_j \cdot ATE(x_j)$ where $x_1, x_2, \ldots$ are the values that X can take, i.e. x_j is a vector of N_G components composed of all zeros but a 1 in the j^{th} component. The $w_j = \frac{Var(D_i|X_i)\cdot P(X_i = x_j)}{E[Var(D_i|X_i)]}$ can be thought of as weights: they are positive and sum to one. This result can be found in Angrist and Pischke (2008).

As an example where a saturated regression might occur, suppose we are conducting a returns-to-schooling study and want to control for a student's gender and their mother's level of education:

$$G_i \in \{\underbrace{\text{"male and mother graduated high school"}}_{\text{Group 1}}, \underbrace{\text{"male and mother didn't graduate high school"}}_{\text{Group 2}},$$
$$\underbrace{\text{"female and mother graduated high school"}}_{\text{Group 3}}, \underbrace{\text{"female and mother did not graduate high school"}}_{\text{Group 4}}\}$$

In this example, $N_G = 4$, and the four values of X_i are map one-to-one with the four possible combinations of two binary variables: one for gender, and one for mother's high school completion.

An important property of a saturated set of controls is that it implies that $\mathbb{E}[D_i|X_i]$ is linear in X_i. To see this, note that since $X_{ji} = \mathbb{1}(G_i = j)$

$$\mathbb{E}[D_i|X_i] = \sum_{j=1}^{N_G} X_{ji} \cdot \gamma_j = X_i'\gamma$$

if we let $\gamma_j = \mathbb{E}[D_i|G_i = j]$. Cool!

To see this, we'll apply the regression anatomy formula to get the coefficient β_D:

$$\rho = \frac{Cov(Y_i, \tilde{D}_i)}{Var(\tilde{D}_i)}$$

where $\tilde{D}_i$ is the residual from a regression of D_i on X_i. Since the conditional expectation of D_i is linear in X_i we know that $\tilde{D}_i = D_i - \mathbb{E}[D_i|X_i]$. Then:

$$\beta_D = \frac{Cov(Y_i, (D_i - \mathbb{E}[D_i|X_i]))}{Var(D_i - \mathbb{E}[D_i|X_i])}$$

Warning: this is about to get messy. Let $\sigma^2_D(x)$ denote the conditional variance of D_i on X_i, i.e. $\sigma^2_D(x) := \mathbb{E}\left[(D_i - \mathbb{E}[D_i|X_i])^2 \middle| X_i = x\right]$. Noting that $\tilde{D}_i$ is mean zero and applying LIE to $\mathbb{E}[\sigma^2_D(X_i)]$, we have:

$$\beta_D = \frac{\mathbb{E}[Y_i(D_i - \mathbb{E}[D_i|X_i])]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]} = \frac{E\left\{\mathbb{E}[Y_i(D_i - \mathbb{E}[D_i|X_i])|X_i]\right\}}{\mathbb{E}[\sigma^2_D(X_i)]}$$

Consider the numerator. Applying the LIE over D_i:

$$\begin{aligned}\mathbb{E}[Y_i(D_i - \mathbb{E}[D_i|X_i])|X_i] &= \sum_{d\in\{0,1\}} P(D_i = d|X_i)\mathbb{E}[Y_i(D_i - \mathbb{E}[D_i|X_i])|D_i = d, X_i] \\ &= P(D_i = 1|X_i)\mathbb{E}[Y_i|D_i = 1, X_i](1 - P(D_i = 1|X_i)) \\ &\qquad - P(D_i = 0|X_i)\mathbb{E}[Y_i|D_i = 0, X_i]P(D_i = 1|X_i)\end{aligned}$$

To simplify notation, let's let $p(X_i) = P(D_i = 1|X_i) = \mathbb{E}[D_i|X_i]$. Note that $\sigma^2_D(X_i) = p(X_i)(1-p(X_i)) = P(D_i = 1|X_i)P(D_i = 0|X_i)$. Thus:

$$\mathbb{E}[Y_i(D_i - \mathbb{E}[D_i|X_i])|X_i] = \sigma^2_D(X_i)\left\{\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]\right\}$$

Thus, we've shown that

$$\beta_D = \frac{\mathbb{E}[\delta(X_i)\sigma^2_D(X_i)]}{\mathbb{E}[\sigma^2_D(X_i)]} \tag{3.30}$$

where $\delta(X_i) := \mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]$. We can think of $\delta(x)$ as a function that "matches" treated ($D_i = 1$) and control ($D_i = 0$) units with the same value of $X_i = x$, and then conducts a simple comparison between the treated and control means for that x.

3.5.2 Multi-valued treatment regressions that are saturated in controls*

Unfortunately, the result of the last section does not extend beyond the case of a binary treatment. Consider for example a regression

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta'_X X_i + \epsilon_i \tag{3.31}$$

where treatment takes on three possible values: 0,1,2, and D_1 is an indicator that treatment takes value 1 and D_2 is an indicator that treatment takes value 2. Assume that X includes a constant–therefore we omit an indicator for category 0.

Goldsmith-Pinkham et al. (2021) show that even if $\mathbb{E}[D_{1i}|X_i]$ and $\mathbb{E}[D_{2i}|X_i]$ are each linear in X (for example if X contains a fully-saturated set of controls), then β_1 and β_2 do not generally represent weighted averages of their respective treatment effects. Specifically, if we let $Y_i(0)$, $Y_i(1)$, and $Y_i(2)$ denote potential outcomes for each of the three treatments, β_1 is not a weighted average of $\mathbb{E}[Y_i(1)-Y_i(0)|X = x]$ across values of x, and β_2 is not a weighted average of $\mathbb{E}[Y_i(2) - Y_i(0)|X = x]$. Rather, β_1 is generally "contaminated" by the effects of the third treatment: it contains a second term that measures the effect of treatment 2 versus treatment 0. Similarly, β_1 is contaminated by the effects of treatment 1. One way to avoid the contamination bias problem is to replace Equation (3.31) with a more flexible regression equation that contains interactions between the treaments D_1, D_2 and the covariates X. See Goldsmith-Pinkham et al. (2021) for details.

Chapter 4

Instrumental variables

So far we've considered identification of causal effects under random assignment and more generally, selection-on-observables. Selection on observables is a powerful assumption: if one has the right X variables and controls for them carefully. But it is not always enough. In fact, it's usually not: do we really think we observe *everything* that we need to condition on to render treatment assignment independent of potential outcomes?

One popular method to deal with settings in which we do not have selection-on-observables is the use of *instrumental variables* (IV). You can think of IV as a method that let's selection into treatment depend on unobservables as well as observables. However, it also relies on strong assumptions. As we'll see, one of the cleanest examples where we can use IV is when we have a randomized experiment, but treatment uptake is imperfect and non-random.

4.1 Basic intuition with homogeneous effects

Suppose we have a binary treatment and treatment effects are homogenous across units:

$$Y_i(1) - Y_i(0) = \Delta$$

for some number Δ. We know from Section 1.6 that we can represent this with a regression equation

$$\begin{aligned} Y_i &= \underbrace{\mathbb{E}[Y_i(0)]}_{\beta_0} + \underbrace{\Delta}_{\beta_1} \cdot D_i + \underbrace{Y_i(0) - \mathbb{E}[Y_i(0)]}_{\epsilon_i} \\ &= \beta_0 + \beta_1 D_i + \epsilon_i \end{aligned}$$

where $\beta_1 = \Delta$, i.e. the slope coefficient on D_i is equal to (common) treatment effect Δ.

Recall that we have a selection bias problem when $\mathbb{E}[Y_i(0)|D_i = 0] \neq \mathbb{E}[Y_i(0)|D_i = 0]$, or equivalently $Cov(D_i, Y_i(0)) \neq 0$. If this is true then D_i is correlated with the error term ϵ_i.

Exercise: Show that $Cov(D_i, Y_i(0)) \neq 0$ if and only if $Cov(D_i, \epsilon_i) \neq 0$ in the above equation.

Recall that the linear regression model is built on the idea that $\mathbb{E}[\epsilon_i|D_i] = 0$, which implies that $Cov(D_i, \epsilon_i) = 0$. When this equality fails, OLS will not generally give consistent estimates of β_1. We refer to the situation $Cov(D_i, \epsilon_i) = 0$ as an *endogeneity problem*. When there is an endogeneity problem, the treatment D_i is often called *endogenous*, meaning (loosely speaking) that it is influenced by other things that cannot be easily excluded from the model, because they are related to our outcome.

4.1.1 The simple math of a single IV

An endogeneity problem can also occur treatment is not binary, and is instead a multivalued or continuous treatment. Suppose we have a general regression of the form:

$$Y_i = \beta_0 + \beta_1 S_i + \epsilon_i \tag{4.1}$$

where β_1 is of interest but $Cov(S_i, \epsilon_i) \neq 0$ (note that this nests the case we started with if $S_i = D_i$). In this case we say that the treatment S_i is endogenous.

Suppose that we are able to find a third observable variable Z_i, that unlike S_i is uncorrelated with the error term ϵ_i, i.e.

$$Cov(Z_i, \epsilon_i) = 0$$

Then notice that by substituting (4.1) and using linearity of the covariance operator:

$$Cov(Y_i, Z_i) = Cov(\beta_0 + \beta_1 S_i + \epsilon_i) = 0 + \beta_1 \cdot Cov(S_i, Z_i) + \cancel{Cov(\epsilon_i, Z_i)} = \beta_1 \cdot Cov(S_i, Z_i)$$

But $Cov(S_i, Z_i)$ is observable! That means so long as it is non-zero, we can solve for β_1 as:

$$\beta_1 = \frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)} \tag{4.2}$$

This means that even in the presence of selection bias, we can overcome it to identify β_1, provided that we have a variable Z_i that satisfies the two properties that we used:

1. $Cov(Z_i, \epsilon_i) = 0$. This is often referred to as instrument *validity*.
2. $Cov(Z_i, S_i) \neq 0$. This is often referred to as instrument *relevance*.

We'll refer to Z_i as an *instrumental variable*, and IV, or just an *instrument* for short.

4.1.2 Interpreting an IV though a causal graph

What kinds of variables might qualify as valid instruments? The first step is to find a variable Z for which it's at least plausible to believe that Z *only* influences Y *through* its effect on S. Otherwise, it would hard to be sure that the covariance between Z and Y does not come from the same unobservables that produce a correlation between S and ϵ.

The idea that Z only influences Y through S is called the IV *exclusion restriction*. Some people use *exclusion restriction* synonymously with instrument validity, but really you should think of instrument validity as comprising two assumptions: i) the exclusion restriction, ii) that Z is not statistically related to unobserved confounders. This will be made explicit when we get to IV with heterogeneous effects.

To illustrate how the exclusion restriction can help us, consider the diagram in Figure 4.1. Here, the variables Z, S and Y are observed, and we assume that the instrument Z has a causal effect of some kind on S, indicated by the first solid arrow. The second solid arrow indicates the causal effect of S on Y, which is the relationship we would like to measure.

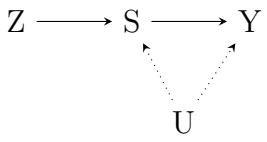


Figure 4.1: A causal diagram depicting the logic of an IV.

Unfortunately, there are one or more unobserved *confounders* U, which influence both S and Y. Since U is unobserved, the covariation between U and Y is captured by the error term ϵ in Equation 4.1. Endogeneity, or $Cov(S_i, \epsilon_i) \neq 0$, occurs because U also affects S.

Intuitively speaking, if we "wiggle" Z, this wiggles S which wiggles Z. But since there is no arrow from Z or S to U, this does not "wiggle" U. The variation in Y induced by variation in Z is free from the confounding variation in U, and we can identify β_1. This is awesome!

Note: By the way, if you're interested in these types of causal diagrams, there is a whole literature that uses them to study the identification of causal effects. The graphs are referred to formally as *directed acyclic graphs* or DAGs. The directed part just means that the arrows have directions, and the acyclic part that you cant have A causing B causing A causing B and so on, ad infinium. DAGs can be used to reason about causality using a set of rules referred to as the "do-calculus". A canonical text is Pearl

(2009). In the context of Figure 4.1, the fact that U influences both S and Y creates a so-called *backdoor path* from S to Y, generating an endogeneity problem. The DAG approach shows that we get identification of the effect of S on Y using Z because the only backdoor paths from Z to Y go through S.

4.1.3 Example: the returns to schooling

To see the magic of an IV, let's consider a particular application in which our treatment variable S_i is years of schooling, and we're interested in identifying the causal effect β_1 of increasing years of schooling by one. We assume this relationship is linear, so that we can write

$$Y_i = \beta_0 + \beta_1 S_i + \epsilon_i \tag{4.3}$$

where Y_i is log earnings. However, $Cov(S_i, \epsilon_i) \neq 0$, because earnings are also determined by "ability" A_i (e.g. measured by a standardized test), which is unobserved in our data (therefore, we cannot pursue a selection-on-observables strategy using A_i as a control variable). While years of schooling serves as S in Figure 4.1, ability A serves as the unobserved confounder A.

To make everything simple, we can suppose that the relationship between A and ϵ is also linear, i.e. that

$$\epsilon_i = \gamma A_i + \nu_i$$

for some coefficient γ and ν_i that is uncorrelated with both A_i and S_i and represents an error term in the relationship between A and ϵ. Given this, we can rewrite Eq. (4.3) as:

$$Y_i = \beta_0 + \beta_1 + \gamma A_i + \nu_i$$

which makes clear that if we did in fact observe ability A_i, we could estimate β_1 through multiple linear regression of Y on A and S, using A as a control variable.

Now suppose we have an instrument Z_i that provides an incentive for students to attend an additional year of schooling. For example, it could be a scholarship for university that is allocated to some students and not to others. If the scholarship Z_i is uncorrelated with ability A_i, then we have a valid IV, since then:

$$Cov(Z_i, \epsilon_i) = \gamma \cdot Cov(Z_i, A_i) + Cov(Z_i, \nu_i) = \gamma \cdot 0 + 0 = 0$$

The instrument is *relevant* provided that it actually works in incentivizing some students to increase their years of schooling, i.e. $Cov(Z_i, S_i) \neq 0$.

4.1.4 IV as a ratio of two regression coefficients

Recall from Equation 4.2 that if we have a valid IV for S in model (4.1), we can identify it through the quantity:

$$\beta_1 = \frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)}$$

which is estimable from data on (Y, S, Z).

Note that we could rewrite this as

$$\beta_1 = \frac{Cov(Y_i, Z_i)}{Var(Z_i)} \cdot \frac{Var(Z_i)}{Cov(S_i, Z_i)}$$

i.e., the coefficient from a simple linear regression of Y on Z, divided by the coefficient from a simple linear regression of S on Z. The first of these regressions is often referred to as the *reduced-form* regression:

$$Y_i = \rho_0 + \rho_1 Z_i + u_i \tag{4.4}$$

The regression coefficient that ρ is divided by is that of the so-called *first-stage* regression:

$$S_i = \pi_0 + \pi_1 Z_i + V_i \tag{4.5}$$

With this notation, $\beta_1 = \rho_1 / \pi_1$. We can make sense of this as follows. When we "wiggle" the instrument Z, this wiggles S by π_1, and each wiggle of S wiggles Y by β_1. Thus, each wiggle of Z is associated with $\rho_1 = \pi_1 \cdot \beta_1$ wiggles of Y. To determine β_1, we thus need to divide the reduced-form regression coefficient

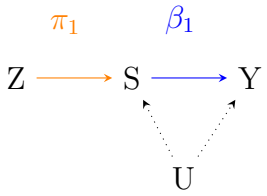


Figure 4.2: If we let ρ denote the coefficient from a regression of Y on Z, then $\beta_1 = \rho_1/\pi_1$.

ρ_1 by the first-stage regression coefficient π_1. This is depicted in terms of the causal graph from Section 4.1.2 Figure 4.2.

Note finally that if the instrument is a binary variable, then both the reduced-form and first-stage coefficients take the form of differences is means, and

$$\beta_1 = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]} \tag{4.6}$$

This expression is often referred to as a *Wald ratio*. With a binary instrument, one does not even need to run any regressions to implement IV: β_1 is identified just from computing four conditional expectations. In practice however, we want standard errors, and a regression type-framework will be a convenient way to get them. This leads to the so=called *two stage-least squares estimator*, which generalizes the idea of estimating $\beta_1 = \rho/\pi$ by applying OLS to both the first-stage and the reduced-form regressions in turn.

4.2 IV with heterogeneous treatment effects

4.2.1 Identifying the ATE when there is no selection on gains

Our introduction to instrumental variable in Section 4.1 has focused on a setting with homogeneous treatment effects, so that we can write $Y_i = \beta_0 + \beta_1 S_i + \epsilon_i$ where the parameter of interest is β_1 (in a setting with binary treatment D_i, we can let $S_i = D_i$ and then β_1 is equal to the treatment effect).

In most practical applications however, the assumption that every individual i has the same treatment effect is pretty unrealistic. After all, we only have to worry about endogoneiety/selection bias because $Y_i(0)$ differs between individuals (otherwise, it would have to be the case that $\mathbb{E}[Y_i(0)|D_i = 1] = \mathbb{E}[Y_i(0)|D_i = 0]$, since $Y_i(0)$ would be a degenerate random variable).

Fortunately, the results of the last section generalize nicely provided that there is no "selection on gains": that is, treatment effects are not correlated with treatment status.

Definition 4.1. *With a binary treatment D_i, we say that **no selection on gains** (NSOG) is satisfied if $\mathbb{E}[\Delta_i|D_i = 1] = \mathbb{E}[\Delta_i|D_i = 0]$, where $\Delta_i = Y_i(1) - Y_i(0)$. More generally, with treatment variable S_i with support $\mathcal{S}$, we say NSOG holds if $\mathbb{E}[\Delta_i|S_i = s]$ is the same for all values $s \in \mathcal{S}$.*

Let us consider a binary treatment, and further assume that $\mathbb{E}[\Delta_i|D_i = d, Z_i = z] = \mathbb{E}[\Delta_i|D_i = d]$. This is a natural assumption given the exclusion restriction depicted in Figure 4.1. Since Z only influences Y via D, knowing the value of Z does not change our expectations about potential outcomes once we've fixed a value of d. Thus altogether, NSOG and this exclusion assumption imply that

$$\mathbb{E}[\Delta_i|D_i = d, Z_i = z] = \Delta,$$

for all d and z, where we let $\Delta = \mathbb{E}[\Delta_i|D_i = 1] = \mathbb{E}[\Delta_i|D_i = 0]$. Notice that Δ is also equal to the average treatment effect.

Now to see how NSOG helps, let us generate an equation for the realized value of Y_i as before, but now allowing treatment effect heterogeneity:

$$\begin{aligned}
Y_i &= Y_i(0) + \Delta_i \cdot D_i \\
&= Y_i(0) + \Delta \cdot D_i + D_i \cdot (\Delta_i - \Delta) \\
&= \underbrace{\mathbb{E}[Y_i(0)]}_{\beta_0} + \underbrace{\Delta}_{\beta_1} \cdot D_i + \underbrace{Y_i(0) - \mathbb{E}[Y_i(0)] + D_i \cdot (\Delta_i - \Delta)}_{\epsilon_i} \\
&= \beta_0 + \beta_1 D_i + \epsilon_i
\end{aligned}$$

Now consider an instrument Z_i such that $Cov(Z_i, Y_i(0)) = 0$. By NSOG, $\mathbb{E}[D_i \cdot (\Delta_i - \Delta)] = \mathbb{E}[D_i] \cdot \mathbb{E}[\Delta_i - \Delta] = 0$, and hence

$$\begin{aligned} Cov(Z_i, \epsilon_i) &= \mathbb{E}[Z_i \cdot D_i \cdot (\Delta_i - \Delta)] \\ &= \sum_{d,z} P(D_i = d, Z_i = z) \cdot (\mathbb{E}[\Delta_i | D_i = d, Z_i = z] - \Delta) \\ &= \sum_{d,z} P(D_i = d, Z_i = z) \cdot (\mathbb{E}[\Delta_i | D_i = d] - \Delta) \\ &= \sum_{d,z} P(D_i = d, Z_i = z) \cdot (\Delta - \Delta) = 0 \end{aligned}$$

where we have used the additional "exclusion restriction" assumption that $\mathbb{E}[\Delta_i | D_i = d, Z_i = z] = \mathbb{E}[\Delta_i | D_i = d]$.

We have thus shown that $Cov(Z_i, \epsilon_i) = 0$ even in this extended model with heterogeneous treatment effects (and NSOG), in which ϵ_i captures heterogeneity both in $Y_i(0)$ and in treatment effects Δ_i. Thus, in this setting the IV estimand $\beta_1 = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)}$ from Eq (4.2) identifies the average treatment effect. For a generalization of this argument to the case of a multi-valued or continuous treatment, see Kolesár (2013), who refers to NSOG as "constant average treatment effects".

4.2.2 The local average treatment effects (LATE) model

The last section showed how our results that assume homogenous treatment effects generalize to a setting with no selection on gains (NSOG): when NSOG holds the IV estimand captures the average treatment effect. However, NSOG can itself be a strong assumption. For example, individual's may be inclined to choose the value of the value of treatment that benefits them the most (this kind of behavior is often referred to as a *Roy model*). In it's simplest form, a Roy model says that agents select into whichever value of treatment maximizes their outcome, the *same* outcome that we as a researcher are interested in. In this case, we would have

$$D_i = \mathbb{1}(Y_i(1) \geq Y_i(0)) = \mathbb{1}(\Delta_i \geq 0)$$

In this model, there's no way that NSOG would hold, because Δ_i will be negative for all untreated individuals, and positive for all treated individuals, thuse $Cov(\Delta_i, D_i) > 0$. One can extend the Roy model to allow idiosyncratic noise to individual's choices, e.g. $\mathbb{1}(\Delta_i \geq U_i)$ where U_i varies by individual. But the message is the same: NSOG is a restrictive assumption for settings in which individual's self select into treatment. Can we relax our assumptions even further, to allow for individuals to select on their individual gains?

The answer is yes, and in seeing this we encounter the canonical "local average treatment effects" framework for IV introduced by Imbens and Angrist (1994). The LATE model provides a key result for the interpretation of IV estimates. In particular, it shows that under very minimal assumptions, we can understand instrumental variables as telling us about treatment effects for a certain subset of the population of interest, whose value of treatment is responsive to the value of the IV. We call these individuals "compliers", and the main result is that IV identifies the "local" average treatment effect, just among these compliers.

To present the LATE model, it is helpful to use the language of a randomized controlled trial in which there is imperfect take-up of the treatment. In this setting, both the treatment and the instrument are binary, taking values of 0 or 1. As before, those with $D_i = 1$ are "treated" and those with $D_i = 0$ are "untreated", representing a control group. Those with $Z_i = 1$ are "assigned" to the treatment arm of the experiment, and those with $Z_i = 0$ are instead "assigned" to the control arm. While this language is motivated by a true experiment in which some randomization device is used to actually assign individuals to treatment or control, we also use this language in non-experimental settings. The reason for this is that the underlying assumptions made in a clinical trial (in which not all individuals who are assigned to treatment actually become treated) are sometimes plausible in non-experimental settings as well.

For example suppose some students are offered a scholarship that will pay for their university tuition, indicated by $Z_i = 1$, and selection for the scholarship is random. Those individuals who are not offered the scholarship instead have $Z_i = 0$. However, not all individuals who are offered the scholarship end up going to university. We let $D_i = 1$ indicate that i goes to university, and $D_i = 0$ if they do not. Since not all scholarship recipients go to university, we have that $P(D_i = 0 | Z_i = 1) > 0$. On the other hand, many students enroll in university even without the scholarship, so that $P(D_i = 01 | Z_i = 0) > 0$ as well.

To make sense of these patterns of behavior, let us introduce the notion of *potential treatments* $D_i(z)$. Recall that potential outcomes $Y_i(d)$ say what value our outcome variable Y would take for individual i if their treatment status were $d \in \{0, 1\}$. Potential treatments tell us what value our *treatment variable* D_i would take if their *instrument value* were equal to z. We can make sense of this notation through Figure 4.1: Z has a causal effect on the treatment S or D, which in turn has a causal effect on Y. Potential treatments are simply potential outcomes for the first arrow from the instrument to the treatment. An individual's actual, or *realized* treatment D_i is related to their assigned value of Z_i through $D_i = D_i(Z_i)$.

To formalize our notion that Z *only* affects Y through D, let us extend out potential outcomes notation to $Y_i(d, z)$, which says what the value of Y would be for individual i if their treatment value were d *and* their instrument value were z. We'll assume that $Y_i(d, z)$ does not depend on z, in which case we can use the simpler notation $Y_i(d)$ as before.

For each z, $D_i(z)$ is either equal to zero or is equal to one. Thus, there are four conceivable groups within the population, which we will refer to as "selection groups". The four selection groups are given in the table below. The first group in this table, the "never-takers", do not go to university if they do

Name	**Meaning**
"never-takers"	$D_i(0) = 0$ & $D_i(1) = 0$
"always-takers"	$D_i(0) = 1$ & $D_i(1) = 1$
"compliers"	$D_i(0) = 0$ & $D_i(1) = 1$
~~"defiers"~~	~~$D_i(0) = 1$ & $D_i(1) = 0$~~

not get the scholarship, and they still do not attend university even if they are offered the scholarship. "Always-takers", on the other hand, go to university even if they do not get the scholarship. "Compliers" go to university only if they receive the scholarship, i.e. $D_i(0) = 0$ and $D_i(1) = 1$. Finally, one group in the table is crossed out: "defiers" would only go to university of they *did not* receive a scholarship, and would not go if they did receive the scholarship (i.e. $D_i(0) = 1$ and $D_i(1) = 0$). The LATE model assumes that there are no such defiers. This seems like a fairly innocuous assumption in the context of our hypothetical scholarship program: who would go to university only if it were not made cheaper for them?

Note that in the context of a true clinical trial, we might also be willing to rule out always-takers. If an experiment is testing a new drug that is only available through the experiment, there would not be any individuals who managed to be treated when not assigned to the treatment arm of the study (since they can't obtain the drug any other way). However, always-takers are an important segment of the population in most non-experimental settings, and the LATE model allows them to be present.

Given the potential-treatments notation, the LATE model assumptions are:

1. **IV Independence:** $(Y_i(d, z), D_i(z)) \perp Z_i$ for all d, z
2. **Exclusion:** $Y_i(d, z) = Y_i(d)$ for all d, z
3. **Monotonicity:** $D_i(1) \geq D_i(0)$
4. **Relevance:** $P(D_i(1) > D_i(0)) > 0$

The independence assumption says that the instrument Z is as good as randomly assigned, in the sense that it is statistically independent of both potential outcomes and potential treatments. Exclusion states $Y_i(d, z)$ doesn't change with z if d is held fixed, reflecting the idea that Z only effects Y *through* D. We can combine Assumptions 1 and 2 to write

$$(Y_i(0), Y_i(1), D_i(0), D_i(1)) \perp Z_i$$

and think of this as our *instrument validity* assumption. The fourth condition says that a positive proportion of the population are compliers, and plays the role of the *instrument relevance* condition from Section 4.1. Note that $D_i(1) > D_i(0)$ is the same as $D_i(0) = 0, D_i(1) = 1$, which indicates that i is a complier.

Our final assumption, "monotonicity", is a new one that we did not need in the homogenous effects or NSOG models. It states that the causal effect of the instrument on treatment status is to move all

units weakly in the *same direction*. That is, we can't have some individual i for whom $D_i(0) = 0$ and $D_i(1) = 1$, and some other individual j for whom $D_j(0) = 1$ and $D_j(1) = 0$. These latter individuals would be "defiers", and we rule them out by assumption. Note that the direction of the weak inequality in Assumption 3 is arbitrary. If the instrument taking a value of one were to move all units *out* of treatment or not at all (i.e. there are defiers but no compliers), we could simply redefine the instrument by swapping the labels of $z = 0$ and $z = 1$. What's important is that there are not *two-way flows*, both into and out of treatment, when it is moved from one value to the other.

The famous result of Imbens and Angrist (1994) is that under the LATE assumptions 1-3, the Wald ratio of Eq. (4.6) identifies the average treatment effect among the compliers, often referred to as "the LATE":

$$\frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]} = \mathbb{E}[\Delta_i|D_i(1) > D_i(0)] \tag{4.7}$$

where $\Delta_i = Y_i(1) - Y_i(0)$ and note that the event $D_i(1) > D_i(0)$ is the same as $D_i(0) = 0, D_i(1) = 1$, which indicates that i is a complier. Note that the expression above is also equal to $Cov(Y_i, Z_i)/Cov(D_i, Z_i)$. Thus the expression that gives the LATE coincides with our expression for β_1, the average treatment effect under NSOG, and the common treatment effect when treatment effects are fully homogenous.

What if monotonicity does not hold? You might be curious how robust the LATE theorem is to departures from the important assumption of monotonicity—"no defiers". We know from Section 4.2.1 that if we have treatment effect heterogeneity but no selection on gains, the LHS of Eq. (4.7) captures the ATE, even if there are defiers. This includes homogenous treatment effects as a special case, in which the LHS of Eq. (4.7) captures the homogenous treatment effect.

But is there any thing we can say if treatment effects are heterogenous and we have selection on gains? Chaisemartin (2017) shows that the answer is yes, under a more general assumption he calls the "compliers-defiers" assumption. I will not state the general complier-defiers assumption here, but will just give you my favorite sufficient condition for it. If for any value y there are more compliers having treatment effect y then there are defiers having treatment effect y, i.e. $P(i \text{ is complier}|\Delta_i = y) \geq P(i \text{ is defier}|\Delta_i = y)$, then the LHS of Eq. (4.7) captures a local average treatment effect among a *subgroup* of the compliers, whose size is $P(i \text{ is complier}) - P(i \text{ is defier})$. This suggests that if there are some defiers in the population, but not too many, we can still interpret the IV estimand as an average treatment effect among the "surviving compliers" within this subgroup.

Proof of the LATE theorem

We will now see why Eq. (4.7) holds, proceeding in several steps. Let us denote $p_n = P(i \text{ is a never-taker})$, $p_a = P(i \text{ is an always-taker})$, and $p_c = P(i \text{ is a complier})$. Denote the relative proportions of the three selection groups in our population. By monotonicity, we know that $p_d = P(i \text{ is a defier}) = 0$ and so $p_n + p_a + p_c = 1$.

Note that $D_i(1)$ and $D_i(0)$ are each random variables, and by the independence assumption $P(D_i(0) = d, D_i(1) = d'|Z_i = z)$ doesn't depend on z, for any values $d, d' \in \{0, 1\}$. With this in mind, consider the quantity $E[D_i|Z_i = 1]$. By independence and then the law of iterated expectations over both $D_i(1)$ and $D_i(0)$:

$$\begin{aligned}
\mathbb{E}[D_i|Z_i = 1] &= \mathbb{E}[D_i(1)|Z_i = 1] = \mathbb{E}[D_i(1)] \qquad (4.8)\\
&= p_n \cdot \mathbb{E}[D_i|D_i(0) = D_i(1) = 0] + p_a \cdot \mathbb{E}[D_i|D_i(0) = D_i(1) = 1]\\
&\qquad + p_c \cdot \mathbb{E}[D_i|D_i(1) = 1, D_i(0) = 0]\\
&= p_n \cdot \mathbb{E}[D_i(1)|D_i(0) = D_i(1) = 0] + p_a \cdot \mathbb{E}[D_i(1)|D_i(0) = D_i(1) = 1]\\
&\qquad + p_c \cdot \mathbb{E}[D_i(1)|D_i(1) = 1, D_i(0) = 0]\\
&= p_n \cdot 0 + p_a \cdot 1 + p_c \cdot 1\\
&= p_a + p_c \qquad (4.9)
\end{aligned}$$

In the second equation, we have used that $D_i = D_i(Z_i)$, and then independence to remove the conditioning on Z_i. We can see from the above that the share of individuals that go to university, among

those offered the scholarship, tells us the proportion of the population that are either always-takers or are compliers. By similar steps:

$$\begin{aligned}
\mathbb{E}[D_i|Z_i=0] &= \mathbb{E}[D_i(0)|Z_i=0] = \mathbb{E}[D_i(0)] && (4.10)\\
&&& (4.11)\\
&= p_n \cdot \mathbb{E}[D_i|D_i(0)=D_i(1)=0] + p_a \cdot \mathbb{E}[D_i|D_i(0)=D_i(1)=1] \\
&\qquad + p_c \cdot \mathbb{E}[D_i|D_i(1)=1, D_i(0)=0] \\
&= p_n \cdot \mathbb{E}[D_i(0)|D_i(0)=D_i(1)=0] + p_a \cdot \mathbb{E}[D_i(0)|D_i(0)=D_i(1)=1] \\
&\qquad + p_c \cdot \mathbb{E}[D_i(0)|D_i(1)=1, D_i(0)=0] \\
&= p_n \cdot 0 + p_a \cdot 1 + p_c \cdot 0 \\
&= p_a && (4.12)
\end{aligned}$$

Taking the difference between (4.9) and (4.12), we see that we can identify the share of compliers in the population:

$$\mathbb{E}[D_i|Z_i=1] - \mathbb{E}[D_i|Z_i=0] = p_c \tag{4.13}$$

We'll now use an analgous set of steps to show that the numerator of Eq. (4.7) is equal to $p_c \cdot \mathbb{E}[\Delta_i|D_i(1) > D_i(0)]$, thus establishing the result provided that $p_c > 0$ (Assumption 4).

Consider the quantity $\mathbb{E}[Y_i|Z_i=1]$. By the law of iterated expectations and then independence:

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i=1] = p_n &\cdot \mathbb{E}[Y_i|Z_i=1, D_i(0)=0, D_i(1)=0] \\
+ p_a &\cdot \mathbb{E}[Y_i|Z_i=1, D_i(0)=1, D_i(1)=1] \\
+ p_c &\cdot \mathbb{E}[Y_i|Z_i=1, D_i(0)=0, D_i(1)=1]
\end{aligned}$$

Now, having conditioned on Z_i as well as a unit's potential treatments, we know their realized treatment $D_i = D_i(Z_i)$, and hence which potential outcome we are observing in Y_i:

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i=1] = p_n &\cdot \mathbb{E}[Y_i(0)|Z_i=1, D_i(0)=0, D_i(1)=0] \\
+ p_a &\cdot \mathbb{E}[Y_i(1)|Z_i=1, D_i(0)=1, D_i(1)=1] \\
+ p_c &\cdot \mathbb{E}[Y_i(1)|Z_i=1, D_i(0)=0, D_i(1)=1]
\end{aligned}$$

The great thing about having replaced the Y_i's by the corresponding potential outcomes is that the potential outcomes themselves are independent of the instrument Z_i, so we can drop the conditioning on Z_i, as before when we were considering $\mathbb{E}[D_i|Z_i=1]$.[1] Thus:

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i=1] = p_n &\cdot \mathbb{E}[Y_i(0)|D_i(0)=0, D_i(1)=0] \\
+ p_a &\cdot \mathbb{E}[Y_i(1)|D_i(0)=1, D_i(1)=1] \\
+ p_c &\cdot \mathbb{E}[Y_i(1)|D_i(0)=0, D_i(1)=1] \qquad (4.14)
\end{aligned}$$

Now following the same logic for $\mathbb{E}[Y_i|Z_i=0]$

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i=0] &= p_n \cdot \mathbb{E}[Y_i(0)|Z_i=0, D_i(0)=0, D_i(1)=0] \\
&\quad + p_a \cdot \mathbb{E}[Y_i(1)|Z_i=0, D_i(0)=1, D_i(1)=1] \\
&\quad + p_c \cdot \mathbb{E}[Y_i(0)|Z_i=0, D_i(0)=0, D_i(1)=1] \\
&= p_n \cdot \mathbb{E}[Y_i(0)|D_i(0)=0, D_i(1)=0] \\
&\quad + p_a \cdot \mathbb{E}[Y_i(1)|D_i(0)=1, D_i(1)=1] \\
&\quad + p_c \cdot \mathbb{E}[Y_i(0)|D_i(0)=0, D_i(1)=1] \qquad (4.15)
\end{aligned}$$

[1] Consider a term $\mathbb{E}[Y_i(d)|Z_i=z, D_i(0)=d, D_i(1)=d^*]$ with specific values of d, z and z^*. Really what we're using is the joint independence condition $(Y_i(d), D_i(0), D_i(1)) \perp Z_i$. Assume for simplicity that $Y_i(d)$ has discrete support. Then $\mathbb{E}[Y_i(d)|Z_i=z, D_i(0)=d, D_i(1)=d^*]$ can be written as

$$\sum_y y P(Y_i(d)=y|Z_i=z, D_i(0)=d, D_i(1)=d^*) = \sum_y y \frac{P(Y_i(d)=y, Z_i=z, D_i(0)=d, D_i(1)=d^*)}{P(Z_i=z, D_i(0)=d, D_i(1)=d^*)}$$

By the independence condition this is equal to

$$\sum_y y \frac{\cancel{P(Z_i=z)}P(Y_i(d)=y, D_i(0)=d, D_i(1)=d^*)}{\cancel{P(Z_i=z)}P(D_i(0)=d, D_i(1)=d^*)} = \sum_y y P(Y_i(d)=y|D_i(0)=d, D_i(1)=d^*) = \mathbb{E}[Y_i(d)|D_i(0)=d, D_i(1)=d^*]$$

The always-taker and never-taker terms cancel out when we consider the difference between (4.14) and (4.15), but the compliers term remains:

$$\begin{aligned}\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0] &= p_c\left(\mathbb{E}[Y_i(1)|D_i(0)=0,D_i(1)=1]-\mathbb{E}[Y_i(0)|D_i(0)=0,D_i(1)=1]\right)\\ &= p_c\cdot\mathbb{E}[Y_i(1)-Y_i(0)|D_i(0)=0,D_i(1)=1]\end{aligned}$$

This establishes Eq. 4.7.

4.2.3 Covariates and characterizing the complier population†

We've seen from Eq. (4.13) that we can count how many compliers p_c there are in the population, since $p_c\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]$ and the RHS is identified. However, we can't say exactly *who* those compliers are. Because of the fundamental problem of causal inference, we only see $D_i(0)$ or $D_i(1)$ for a given individual (depending on their value of Z_i), and never both. Since the LATE result tells us that IV only identifies the average treatment effect among compliers (and not the whole population), understanding who the compliers are would be very useful in making sense of whether they are representative of the population as a whole, rather than being some kind of special group that is not that interesting.

It turns out that there is more we can say about the compliers in the LATE model, if we have covariates X_i at our disposal. In particular, we can estimate average complier characteristics $\mathbb{E}[X_i|D_i(1)>D_i(0)]$. These can then be compared with e.g. $\mathbb{E}[X_i]$ to assess how different the compliers are from the population as a whole, along the dimensions captured by X_i.

For this result, we need to augment the LATE model with a *conditional independence* assumption, rather than the unconditional independence assumption in Section 4.2.2:

Conditional IV independence: $\{(Y_i(d),D_i(z))\perp Z_i\}\,|X_i$ for all d, z.

Conditional independence is a natural assumption if we have a randomly assigned instrument, and X_i is a collection of baseline characteristics X_i that are not themselves affected by treatment (e.g. age). If on the other hand X_i is affected by treatment, it could represent a "bad-control" and $\{(Y_i(d),D_i(z))\perp Z_i\}\,|X_i$ may fail even if $(Y_i(d),D_i(z))\perp Z_i$ holds.[2]

Another context in which we might have conditional IV independence is when treatment assignment Z_i is not unconditionally independent of potential outcomes/potential treatments, but is instead random only when conditioning on the X_i. This is what we would expect in a *stratified* randomized controlled trial (RCT), in which the probability of $Z_i=1$ may differ between strata $X_i=x$, but is random within each one.

Abadie (2003) shows that in the LATE model with conditional independence:

$$\mathbb{E}[X_i|D_i(1)>D_i(0)]=\mathbb{E}[\kappa_i\cdot X_i]/\mathbb{E}[\kappa_i]\quad\text{where}\quad\kappa_i=1-\frac{D_i\cdot\mathbb{1}(Z_i=0)}{P(Z_i=0|X_i)}-\frac{(1-D_i)\cdot\mathbb{1}(Z_i=1)}{P(Z_i=1|X_i)}$$

This result shows that the complier-mean of covariate X_i can be estimated so long as we can estimate the quantity κ_i for each i. This in turn requires estimating $P(Z_i=1|X_i=x)$, the conditional propensity of treatment assignment among individuals with $X_i=x$. In a stratified RCT, this quantity may be known ex-ante based on the experimental procedure (e.g. men are assigned to treatment with probability 0.4 and women are assigned to treatment with probability 0.7). In most non-experimental applications, it would have to be estimated from the data, but it is identified given an i.i.d sample of (Y_i,D_i,Z_i,X_i).

Another application of conditional IV independence is to perform heterogeneity analysis on the LATE. Conditioning all of the expectations in Eq. (4.6) on $X_i=x$, we can identify a *conditional LATE*

$$\mathbb{E}[\Delta_i|D_i(1)>D_i(0),X_i=x]=\frac{\mathbb{E}[Y_i|Z_i=1,X_i=x]-\mathbb{E}[Y_i|Z_i=0,X_i=x]}{\mathbb{E}[D_i|Z_i=1,X_i=x]-\mathbb{E}[D_i|Z_i=0,X_i=x]}\tag{4.16}$$

among those individuals who have observable characteristics $X_i=x$, and are also compliers.

[2] In the presence of covariates X_i we mantain our other assumptions 2-4 from Section 4.2.2 also conditional on X_i. In particular, this requires for the direction of "compliance" to be the same for all values of X_i, i.e. we can't have $P(D_i(1)\geq D_i(0)|X_i=x)=1$ for some value x while $P(D_i(1)\leq D_i(0)|X_i=x)=1$ for some other value x'. Our relevance assumption will also be strengthened to $0<P(Z_i=1|X_i)<1$ with probability one, an analog of the common support condition introduced in the context of selection-on-observables. See Abadie (2003) for details.

4.2.4 Connection to latent-index models*

To get the LATE theorem, we have made assumptions about potential outcomes/treatments and their distributions, but we haven't committed to an explicit model of how these outcomes come about. An alternative/complimentary approach might characterize the selection process by constructing a "structural" model of who chooses treatment.

For instance, we might think that each unit i is a utility-maximizing agent who's utility is

$$u_i = \begin{cases} \gamma_0 + \gamma_1 Z_i & \text{if they receive treatment } (i.e.\ D_i = 1) \\ U_i & \text{if they don't } (i.e.\ D_i = 0) \end{cases}$$

Agents will choose treatment when it gives them higher utility, and so they will choose:

$$D_i = \mathbb{1}(\gamma_0 + \gamma_1 Z_i > U_i)$$

where we've assumed that ties go to non-treatment.

In this model, heterogeneity among D_i comes from agent's having different values of the instrument, as well as a different "random utility" U_i in the non-treatment state. If γ_1 is positive, the instrument incents individuals towards treatment, since:

$$D_i(0) = \mathbb{1}(\gamma_0 > U_i) \quad \text{and} \quad D_i(1) = \mathbb{1}(\gamma_0 + \gamma_1 > U_i)$$

so the monotonicity condition is immediately satisfied: $D_i(1) \geq D_i(0)$. If γ_1 were negative, we'd have monotonicity in the other direction.

An important result of Vytlacil (2002) establishes that the LATE model is in fact *equivalent* to a latent-index model of treatment along with standard IV assumptions. In other words, whenever you are willing to accept that monotonicity holds (as well as the other three assumptions from Section 4.2.2), there exists a latent index model of the form

$$D_i(z) = \mathbb{1}(g(z) \geq U_i) \tag{4.17}$$

that can represent the setting equally well. This result holds up even if the instrument Z_i is not binary (e.g. it takes on many values), or may even be a vector. In the above, $g(z)$ is some function of the instrument(s), common to all i. In our example, we took g to be linear, i.e. $g(z) = \gamma_0 + \gamma_1 \cdot z$ for some γ_0 and γ_1.

4.2.5 Beyond a binary instrument: many LATE's and marginal treatment effects*

Our discussion of the LATE model has focused on a setting in which both the treatment and the outcome variable are binary, taking just two values. This section maintains the setup of a binary treatment but now considers an instrument Z_i that may take on many values, or be continuous. If we have multiple instruments Z_1, Z_2, etc. for our treatment, we can take Z_i to be a *vector* $Z_i = (Z_{1i}, Z_{2i}, \dots)$ comprised of all of them. Let $\mathcal{Z}$ be the support of Z_i. In this context we can generalize our IV monotonicity assumption as follows:

IV monotonicity for arbitrary Z: For any z, z' in $\mathcal{Z}$, either $D_i(z') \geq D_i(z)$ for all i or $D_i(z) \leq D_i(z')$ for all i.

This generalized monotonicity assumption says that if we take any two values z and z' of our instrument(s), all individuals i move in the same direction (either into or out of treatment) when their counterfactual value of Z is changed from z to z'. Consider for example an instrument Z_i that takes on four values. Monotonicity means that we can choose labels for these values z_1, z_2, z_3 and z_4 such that if any individual i would take treatment if $Z_i = z_j$, they would also take treatment if $Z_i = z_{j+1}$. We might represent this by a "chain" of instrument values:

In this diagram, an arrow from z to z' means that anyone who would take treatment at z would also take treatment at z' (don't confuse this with the DAG arrows from Figure 4.1! In that case, an arrow represents the presence of a causal effect). Since $D_i(z_4) \geq D_i(z_3)$ and $D_i(z_3) \geq D_i(z_2)$ imply $D_i(z_3) \geq D_i(z_2)$, we can connect these arrows and lay all of the values z_j out in a chain.

As an example, suppose that a the tuitions at various universities in the country of Econometrica can only be one of four values $z_4 = \$0$, $z_3 = \$1000$, $z_2 = \$5000$, or $z_1 = \$10,000$ (alternatively, we can think of these instrument values as reflecting tuition ranges). Tuition is randomly assigned across students, and we consider this tuition to be an instrument for whether students in Econometrica attend university. The interpretation of the above figure is that any student who would go to university if it cost $\$10,000$ would still attend if it was $\$5,000$, or if it were $\$1,000$ or free. Likewise, any student who would attend university if it were $\$5,000$ would still attend if it were $\$1,000$, or free. And so on. The idea is that while student may differ in many ways that influence their decision of whether to attend university, any one student i's treatment status would be monotonically decreasing in counterfactual tuition rates.

> *Note:* IV monotonicity can be a very strong assumption when Z_i is a vector. For instance, if we have two binary instruments for university attendance (tuition and proximity), monotonicity requires that either i) all students who would go to university if it were close but expensive would also go if it were instead far but cheap, or ii) all students who would go to university if it were far but cheap would also go if it were instead close and expensive. In this context a more natural monotonicity assumption might be that all students who would go to university if it were expensive would also go if it were cheap (regardless of whether it is close or far), and all students who would go to university of it were far would also go if it were close (regardless of whether it is cheap or expensive). See Mogstad et al. (2021) shows that under this more natural assumption, referred to as *partial monotonicity* or *vector monotonicity*, conventional IV estimands based upon the typical IV monotonicity assumption can be misleading. Goff (2020) shows how certain local average treatment effects are nevertheless identified, and how one can estimate them.

A simple extension of the LATE result given in Section 4.2.2 shows that under the general IV monotonicity assumption, for any pair of instrument values z' and z such that $P(D_i(z') > D_i(z)) > 0$:

$$\frac{\mathbb{E}[Y_i|Z_i = z'] - \mathbb{E}[Y_i|Z_i = z]}{\mathbb{E}[D_i|Z_i = z'] - \mathbb{E}[D_i|Z_i = z]} = \mathbb{E}[\Delta_i|D_i(z') > D_i(z)] \tag{4.18}$$

Eq. (4.7) in the binary instrument setting is a special case of this when $z' = 1$ and $z = 0$. (4.18) is given as Theorem 1 in Imbens and Angrist (1994).

Consider for example a discrete instrument (for example, if Z_i is the number of years of schooling that i's mother completed). Then Eq. (4.18) shows that we can identify a local average treatment effect along each link in the chain: the local average treatment effect among individuals who would go to university if their mother had 11 years of education but not 10, the local average treatment effect among individuals who would go to university if their mother had 12 years of education but not 11, the local average treatment effect among individuals who would go to university if their mother had 13 years of education but not 12, and so on.

Theorem 2 in Imbens and Angrist (1994) shows that with discrete instruments, we can aggregate over all of these individual LATE's by using $\mathcal{P}(Z_i)$ as our instrument, where $\mathcal{P}(z) := \mathbb{E}[D_i|Z_i = z]$ is the propensity score function. In particular, $Cov(Y_i, \mathcal{P}(Z_i))/Cov(D_i, \mathcal{P}(Z_i))$ yields a weighted average of $\mathbb{E}[\Delta_i|D_i(z_{j+1}) > D_i(z_z)]$ across the j. Under certain assumptions, this estimand is exactly what is captured by the two-stage least squares estimator, which we'll study in Section 4.3.

When we have access to a continuous instrument, our many LATE's from Eq. (4.18) can yield a continuum of local average treatment effects, referred to as *marginal treatment effects*. Suppose we have a single continuous instrument, and that monotonicity holds in the direction of increasing values of that instrument: $D_i(z') \geq D_i(z)$ for any $z' > z$. Then we can take the limit of Eq. (4.18) as $z' \downarrow z$ to obtain:

$$\frac{\frac{d}{dz}\mathbb{E}[Y_i|Z_i = z]}{\frac{d}{dz}\mathbb{E}[D_i|Z_i = z]} = \lim_{z' \downarrow z} \frac{\mathbb{E}[Y_i|Z_i = z'] - \mathbb{E}[Y_i|Z_i = z]}{\mathbb{E}[D_i|Z_i = z'] - \mathbb{E}[D_i|Z_i = z]} = \mathbb{E}[\Delta_i|z = \inf_z : D_i(z) = 1] \tag{4.19}$$

The first equality divides both the numerator and denominator by $z' - z$ and uses the definition of the limit. Note that the LHS of (4.19) is identified from a regression of the outcome on the instrument, while the denominator is identified from a regression of the treatment on the instrument. Neither of these regression functions are guaranteed to be linear however, so in practice flexible or non-parametric methods should be used to estimate them. The RHS of (4.19) is the average treatment effect among individuals for whom the first value of z for which they begin to take treatment is z.

One can thus estimate $\mathbb{E}[\Delta_i | z = \inf_z : D_i(z) = 1]$ as a function of z, with a single continuous instrument. When Z_i is a vector, we need some way to collapse our instruments into a single scalar to define an analagous funciton. This can be done through the propensity score function, since $\mathcal{P}(z') \geq \mathcal{P}(z)$ exactly when $D_i(z') \geq D_i(z)$. Since $\mathcal{P}(z) = \mathbb{E}[D_i | Z_i = z]$ takes values on the unit interval, we define the *marginal treatment effect function* $MTE(p)$ as a function of $p \in (0, 1)$:

$$MTE(p) = \mathbb{E}[\Delta_i | U_i = p]$$

where we can for each individual i define U_i to be $\inf_{z:D_i(z)=1} \mathcal{P}(z)$. For the MTE to be defined at p, there must exist a value $z \in \mathcal{Z}$ such that the proportion of individuals who would take treatment at $Z_i = z$ is exactly p, and there must be individuals for whom z is the "first" value of Z at which they take treatment (where "first" is measured increasing order of the propensity score). In this case $MTE(p) = \frac{\frac{d}{dz}\mathbb{E}[Y_i | P(Z_i)=p]}{\frac{d}{dz}\mathbb{E}[D_i | P(Z_i)=p]}$, which can be estimated from the data.

In understanding the MTE function, the scalar U_i can be interpreted as a latent "reluctance" against treatment. An individual with a higher value of U_i requires a "higher" value of the instrument(s) to take treatment. In fact, the conventional approach in the marginal treatment effects literature is to define U_i explicitly as a relative cost of treatment in a latent index model, such as the one considered in Section 4.2.4, e.g.

$$D_i(z) = \mathbb{1}(\mathcal{P}(z) \geq U_i)$$

where without loss of generality the function g in Eq. (4.17) can be taken to be the propensity score function. See e.g. Heckman and Vytlacil (2005) for details. My definition $U_i = \inf_{z:D_i(z)=1} \mathcal{P}(z)$ is unconventional, but is equivalent if one starts from the IV monotonicity assumption rather than from an explicit selection model Eq. like Eq. (4.17).

4.2.6 Potential outcome distributions and quantile treatment effects in the LATE model*

Average treatment effects are a convenient and intuitive summary of heterogeneous treatment effects. In the proceeding sections, we've seen how local average treatment effects are identified in IV settings (with a binary treatment), whether we have a single binary instrument or even a continuous instrument or collection of instruments.

But when treatment effects Δ_i are highly heterogeneous within the population of compliers, the average could be misleading, even when it is conditioned on covariates as in Eq. (4.16). In the extreme case, imagine that treatment has a huge effect just for some small subgroup of the compliers. Then we might see a substantially positive LATE, even if treatment has a very small or even negative effect for most of the compliers. Is there any way to empirically distinguish this case from one in which all the compliers had the same treatment effect?

It turns out that we have a great tool at our disposal to "move beyond the mean" – under the standard LATE assumptions of independence, exclusion, and monotonicity, we can actually determine the effect of treatment on the whole distribution of Y among compliers.

This generalizes the discussion in Section 1.8 in which we assumed random assignment. Recall that

$$E[\mathbb{1}(Y_i \leq y) | D_i = 1] = P(Y_i(1) \leq y | D_i = 1) = F_{Y(1)|D=1}(y)$$

Thus the LHS, which can be estimated from the data, tells us something about the conditional distribution of the treated potential outcome. Under random assignment, this is in turn equal to $F_{Y(1)}(y)$, and a similar argument let's us identify the CDF of $Y(0)$.

With an instrumental variable, we can only estimate the distribution of each potential outcome among compliers, and not their unconditional distributions as with random assignment. To do this, we can make use of a general result from Abadie (2002) that also underlies our ability to capture average complier characteristics (Section 4.2.7). In particular, Lemma 2.1 of Abadie (2002) shows that

Lemma. *Let $g(y)$ be a function and make the standard LATE model assumptions. Then:*

$$E[g(Y_i(1))|D_{1i} > D_{0i}] = \frac{E[D_i g(Y_i)|Z_i = 1] - E[D_i g(Y_i)|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

and

$$E[g(Y_i(0))|D_{1i} > D_{0i}] = \frac{E[(1 - D_i)g(Y_i)|Z_i = 1] - E[(1 - D_i)g(Y_i)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]}$$

The result implies that if we pick some possible value y for Y_i, and let $g(Y_i) = \mathbb{1}[Y_i \leq y]$, then the CDFs of $Y(0)$ and $Y(1)$ conditional on being a complier are each identified:

$$F_{Y(1)|D_1>D_0}(y) = \frac{E[D_i \mathbb{1}(Y_i \leq y)|Z_i = 1] - E[D_i \mathbb{1}(Y_i \leq y)|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

and

$$F_{Y(0)|D_1>D_0}(y) = \frac{E[(1 - D_i)\mathbb{1}(Y_i \leq y)|Z_i = 1] - E[(1 - D_i)\mathbb{1}(Y_i \leq y)|Z_i = 0]}{E[(1 - D_i)|Z_i = 1] - E[(1 - D_i)|Z_i = 0]}$$

where for clarity, we for each $d \in \{0, 1\}$ denote $P(Y_i(d) \leq y|D_i(1) > D_i(0))$ by $F_{Y(d)|D_1>D_0}(y)$.

The RHS of the above two equations can be estimated from the data for each value of y. If we repeat this computation for all values of y, then we know the whole distribution function of each potential outcome, conditional on being a complier.

Note that this type of result isn't specific to the IV research design: you might be interested to know that something analogous can also be done in an RDD setup (see Frandsen et al. 2012), and under more complicated assumptions in a difference-in-differences design too (Callaway 2015).

One thing that having $F_{Y_1|D_1>D_0}(y)$ and $F_{Y_0|D_1>D_0}(y)$ lets us compute is so-called *quantile treatment effects* (QTEs) among the compliers. For notational simplicity, let's drop the conditioning on being a complier: $D_{i1} > D_{i0}$. The (unconditional) QTE is defined as

$$QTE(u) = F_1^{-1}(u) - F_0^{-1}(u)$$

where F_d^{-1} is the quantile function associated with potential outcome Y_d: $F_d^{-1}(u) = \inf\{y : P(Y_{id} \leq y) \geq u\}$ is the u^{th} quantile of Y_d, and u is a specified quantile level $u \in (0, 1)$.

Note that the QTEs are causal: they tell us about the difference between the distribution of $Y(1)$ and $Y(0)$ (as opposed to the distributions $Y(1)|D_i = 1$ and $Y(0)|D_i = 0$, which might be confounded by selection/endogeneity). Nevertheless, the QTEs do not tell us directly about the individual treatment effects Δ_i or their distribution, without further assumptions. The reason is that unlike the expectation function, the quantile function is not linear–thus: $QTE_i(u) \neq F_\Delta^{-1}(u)$.

There is a notable exception: if we assume that each students' *rank* were the same in both the treated and untreated distributions: $F_0(Y_{i(0)}) = F_1(Y_i(1))$ for all i, then the u-quantile treatment effect is equal to the treatment effect for a student with rank u. However, this is a strong assumption (referred to as *rank invariance*) that's hard to justify in general. Without additional assumptions such as rank invariance, the marginal distributions $F_1(y)$ and $F_0(y)$ do generally place bounds on the distribution of treatment effects, which are sometimes informative. See for example Fan and Park (2009).

4.2.7 The LATE framework beyond a binary treatment*

We've so far focused on a binary treatment variable when considering IV with heterogeneous treatment effects. However, some of the results of this section carry over to treatments that are not binary.

Angrist and Imbens (1995) for example studies the LATE model in which we have an ordered discrete treatment variable S_i, such as years of schooling, and a binary instrument. Suppose that S takes as values the integers 0 to J, where $Y_i(s)$ denotes potential outcomes when $S_i = s$. We let $S_i(z)$ denote our potential treatments, depending on instrument value $z \in \{0, 1\}$. Analogously to the monotonicity assumption $D_i(1) \geq D_i(0)$ from the binary-treatment case, assume:

IV monotonicity for ordered discrete treatment and binary instrument: For all i: $S_i(1) \geq S_i(0)$.

Note that we could instead accommodate $S_i(1) \leq S_i(0)$ for all i by simply re-labeling the instrument values. Angrist and Imbens (1995) show that under the above monotonicity assumption:

$$\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[S_i|Z_i=1]-\mathbb{E}[S_i|Z_i=0]} = \sum_{s=1}^{J} \omega_s \cdot \mathbb{E}[Y_i(s)-Y_i(s-1)|S_i(1)\geq s > S_i(0)] \tag{4.20}$$

where $\omega_s = \frac{S_i(1)\geq s > S_i(0)}{\sum_{j=1}^{J} P(S_i(1)\geq j > S_i(0))}$. This result shows that the Wald estimand yields a weighted average over . It is easy to check that $\sum_{s=1}^{J} \omega_s = 1$ so the weights add up to one (and are positive). Eq. (4.20) nests the familiar LATE formula (4.7) when the treatment variable takes just two values, in which case $J = 1$ and there is just one term in the sum.

Angrist et al. (2000) extend this logic to a treatment variable that is *continuous*, such as when the treatment variable is a price, and we're for example interested in the elasticity of demand with respect to the price. Let P_i be our treatment variable, where i might indicate a market in the case of demand elasticity. Potential outcomes $Y_i(p)$ might then denote quantities demanded as a function of the price p. Angrist et al. (2000) consider a setup with instrument(s) Z that may not be binary, but suppose that we have the following instrument monotoncity assumption

IV monotonicity with a continuous treatment: For any pair z, z': $P_i(z') \geq P_i(z)$ for all i or $P_i(z') \leq P_i(z)$ for all i.

In their application Z_i is the weather, which is assumed to shift supply but not demand for fish. They then show that, analogously to (4.18) and (4.20):

$$\frac{\mathbb{E}[Y_i|Z_i=z']-\mathbb{E}[Y_i|Z_i=z]}{\mathbb{E}[P_i|Z_i=z']-\mathbb{E}[P_i|Z_i=z]} = \int \omega(p) \cdot \mathbb{E}\left[\frac{dY_i(p)}{dp}\middle| P_i(z')\geq p > P_i(z)\right] \cdot dp \tag{4.21}$$

where $\omega(p) = \frac{P(P_i(z')\geq p > P_i(z))}{\int P(P_i(z')\geq p' > P_i(z))\cdot dp'}$. A Wald ratio comparing two instrument values z' and z (such that $P_i(z') \geq P_i(z)$) identifies a weighted average of the derivative treatment effect $\frac{dY_i(p)}{dp}$, among "complier" markets whose price p is affected by the shift of instrument value from z to z'. These weights are positive and integrate to unity, since $\int P(P_i(z') \geq p > P_i(z)) \cdot dp = 1$. Angrist and Imbens (1995) also extend this expression to allow for covariates.

Moreover, the authors show that if we have a single continuous instrument having density f_z, then for any function $g(z)$ of the instrument:

$$\frac{Cov(Y_i, g(Z_i))}{Cov(P_i, g(Z_i)} = \int \lambda(z) \cdot \mathbb{E}\left[\frac{dY_i(p)}{dp}\bigg|_{p=P_i(z)}\right] \cdot dz$$

where $\lambda(z) = \frac{\alpha(z)}{\int \alpha(z')dz'}$ with $\alpha(z) = \frac{dP_i(z)}{dz} \cdot \int_z^{\infty} (g(y) - \mathbb{E}[g(Z_i)]) \cdot f_z(y) \cdot dy$. The weighting function here integrates to unity, but whether or not it is positive depends on the function g. One choice that guarantees it will be positive is $g(z) = \mathbb{E}[P_i|Z_i = z]$. With this choice, the result shows that LHS of the above captures a weighted average of the causal derivative function $dY_i(p)/dp$ across different prices p.

So far we've only considered treatment variables that are *ordered*, whether they are binary, discrete, or continuous. What about when the treatment variable does not have any natural order to it, for example when the treatment variable is something like occupation or field of study? (Heckman and Pinto, 2018) extend the notion of monotonicity to these settings and develop identification results.

4.3 The two stage least squares estimator

The proceeding sections have considered *identification* results unders IV assumptions. This section considers *estimation*.

4.3.1 Prelude: estimating the Wald ratio with a single binary instrument

Throughout most of the identification results above, we focused on understanding the quantity $\frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$, where we use S_i to denote our treatment variable (which may be binary, discrete or continuous). Here

Z_i is a scalar instrument of some kind. When Z_i is binary, we know that $\frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$ takes the form of a Wald ratio:

$$\frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]}$$

With an i.i.d. sample $(Y_i, S_i, Z_i)_{i=1}^n$, each of these four conditional expectations can be consistently estimated by their sample analog, e.g.

$$\hat{\mathbb{E}}[Y_i|Z_i = 1] = \frac{\sum_{i=1}^n Y_i \cdot \mathbb{1}(Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)}$$

By the LLN and the continuous mapping theorem (see Section B.4), we can replace each of the conditional expectations in the Wald ratio by their sample analog's to get a consistent estimator of $\frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$:

$$\frac{\hat{\mathbb{E}}[Y_i|Z_i = 1] - \hat{\mathbb{E}}[Y_i|Z_i = 0]}{\hat{\mathbb{E}}[S_i|Z_i = 1] - \hat{\mathbb{E}}[S_i|Z_i = 0]} \xrightarrow{p} \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[S_i|Z_i = 1] - \mathbb{E}[S_i|Z_i = 0]}$$

4.3.2 The two stage least-squares estimator with a single instrument

Now how does this generalize to the case in which Z_i may take many values, rather than being binary? Recall from Section 4.1 that our parameter of interest, the ratio $\frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$, is equal to the coefficient β_1 from a regression

$$Y_i = \beta_0 + \beta_1 S_i + U_i \tag{4.22}$$

In Section 4.1.4, we saw that we can write β_1 as the ratio of a simple linear regression coefficient of the outcome on the instrument $\frac{Cov(Y_i,Z_i)}{Var(Z_i)}$ to the simple linear regression coefficient of the treatment variable on the instrument $\frac{Cov(S_i,Z_i)}{Var(Z_i)}$. The latter of these two regressions is referred to as the *first-stage* regression.

$$S_i = \pi_0 + \pi_1 Z_i + V_i \tag{4.23}$$

The two-stage least squares estimator (2SLS) is constructed by first estimating the *first stage* Eq. (4.23) by OLS, and then using these estimates to define the *fitted* or *predicted* value $\hat{S}_i$ for each observation i. Next, Equation (4.22) is estimated by OLS, but using $\hat{S}_i$ rather than S_i as the regressor. This is the *second stage* of 2SLS. This delivers a vector of estimated coefficients $\hat{\beta}_{2SlS} = (\hat{\beta}_0, \hat{\beta}_1)'$, of which the second component is our estimate of β_1.

To see why this works, recall from Section 3.3 that the OLS fitted value $\hat{S}_i$ is defined as $\hat{\pi}_0 + \hat{\pi}_1 Z_i$, where $\hat{\pi}_0, \hat{\pi}_1$ are the OLS estimates of the coefficients in Eq. (4.23). Then

$$\hat{\beta}_1 = \frac{\widehat{Cov}(Y_i, \hat{S}_i)}{\widehat{Var}(\hat{S}_i)} = \frac{\hat{\pi}_1 \cdot \widehat{Cov}(Y_i, Z_i)}{\hat{\pi}_1^2 \cdot \widehat{Var}(Z_i)} = \frac{\hat{\rho}_1}{\hat{\pi}_1}$$

where we've used that $\hat{S}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ and that the constant term does not contribute to the covariance or variance. The quantity in the third equality, $\hat{\rho}_{2sls,1} = \frac{\widehat{Cov}(Y_i,Z_i)}{\widehat{Var}(Z_i)}$, is the OLS coefficient on Z_i in the reduced-form" regression of Y_i on Z_i introduced in Section 4.1.4. Since $\hat{\rho}_1$ and $\hat{\pi}_1$ are each consistent estimators of their population counterparts, we have by the continuous mapping theorem that $\hat{\rho}_{2sls,1} \xrightarrow{p} \beta_1 := \frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$.

Note: if you implement 2SLS "by hand", manually computing $\hat{S}_i$ (using e.g. the Stata `predict` command) and then run the second-stage regression, the point estimates will be those of 2SLS but the standard errors will not be valid, because they do not account for the fact that $\hat{S}_i$ is estimated from the data. See the end of Section 4.3.3 for a discussion of standard errors for the 2SLS estimator.

4.3.3 The general 2SLS estimator with multiple treatments, instruments and covariates†

Throughout our study of the local average treatment effects model, we found that the quantity $\frac{Cov(Y_i,Z_i)}{Cov(S_i,Z_i)}$ captured a meaningful average of causal effects, even when treatment S_i takes on many values and/or is continuous. In the last section, we've seen the 2SLS approach to estimating this quantity. But what

about when we have multiple treatment variables, and/or multiple instruments, and possibly covariates that are necessary to control for?

Generalizing LATE results to a case in which treatment is unordered or in which there are multiple treatment variables is not straightforward, and is an active area of research (see Heckman and Pinto, 2018; Lee and Salanié, 2018; Kirkeboen et al., 2016; Kline and Walters, 2016 for some examples). But while heterogeneous treatment effects are complicated, one can still use IV in very general settings by assuming that treatment effects are homogenous, and linear when S_i contains continuous components.

To this end, consider the equation:

$$Y_i = S_i'\beta_S + X_i'\beta_X + U_i \tag{4.24}$$

in which our parameter of interest is the vector β_S of coefficients on our treatment variables S_i, and the variables X_i are covariates, which we can assume to include a constant (this prevents us from needing an intercept term β_0). Suppose that S_i has k components, which we'll refer to as our k *endogenous* variables. We call these variables endogenous because $Cov(S_i, U_i) \neq \mathbf{0}$, where since S_i is now a vector we use the notation that $\mathbf{0}$ is a vector of zeros (in this case, having k components). Eq. (4.24) is referred to as the outcome equation or the structural equation.

We aim to identify β_S using a vector of m instruments Z_i, where $Cov(Z_i, U_i) = \mathbf{0}$. We will use the following terminology:

- In the case that $m < k$ say that the model is *underidentified*. In this case we lack sufficient instruments to identify the k components of β_S, and the 2SLS estimator will not be defined.
- In the case that $m > k$ say that the model is *overidentified*. We have more instruments than we have endogenous variables.
- In the case that $m = k$ say that the model is *just* identified or *exactly* identified. We have exactly the same number of instruments as endogenous variables.

The typical way to motivate Eq. (4.24) is to think of it as a structural model that determines Y_i, in which β_S and β_X denote the causal effects of treatments S and covariates X on Y, i.e. $Y_i(s, w) = s'\beta_S + w'\beta_X + U_i$. On this view, the X are sometimes referred to as "included exogenous variables", since they show up in the structural equation (4.24) but are uncorrelated with the error term. In this jargon, the instruments are called "excluded exogenous variables" since they do not appear in the outcome equation (4.24). However, it is not necessary to treat the covariates as exogenous to use the 2SLS estimator. Rather, we will think of them as controls that are necessary for the instruments to be valid, in the sense of the conditional IV independence assumption introduced in Section 4.2.7. Below, we will state a full set of assumptions under which 2SLS can be used.

As before, the 2SLS approach will be to replace S_i in Eq. (4.24) by its predicted value in a first stage OLS regression. When there is just one endogenous variable ($k = 1$), we simply augment Equation 4.23 to incorporate the same covariates as the outcome equation:

$$S_i = Z_i'\pi_Z + X_i'\pi_X + V_i \tag{4.25}$$

and the predicted value is $\hat{S}_i = Z_i'\hat{\beta}_Z + X_i'\hat{\beta}_X$. It is important to include the covariates X in (4.25), as the 2SLS estimator will not be consistent without them, even under conditional IV independence.

When $k > 1$, we instead have a first stage equation for *each* of the endogenous variables:

$$S_{1i} = Z_i'[\pi_Z]_1 + X_i'[\pi_Z]_1 + V_{1i} \tag{4.26a}$$

$$S_{2i} = Z_i'[\pi_Z]_2 + X_i'[\pi_Z]_2 + V_{2i} \tag{4.26b}$$

$$\vdots$$

$$S_{ki} = Z_i'[\pi_Z]_k + X_i'[\pi_Z]_k + V_{ki} \tag{4.26c}$$

where π_Z is a $k \times m$ matrix and $[\pi_Z]_j$ indicates it's j^{th} row. Similarly, if there are p covariates (including a constant), π_X is a $k \times p$ matrix of first stage coefficients for the covariates.

From Chapter 3 that for each equation $j = 1 \dots k$, we know that the OLS estimator of $([\pi_Z]_j, [\pi_X]_j)$ is given by

$$\begin{pmatrix} [\hat{\pi}_Z]_j \\ [\hat{\pi}_X]_j \end{pmatrix} = ([\mathbf{Z}, \mathbf{X}]'[\mathbf{Z}, \mathbf{X}])^{-1}[\mathbf{Z}, \mathbf{X}]'\mathbf{S}_j$$

where $\mathbf{S}_j$ is an $n \times 1$ vector of observations of treatment S_j, and n is the sample size. Here where $\mathbf{Z}$ is an $n \times m$ matrix with rows Z_i', $\mathbf{Z}$ is an $n \times p$ matrix with rows X_i', and $[\mathbf{Z}, \mathbf{X}]$ is then an $n \times (l + p)$ matrix formed from all the RHS regressors: the m instruments and the p covariates.

The $n \times 1$ vector of fitted values $\hat{\mathbf{S}}_j$ for endogenous variable j can then be written as

$$\hat{\mathbf{S}}_j = [\mathbf{Z}, \mathbf{X}] \begin{pmatrix} [\hat{\pi}_Z]_j \\ [\hat{\pi}_X]_j \end{pmatrix} = [\mathbf{Z}, \mathbf{X}]'([\mathbf{Z}, \mathbf{X}]'[\mathbf{Z}, \mathbf{X}])^{-1}[\mathbf{Z}, \mathbf{X}]'\mathbf{S}_j = \mathbf{P_{ZX}}\mathbf{S}_j$$

where we define the $n \times n$ matrix $\mathbf{P_{ZX}} = [\mathbf{Z}, \mathbf{X}]([\mathbf{Z}, \mathbf{X}]'[\mathbf{Z}, \mathbf{X}])^{-1}[\mathbf{Z}, \mathbf{X}]'$. This matrix represents a *projector* matrix into the subspace of $\mathbb{R}^n$ spanned by the instruments and covariates.

Now let us return to the structural equation (4.24). If we collect across all i, we have

$$\mathbf{Y} = \mathbf{S}\beta_S + \mathbf{X}\beta_X \ + \mathbf{U} = [\mathbf{S}, \mathbf{X}] \begin{pmatrix} \beta_S \\ \beta_X \end{pmatrix} + \mathbf{U} \tag{4.27}$$

where $\mathbf{S}$ is an $n \times k$ matrix with rows S_i'. The 2SLS estimator results from replacing each S_{ij} by it's corresponding predicted value $\hat{S}_{ij}$ and performing OLS. Collecting across all $j = 1 \dots k$, we can define an $n \times k$ matrix $\hat{\mathbf{S}}$ with rows $\hat{\mathbf{S}}_j'$, i.e. $\hat{\mathbf{S}} = \mathbf{P_{ZX}}\mathbf{S}$.

The 2SLS estimator is then:

$$\begin{aligned} \hat{\beta}_{2sls} = \begin{pmatrix} \hat{\beta}_{2sls,S} \\ \hat{\beta}_{2sls,X} \end{pmatrix} &= ([\hat{\mathbf{S}}, \mathbf{X}]'[\hat{\mathbf{S}}, \mathbf{X}])^{-1}[\hat{\mathbf{S}}, \mathbf{X}]'\mathbf{Y} \\ &= ([\mathbf{P_{ZX}}\mathbf{S}, \mathbf{X}]'[\mathbf{P_{ZX}}\mathbf{S}, \mathbf{X}])^{-1}[\mathbf{P_{ZX}}\mathbf{S}, \mathbf{X}]'\mathbf{Y} \end{aligned}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations of the outcome variable. A useful identity is that $\mathbf{P_{ZX}}\mathbf{X} = \mathbf{X}$, because the covariates X are included in the projector matrix $\mathbf{P_{ZX}}$. Furthermore, the matrix $\mathbf{P_{ZX}}$ is symmetric and idempotent, meaning that $\mathbf{P_{ZX}}'\mathbf{P_{ZX}} = \mathbf{P_{ZX}}\mathbf{P_{ZX}} = \mathbf{P_{ZX}}$. These properties allow us to rewrite $[\mathbf{P_{ZX}}\mathbf{S}, \mathbf{X}]$ as $\mathbf{P_{ZX}}[\mathbf{S}, \mathbf{X}]$ and then simplify the 2SLS estimator to:

$$\hat{\beta}_{2sls} = ([\mathbf{S}, \mathbf{X}]'\mathbf{P_{ZX}}[\mathbf{S}, \mathbf{X}])^{-1}[\mathbf{S}, \mathbf{X}]'\mathbf{P_{ZX}}\mathbf{Y} \tag{4.28}$$

One can then show that under standard conditions for the IV model, $\hat{\beta}_{2sls} \xrightarrow{p} \beta$, where $\beta = (\beta_S, \beta_X)'$ are the coefficients in the structural equation (4.24).

To derive standard errors for the 2SLS estimator, one can begin by noting that by substituting (4.27) into(4.28):

$$\hat{\beta}_{2sls} = \beta + ([\mathbf{S}, \mathbf{X}]'\mathbf{P_{ZX}}[\mathbf{S}, \mathbf{X}])^{-1}[\mathbf{S}, \mathbf{X}]'\mathbf{P_{ZX}}\mathbf{U} \tag{4.29}$$

where $\beta = \begin{pmatrix} \beta_S \\ \beta_X \end{pmatrix}$. Analogously to the case of OLS, the statistical properties of $\hat{\beta}_{2sls}$ arise from the second term, which depends upon the unobserverd errors $\mathbf{U}$. We omit an explicit formula for estimated the standard errors, but note that they follow from the standard theory for *generlized method of moments* estimators (see box below), and statistical packages (e.g. `ivregress2` in Stata) will calculate them for you.

> The 2SLS estimator is an example of the *generalized method of moments* (GMM) estimator, which tries to solve sample analogs of the $m + p$ moment conditions $\mathbb{E}[(Z_i, X_i)'U_i] = \mathbb{E}[(Z_i, X_i)'(Y_i - S_i'\beta_S - X_i'\beta_X)] = \mathbf{0}$. Each of the $m + p$ *moment conditions* is implied by the exogeneity of Z and X. When the model is overidentified, there is generally not any $\hat{\beta}_S$ and $\hat{\beta}_X$ that can set all of these $m + p$ equations to exactly zero. GMM proceeds by minimizing the size of deviations from the above equation, where a weighting matrix is used to aggregate over the various moment conditions. 2SLS corresponds to the choice $\mathbf{P_{ZX}}$ as the weighting matrix.

When the model is just identified, such that $l = k$, we can decompose $[\mathbf{S}, \mathbf{X}]'\mathbf{P_{ZX}}[\mathbf{S}, \mathbf{X}]$ as the product of three $k + p \times k + p$ matrices: $[\mathbf{S}, \mathbf{X}]'[\mathbf{Z}, \mathbf{X}]$, $([\mathbf{Z}, \mathbf{X}]'[\mathbf{Z}, \mathbf{X}])^{-1}$, and $[\mathbf{Z}, \mathbf{X}]'[\mathbf{S}, \mathbf{X}]$. We can then use the

matrix identity that $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ to rewrite the 2SLS estimator in the just-identified case as:

$$\begin{aligned}\hat{\beta}_{2sls} =&([\mathbf{Z},\mathbf{X}]'[\mathbf{S},\mathbf{X}])^{-1}([\mathbf{Z},\mathbf{X}]'[\mathbf{Z},\mathbf{X}])\cancel{([\mathbf{S},\mathbf{X}]'[\mathbf{Z},\mathbf{X}])^{-1}}\cancel{[\mathbf{S},\mathbf{X}]'[\mathbf{Z},\mathbf{X}]}([\mathbf{Z},\mathbf{X}]'[\mathbf{Z},\mathbf{X}])^{-1}[\mathbf{Z},\mathbf{X}]'\mathbf{Y}\\ =&([\mathbf{Z},\mathbf{X}]'[\mathbf{S},\mathbf{X}])^{-1}\cancel{([\mathbf{Z},\mathbf{X}]'[\mathbf{Z},\mathbf{X}])}\cancel{([\mathbf{Z},\mathbf{X}]'[\mathbf{Z},\mathbf{X}])^{-1}}[\mathbf{Z},\mathbf{X}]'\mathbf{Y}\\ =&([\mathbf{Z},\mathbf{X}]'[\mathbf{S},\mathbf{X}])^{-1}[\mathbf{Z},\mathbf{X}]'\mathbf{Y}\end{aligned}$$

This provides a simpler formula compared with the general case of Eq. (4.28) in which $l \geq k$. Note that this formulation is a nice way to see that when we use the S_i as instruments for themselves (i.e. $\mathbf{Z} = \mathbf{S}$), the estimator coincides exactly with the OLS estimates in the model $Y_i = S_i'\beta_S + X_i'\beta_X + U_i$:

$$\hat{\beta}_{OLS} = ([\mathbf{S},\mathbf{X}]'[\mathbf{S},\mathbf{X}])^{-1}[\mathbf{S},\mathbf{X}]'\mathbf{Y}$$

4.3.4 2SLS issues: functional form and weak instruments*

This section details two potential problems with the 2SLS estimator, which can prevent it from delivering meaningful results about the causal effects of the endogenous variables S on Y.

Treatment effect heterogeneity*

Recall that homogeneous treatment effects is typically a very strong assumption, and our outcome equation (4.24) seems to impose that the effect of S on Y is described by $s'\beta_S$ for all units i. When Eq. (4.24) does not in fact provide a formula for potential outcomes, what does 2SLS estimate?

First, our analysis of no-selection-on-gains (NSOG) from Section 4.2.1 extends to the general setting in which we might use the 2SLS estimator. Roughly speaking, 2SLS will still be consistent for the average treatment effect when:

1. $\{(\{Y_i(s)\}_s, \{S_i(z)\}_z) \perp\!\!\!\perp Z_i\} | X_i$ where $Y_i(s) = Y_i(s,z)$ (conditional independence and exclusion)
2. $\mathbb{E}[Y_i(s) - Y_i(s_0)|Z_i, S_i, X_i] = s'\beta_S$ (no-selection on gains and linearity of treatment effects).
3. $\mathbb{E}[Y_i(s_0)|X_i = x] = x'\beta_X$ (linearity with respect to covariates)
4. Π_Z has full rank (relevance)

In the second assumption we fix a reference category of treatment s_0, e.g. $s_0 = (0, \dots 0)'$. This assumption imposes NSOG because the average conditional treatment effect from s_0 to s does not depend on the value of treatment S_i. It also does not depend upon the covariates which are also conditioned upon, and is linear in s.

To see that these assumptions are sufficient, let us generate an equation for the realized value of Y_i as we did in Section 4.2.1:

$$\begin{aligned}Y_i = Y_i(S_i) &= Y_i(s_0) + Y_i(S_i) - Y_i(s_0)\\ &= \mathbb{E}[Y_i(s_0)|X_i] + \mathbb{E}[Y_i(S_i) - Y_i(s_0)|Z_i, S_i, X_i]\\ &\qquad + (Y_i(S_i) - Y_i(s_0) - \mathbb{E}[Y_i(S_i) - Y_i(s_0)|Z_i, S_i, X_i] + Y_i(s_0) - \mathbb{E}[Y_i(s_0)|X_i])\\ &= S_i'\beta_S + X_i'\beta_X + U_i\end{aligned}$$

where

$$U_i = \underbrace{Y_i(S_i) - Y_i(s_0) - \mathbb{E}[Y_i(S_i) - Y_i(s_0)|Z_i, S_i, X_i]}_{\text{term A}} + \underbrace{Y_i(s_0) - \mathbb{E}[Y_i(s_0)|X_i]}_{\text{term B}},$$

and we've used assumptions 2 and 3.

Now consider an instrument vector Z_i satisfying Assumption 1. Given that 2SLS is a type of GMM estimator (see e.g. Newey and McFadden 1994, for some general theory), we can establish consistency by first showing that the following conditional moment equality is satisfied (with probability one):

$$\mathbb{E}[U_i|Z_i, X_i] = \mathbb{E}[Y_i - S_i'\beta_S - X_i'\beta_X|Z_i, X_i] = 0$$

To see that this moment condition is satisfied under Assumption 1, note that since $\mathbb{E}[Y_i(S_i)-Y_i(s_0)|Z_i, S_i, X_i]$ does not depend on Z_i, $\mathbb{E}[Y_i(S_i)-Y_i(s_0)|Z_i, S_i, X_i] = \mathbb{E}[Y_i(S_i)-Y_i(s_0)|S_i, X_i]$ and hence $\mathbb{E}[$ term $\text{A}|Z_i, X_i] = 0$. Then by Assumption 1, $\mathbb{E}[Y_i(s_0)|X_i] = \mathbb{E}[Y_i(s_0)|Z_i, X_i]$, and hence $\mathbb{E}[$ term $\text{B}|Z_i, X_i] = 0$ as well. Further technical conditions are required for consistency of the 2SLS estimator, but Assumption 4 guarantees that there is sufficient variation in the instruments for identification, and that the probability limit of the estimator is well-defined under standard regularity conditions.

In general, Assumptions 1-4 are pretty strong. In fact, the assumptions above imply that β_S can be identified even without the Z_i, by using non-linear functions of X_i as instruments. Kolesár (2013) shows that with a single treatment variable Assumption 3 can be replaced by assuming that the conditional expectation of the instruments given the covariates are linear:

3.* $\mathbb{E}[Z_i|X_i = x] = \gamma x$ for some matrix γ (linearity of instruments given covariates).

Kolesár (2013) and Angrist and Imbens (1995) show what 2SLS estimates when there is just a single treatment variable S, but NSOG is relaxed and replaced by a LATE monotonicity assumption. Bhuller and Sigstad (2022) further considers the case with multiple treatments S_i, under a multiple-treatment analog of monotonicity. A key result here is that for 2SLS to uncover a weighted average of LATEs (with positive weights), we need a condition of *no cross effects* that prevents the effects of other treatments from contaminating the 2SLS coefficient for a given treatment.

Misspecification of the role of covariates†

A particularly strong restriction under the assumptions above is 3, which as we've seen above can be replaced by 3.*, linearity of the CEF of instruments given covariates. Blandhol et al. (2022) show that when 3.* is relaxed and NSOG is replaced by the LATE monotonicity assumption, problems arise even in the case with a single binary treatment. Not only do compliers contribute to the 2SLS estimand, but so do always takers, who show up with negative weights. In general, care should be taken when interpreting 2SLS causally in the LATE model, unless the specification of covariates is sufficiently flexible to ensure linearity of $\mathbb{E}[Z_i|X_i]$.

Weak instruments*

The last two subsections have considered specification issues: when treatment effects are heterogeneous and we cannot interpret Eq. (4.24) as a direct model of potential outcomes. Another potential pitfall of the 2SLS estimator is statistical in nature, and occurs when the instrments are relevant but are only weak predictors of the endogenous variables S.

From Eq. (4.29), one can show that the estimator $\hat{\beta}_{2sls}$ is biased in finite samples. For simplicity, consider the case without covariates, in which case

$$\hat{\beta}_{2sls} = \beta_S + (\mathbf{S}'\mathbf{P_Z}\mathbf{S})^{-1}\mathbf{S}'\mathbf{P_Z}\mathbf{U} = \beta_S + (\mathbf{S}'\mathbf{P_Z}\mathbf{S})^{-1}(\pi_Z\mathbf{Z}'\mathbf{U} + \mathbf{V}'\mathbf{P_Z}\mathbf{U})$$

where $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ and we've used that $\mathbf{S} = \mathbf{Z}\pi_Z + \mathbf{V}$.

Note that endogeneity of the treatments S arises from V_i being correlated with U_i (if it weren't, then by the first stage equation S_i would be uncorrelated with U_i). Thus, the third term above represents a bias term: in a finite sample, 2SLS is in fact biased "in the direction" of OLS as the term $\mathbf{V}'\mathbf{P_Z}\mathbf{U}$ will be non-zero.

Under standard *i.i.d* asymptotics, the finite-sample bias goes away as $n \to \infty$ because $\frac{1}{n}\mathbf{P_Z}\mathbf{V} \xrightarrow{p} \mathbf{0}$, and 2SLS is thus consistent. However, this result may provide a poor approximation of the true statistical properties of $\hat{\beta}_{2sls}$, if $\pi_Z \approx \mathbf{0}$. The punchline is that when π_Z is small or close to being singular, conventional GMM confidence intervals for the 2SLS estimator are unlikely to provide a good approximation to the actual sampling distribution of $\hat{\beta}_{2sls}$.

> *Weak instruments asymptotics.* Recall that the idea of a sequence in which $n \to \infty$ is a fiction, a theoretical device designed to deliver an approximation to the finite-sample distribution of an estimator: in this case $\hat{\beta}_{2sls}$. If n is "large enough" this may deliver a good approximation. However, standard asymptotics, in which the DGP is held fixed across all n (and only the size

of the sample is varied), may not provide the best asymptotic approximation. In the literature on weak instruments, one often instead considers an sequence in which we let π_Z depend on n as $\pi_{Zn} = \pi_Z/\sqrt{n}$ (so-called "weak instrument asymptotics"). Under weak instrument asymptotics, one can show that 2SLS is not even consistent. However, the Anderson Rubin test is.

Multiple solutions have arisen to deal with the potential of weak instruments issues. The first is to try to assess empirically whether weak instruments are likely to be a problem by inspecting the first stage regression for the endogenous variables. One common rule of thumb is to inspect the F-statistic of the first stage regression. In a case without covariates and a single endogenous regressor, one can show that $\mathbb{E}[\hat{\beta}_{2sls,1}] - \beta_1 \approx \frac{Cov(U_i,V_i)}{Var(V_i)} \frac{1}{F+1}$, where F is the population analog of the first-stage F statistic (Angrist and Pischke, 2008). Note that $\frac{Cov(U_i,V_i)}{Var(V_i)}$ also captures the bias of OLS when $\pi_Z = 0$. Thus with an F statistic of say 9, the bias of 2SLS will be roughly 10% as bad as that of OLS. In this case, using 2SLS rather than OLS may seem like a reasonable tradeoff. This rule of thumb, like all rules of thumb, should not be taken as being dispositive.

Another approach is to avoid using 2SLS altogether. There exist alternative estimators (for example the limited-information maximum likelihood estimator) that are less sensitive to weak instruments. But one can go further actually compute confidence intervals for β_S directly, without using point estimation techniques at all. If one is willing to do this, the so-called Anderson-Rubin (AR) test is robust to weak instruments issues, in the sense that it has the correct size asymptotically even under weak-instruments asymptotics. In the case of $m = k = 1$ (just identified case with a single endogenous variable) and no covariates (furthermore taking the constant β_0 for simplicity), one can show that under the null that $\beta_s = b$ for any candidate vector b, the Anderson-Rubin test statistic $A(b)$ has a limiting distribution of a standard normal, with the definition:

$$A(b) := \frac{\sum_i Z_i(Y_i - bS_i)}{\sum_i Z_i^2(Y_i - bS_i)^2}$$

This result can be used as a basis for constructing confidence intervals for β_S that are valid even if the instruments are weak. The trick to making the AR test "work" is that the covariance between S_i and Z_i never appears in the denominator of $A(b)$.

Chapter 5

Discontinuity based methods

In this chapter we consider approaches to identification of treatment effects that rely on discontinuities in treatment assignment or in the institutional constraints that economic agents face. We focus on the hallmark discontinuity-based method: the regression discontinuity design.

5.1 The regression discontinuity design

5.1.1 Introduction

The regression discontinuity design (RDD) was first introduced by Thistlethwaite and Campbell (1960) in the context of evaluating the effect of public recognition on the educational and career outcomes of students. In their setting, students scoring above a certain threshold score on a standardized test were given public recognition for their academic achievement.

Let us consider a slightly different example, in which a population of students take a standardized test, with X_i denoting the resulting test score for student i. Suppose that all students with test scores greater than $X_i = c$ receive a scholarship to attend university. If we're interested in the effect, say, of having the scholarship on university enrollment, then the use of the threshold in assigning the scholarship offers a natural experiment. The students with scores just *below* the cutoff and the students with scores just *above* the cutoff a pretty comparable to one another, ex-ante. However those with scores above the threshold have access to the scholarship, while those just below do not. Any difference in the probability of enrolling in university between these groups, the argument goes, must then be due to the scholarship, and not other factors.

5.1.2 Identification in the sharp RDD

Let us formalize the intuition of the RDD described in above using our potential outcomes notation. Letting D_i be an indicator for whether student i received the scholarship or not, suppose that

$$D_i = \mathbb{1}(X_i \geq c) \tag{5.1}$$

where c represents the cutoff used in offering the scholarship. We call the test score X_i the RDD *running variable*. An assignment rule like (5.1) that is a deterministic function of the running variable is referred to as a "sharp" RDD, which we'll contrast with a more general "fuzzy" RDD in the next section.

Now consider our outcome variable Y_i, an indicator for whether student i enrolls in university. Letting our potential outcomes $Y_i(0)$denote i's enrollment decision in the case that they do not receive the scholarship, and $Y_i(1)$ in the case that they do, we have that

$$Y_i = Y_i(D_i) = \begin{cases} Y_i(0) & \text{if } X_i < c \\ Y_i(1) & \text{if } X_i \geq c \end{cases} \tag{5.2}$$

Consider the average value of Y_i as a function of X_i: the enrollment rate as a function of test score. It is reasonable to expect $m(x) := \mathbb{E}[Y_i|X_i = x]$ to be everywhere increasing in x, if students scoring higher on the test are more likely to enroll in university.

For values $x < c$, $m(x)$ is equal to $\mathbb{E}[Y_i(0)|X_i = x]$, the average "untreated" outcome among students with scores just around x. If this increases with x for values to the left of the threshold (as depicted by

the solid orange line on the RHS of Figure 5.1), the increase cannot be due to the treatment (a scholarship offer), since none of these students are offered the scholarship. Instead, an increase in $m(x)$ for $x < c$ is evidence that there is endogeneity between test scores and outcomes: students with higher test scores having higher outcomes on average than those with lower test scores, even without the scholarship.

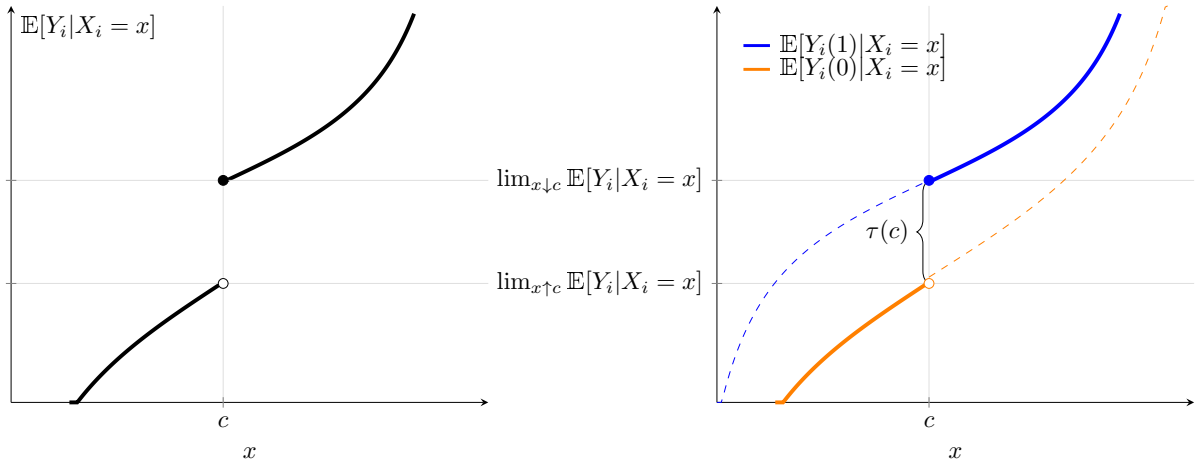


Figure 5.1: Logic of a sharp RDD. Left hand side: the conditional expectation of observed outcome$Y$ with respect to the running variable X. Right hand side, the gap between the blue and orange lines at $x = c$ identifies $\tau(c) := \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$.

Similarly, for $x > c$ the conditional expectation function $m(x)$ captures $\mathbb{E}[Y_i(1)|X_i = x]$. By the same logic, we may expect $m(x)$ to increase with x (as depicted by the solid blue line on the RHS of Figure 5.1) because students who score higher on the exam and receive the scholarship are more likely to enroll in university than students who score lower, even when the latter also receive the scholarship.

The black curve on the LHS of Figure 5.1 depicts the function $m(x)$ over all values of x. This function is *identified* because it is defined in terms of the population distribution of observable quantities (namely Y_i and X_i). Notice that something very special happens right at $x = c$. As we increase x through the threshold c, we switch from observing $Y_i(0)$ potential outcome to observing the $Y_i(1)$ potential outcomes, as we begin to average over students who were offered the scholarship rather than students who did not. If the scholarship has a causal effect on university enrollment, we may expect to see a jump, or *discontinuity* in $m(x)$ at $x = c$.

Refresher: A function $f(x)$ is called *continuous* at $x = x_0$ if $\lim_{x\to x_0} f(x)$ exists and is equal to $f(x_0)$. For $\lim_{x\to x_0} f(x)$ to exist, the limits from the left and from the right must be equal to one another, i.e.

$$\lim_{x\downarrow x_0} f(x) = \lim_{x\uparrow x_0} f(x)$$

If $f(x)$ is discontinuous at $x = x_0$, but the left and right limits are themselves well-defined, then the *discontinuity* in $f(x)$ at $x = x_0$ is the difference between the right and left limits:

$$\lim_{x\downarrow x_0} f(x) - \lim_{x\uparrow x_0} f(x)$$

To further understand the source of the gap depicted in Figure 5.1, let us make the following definitions:

$$m_0(x) := \mathbb{E}[Y_i(0)|X_i = x] \qquad \text{and} \qquad m_1(x) := \mathbb{E}[Y_i(1)|X_i = x]$$

The function $m_0(x)$ is depicted in orange in the right panel of the figure. For values $x < c$, this coincides with the observable CEF $m(x)$. For values $x \geq c$, $m_0(x)$ is unobserved—indicated by a dashed line. The function $m_1(x)$ is depicted in blue. For values $x < c$, it is unobserved (indicated by a dashed blue line), while for values $x \geq c$ it coincides with the observed $m(x)$.

Define

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$$

The function $\tau(x)$ is the average causal effect of a scholarship offer on university enrollment, among students who have a test score of x. For any x, $\tau(x)$ is the difference between the blue curve and the orange curve. If both curves were observed, we could compute $\tau(x)$ for all x. However, there are no values of x for which *both* curves are observable: all units are either treated or are untreated at a given value x, by the deterministic assignment rule 5.1).

The magic of the RDD is to leverage an assumption that $m_0(x)$ is *continuous* at $x = c$ to identify $\tau(x)$ exactly at the single point $x = c$, where the solid portions of the blue and orange curves *almost* overlap. The idea is that while college enrollment is likely to be increasing in test scores (on average), a very small increase in test scores would be associated with only a very small increase in enrollment. By making the test score difference arbitrarily small, we can make the average enrollment difference arbitrarily small as well.

For simplicity, I assume below that both $m_1(x)$ and $m_0(x)$ are continuous in x at $x = c$. Strictly speaking, you only need to assume that one of these is true. The sharp RDD identification result can be stated as follows:

Proposition 5.1 (sharp RDD identification result). *If treatment assignment follows Eq. (5.1) and $m_1(x)$ and $m_0(x)$ are both continuous at $x = c$, then the discontinuity in $m(x) = \mathbb{E}[Y_i|X_i = x]$ yields the local average treatment effect among individuals at the threshold, that is:*

$$\tau(c) = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]$$

Proof. By (5.2), we have that

$$\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1)|X_i = x] = \lim_{x \downarrow c} m_1(x)$$

and

$$\lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \uparrow c} \mathbb{E}[Y_i(0)|X_i = x] = \lim_{x \uparrow c} m_0(x)$$

By continuity of $m_1(x)$ at $x = c$, $\lim_{x \downarrow x} m_1(x) = m_1(c)$. Similarly by continuity of $m_0(x)$ at $x = c$, $\lim_{x \uparrow c} m_0(x) = m_0(c)$. Thus,

$$\begin{aligned}
\lim_{x \downarrow c} m(x) - \lim_{x \uparrow c} m(x) &= m_1(c) - m_0(c) \\
&= \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c] \\
&= \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c] \\
&:= \tau(c)
\end{aligned}$$

□

It's worth reflecting for a moment on the source of identification in the RDD. It does not come from assuming there is no endogeneity in the running variable. Indeed, we have specifically considered an example in which $m_0(x)$ and $m_1(x)$ vary with x. Rather, the RDD assumes that endogeneity gets *small* across *small* differences in the running variable. The magic of the RDD is that right at the threshold, and infitesimal difference in the running variable determines whether an individual is treated or un-treated, leading to an apples-to-apples comparison.

Note: Proposition 5.1 does not say that the ATE is identified. Rather, it shows that $\tau(x)$ is identified *for a single value* of x: in particular at the discontinuity. In Figure 5.1, I've shown a case in which $\tau(x)$ varies considerably with x: it is smaller for small and large values of x, where the orange and blue lines are closer together, while being larger for values of x near c.

The average treatment effect at the threshold $\tau(c)$ might be a poor guide to the overall average treatment effect, and it may not be a directly policy relevant treatment effect parameter. However, it is the one point in the support of x that $\tau(x)$ is identified. Note that if one is interested in forecasting the effect of increasing the threshold by a little bit, then $\tau(c)$ is exactly what you'd like to know, since the individuals with $X_i = c$ are precisely the students that would be identified by this marginal shift to the policy.

Twists on RDD

Discrete running variable: The preceding analysis considers a running variable that is continuous. However, settings often arise in which the support of X_i is discrete, e.g. the integers, and one wants to employ an RDD identification argument. This is possible but requires assumptions that are bit stronger than continuity of the $m_d(x)$ functions. See Kolesár and Rothe (2018) for details.

Many thresholds: Some settings involve discontinuous rules that involve many thresholds, and practitioners seek to combine them by normalizing all thresholds to a common value, e.g. zero. Bertanha (2020) provides a nice analysis of these kinds of settings settings.

Manipulation of the running variable: What if individuals can change their value of X_i in order to make sure they cross the threshold? Does this pose a threat to causal inference? I consider this question in Section 5.1.9.

5.1.3 Identification in the fuzzy RDD*

The last section considered a case in which treatment D_i was a deterministic function of X_i: D_i is equal to one if and only if $X_i \geq c$. What if treatment uptake is not determined entirely on the basis of X_i, but instead features a discontinuous increase at $X_i = c$?

This leads to the so-called *fuzzy* RDD model. To analyze fuzzy RDD settings, we will borrow from the notation and language that we developed for the LATE IV model. In particular, let us say that unit i is *assigned to* treatment if $X_i \geq c$. Denote this by

$$Z_i = \mathbb{1}(X_i \geq c) \tag{5.3}$$

For any unit i, let $D_i(x)$ denote potential treatment for unit i as a function of their value of the running variable. Introduce the notation that

$$D_i^- = \lim_{x \uparrow c} D_i(x) \qquad \text{and} \qquad D_i^+ = \lim_{x \downarrow c} D_i(x) \tag{5.4}$$

Just as with the LATE model, we will assume that there are three groups, defined in terms of their values of D_i^+ and D_i^-:

- *Always-takers* have $D_i^+ = D_i^- = 1$. They would take treatment whether the running variable fell slightly to the right or to the left of c.
- *Never-takers* have $D_i^+ = D_i^- = 0$. They would *not* take treatment whether the running variable fell slightly to the right or to the left of c.
- *Compliers* have $D_i^+ = 1$ and $D_i^- = 0$. If their value of running variable fell slightly to the left of c, they would not take treatment, and if it fell slightly to the right of c they would take treatment.

As an example of the fuzzy RDD model, suppose now we are interested not in the effects of a scholarship on university enrollment, but instead we're interested in the effects of university enrollment on wages at age 30. Now treatment assignment Z_i is an indicator for a scholarship offer, determined by i's test score X_i via Eq. (5.3). The outcome variable Y_i is wages at age 30, and D_i is an indicator for whether i went to university.

Intuitively, always-takers are those students who would go to university irrespective of whether they receive the scholarship. Never-takers do not go to university even if they do receive the scholarship. Compliers are those for whom receipt fo the scholarship makes a difference in their university of enrollment: those who are incented to enrol because of the cost savings (or because of the recognition/positive reinforcement). As with the LATE model, we assume there are no *defiers*: individuals who would only go to university if they did *not* receive the scholarship.

The fuzzy RDD identification result generalizes Proposition 5.1 to the fuzzy setting:

Proposition 5.2 (fuzzy RDD identification result). *Suppose that the following hold:*

- *(Continuity:) for all* $d, d_1, d_2 \in \{0, 1\}$*, the functions*

$$\mathbb{E}[Y_i(d)|X_i = x, D_i^- = d_1 D_i^+ = d_2]$$

and

$$P(D_i^- = d_1 D_i^+ = d_2|X_i = x)$$

are continuous in x *at* $x = c$

- *(Exclusion): potential outcomes* $Y_i(d)$ *do not depend directly on* Z_i*, i.e. whether* x *is over threshold*
- *(Monotonicity): there are no defiers at the threshold, i.e.* $P(D_i^- = 1 D_i^+ = 0|X_i = c) = 0$.
- *(Relevance/first stage):* $P(D_i^- = 1 D_i^+ = 0|X_i = c) > 0$*, i.e. there are compliers at the threshold*

Then:

$$\frac{\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[D_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i|X_i = x]} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_i^+ > D_i^-]$$

Note that the assumptions above are completely analogous to the assumptions underlying the LATE model. However, instead of assuming IV independence, we make continuity assumptions. In the Fuzzy RDD, the causal parameter that is identified is the local average treatment effect among compliers at the threshold:

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_i^+ > D_i^-]$$

This quantity conditions on *two* events: being an individual with a value of $X_i = c$, and being among the individuals who actually switch their treatment status when crossing the threshold.

Note: The fuzzy RDD identification result nests the sharp one (Proposition 5.1) in the case that $D_i = Z_i$, so that all units are compliers, and there are no always- or never-takers. If actual treatment D_i is not observed, one can still use Proposition 5.1 with $D_i = Z_i$ to identify so-called *intent-to-treat* effects: the effect of being assigned to treatment (rather than actually receiving treatment) on the outcome.

Now let's see why Proposition 5.2 holds. Recall that $D_i = D_i(X_i)$, i.e. realized treatment assignment depends on potential treatments and one's value of the running variable.

Proof. Consider e.g. $\lim_{x \downarrow c} m(x)$. By the law of iterated expectations this is:

$$\begin{aligned}
\lim_{x \downarrow c} m(x) &:= \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(D_i(X_i))|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(D_i(x))|X_i = x] \\
&= \lim_{x \downarrow c} P(D_i^- = 0, D_i^+ = 0|D_i = x) \cdot \mathbb{E}[Y_i(D_i(x))|X_i = c, D_i^- = 0, D_i^+ = 0] \\
&\quad + \lim_{x \downarrow c} P(D_i^- = 1, D_i^+ = 1|D_i = x) \cdot \mathbb{E}[Y_i(D_i(x))|X_i = x, D_i^- = 1, D_i^+ = 1] \\
&\quad + \lim_{x \downarrow c} P(D_i^- = 0, D_i^+ = 1|D_i = x) \cdot \mathbb{E}[Y_i(D_i(x))|X_i = x, D_i^- = 0, D_i^+ = 1]
\end{aligned}$$

Noting that in the limit as $x \downarrow c$, $D_i(x)$ approaches D_i^+, we have by continuity that

$$\begin{aligned}
\lim_{x \downarrow c} m(x) &= P(D_i^- = 0, D_i^+ = 0|X_i = c) \cdot \mathbb{E}[Y_i(0)|X_i = c, D_i^- = 0, D_i^+ = 0] \\
&\quad + P(D_i^- = 1, D_i^+ = 1|X_i = c) \cdot \mathbb{E}[Y_i(1)|X_i = c, D_i^- = 1, D_i^+ = 1] \\
&\quad + P(D_i^- = 0, D_i^+ = 1|X_i = c) \cdot \mathbb{E}[Y_i(1)|X_i = c, D_i^- = 0, D_i^+ = 1]
\end{aligned}$$

where we've also used the property that $\lim_{x \downarrow x_0} f(x) \cdot g(x) = f(x_0) \cdot g(x_0)$ when both limits exist.

Repeating the same steps for the limit from below, we have instead that

$$\begin{aligned}
\lim_{x \uparrow c} m(x) &= P(D_i^- = 0, D_i^+ = 0|X_i = c) \cdot \mathbb{E}[Y_i(0)|X_i = c, D_i^- = 0, D_i^+ = 0] \\
&\quad + P(D_i^- = 1, D_i^+ = 1|X_i = c) \cdot \mathbb{E}[Y_i(1)|X_i = c, D_i^- = 1, D_i^+ = 1] \\
&\quad + P(D_i^- = 0, D_i^+ = 1|X_i = c) \cdot \mathbb{E}[Y_i(0)|X_i = c, D_i^- = 0, D_i^+ = 1]
\end{aligned}$$

In the numerator of Proposition 5.2, the always-taker and never-taker terms cancel and we are left only with the compliers:

$$\lim_{x \downarrow c} m(x) - \lim_{x \uparrow x} m(x) = P(D_i^- = 0, D_i^+ = 1 | X_i = c) \cdot \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c, D_i^- = 0, D_i^+ = 1]$$

An analagous argument for the denominator shows that it identifies the proportion of compliers at the threshold, that is if we let $d(x) := \mathbb{E}[D_i | X_i = x]$:

$$\lim_{x \downarrow c} d(x) - \lim_{x \uparrow c} d(x) = P(D_i^- = 0, D_i^+ = 1 | X_i = c)$$

and the result thus holds by relevance. □

5.1.4 Parametric estimation in the RDD

How should one implement sample estimators of the expressions in Propositions 5.1 and 5.2? Here we first consider estimation under the functional form restriction that the relevant estimands $m(x)$ and $d(x)$ are linear on each side of the cutoff. This is not generally a good idea (unless under inspection, the scatter plots actually *look* linear), but provides a good intuition for how to proceed with estimation more generally.

5.1.4.1 Sharp RDD by OLS

Consider the first sharp case, where our goal is to estimate the quantity

$$\lim_{x \downarrow x} m(x) - \lim_{x \uparrow x} m(x)$$

Suppose that $m_0(x)$ and $m_1(x)$ are both linear in x, i.e.

$$m_0(x) = \mathbb{E}[Y_i(0) | X_i = x] = \gamma_0 + \gamma_1 \cdot (x - c) \tag{5.5}$$

and

$$m_1(x) = \mathbb{E}[Y_i(1) | X_i = x] = \lambda_0 + \lambda_1 \cdot (x - c) \tag{5.6}$$

Then, by (5.2):

$$m(x) = \mathbb{E}[Y_i | X_i = x] = \begin{cases} \gamma_0 + \gamma_1 \cdot (x - c) & \text{for } x < c \\ \lambda_0 + \lambda_1 \cdot (x - c) & \text{for } x \geq c \end{cases}$$

and we can thus write

$$\begin{aligned} Y_i &= \gamma_0 \cdot \mathbb{1}(X_i < c) + \gamma_1 \cdot (X_i - c) \cdot \mathbb{1}(X_i < c) + \lambda_0 \cdot \mathbb{1}(X_i \geq c) + \lambda_1 \cdot (X_i - c) \cdot \mathbb{1}(X_i \geq c) + \epsilon_i \\ &= \gamma_0 + \gamma_1 \cdot (X_i - c) + (\lambda_0 - \gamma_0) \cdot \mathbb{1}(X_i \geq c) + (\lambda_1 - \gamma_1) \cdot (X_i - c) \cdot \mathbb{1}(X_i \geq c) + \epsilon_i \end{aligned}$$

where $\mathbb{E}[\epsilon_i | X_i] = 0$. Note that $\mathbb{E}[\epsilon_i | X_i]$ is the same thing as $\mathbb{E}[\epsilon_i | X_i, \mathbb{1}(X_i \geq c), X_i \cdot \mathbb{1}(X_i \geq c)]$. Here we've used that $\mathbb{1}(X_i < c) = 1 - \mathbb{1}(X_i \geq c)$.

Our final estimating equation can thus be written

$$Y_i = \beta_0 + \beta_1 \cdot (X_i - c) + \beta_2 \cdot \mathbb{1}(X_i \geq c) + \beta_3 \cdot (X_i - c) \cdot \mathbb{1}(X_i \geq c) + \epsilon_i \tag{5.7}$$

where $\beta_2 = \lambda_0 - \gamma_0 = \lim_{x \downarrow x} m(x) - \lim_{x \uparrow x} m(x) = \tau(c)$. The treatment effect parameter $\tau(c)$ can therefore be estimated by an OLS regression of the outcome on a constant, the distance $X_i - c$ between the running variable and the cutoff, treatment $D_i = \mathbb{1}(X_i \geq c)$, and an interaction between D_i and $(X_i - c)$. The coefficient on the treatment indicator D_i then provides an estimate $\hat{\tau}(c) = \hat{\beta}_2$. Under the standard conditions for OLS consistency $\hat{\tau}(c) \xrightarrow{p} \tau(c)$.

Note: It can be tempting to omit the interaction β_3 in Eq. (5.10) between D_i and X_i. This is not generally valid: one should allow the function $m(x)$ to have different slopes on either side of the threshold. One case in which dropping β_3 is valid is when treatment effects are assumed to be homogenous, in which case $m_1(x) = m_0(x) + \Delta$ for the homogenous treatment effect Δ, and hence $\lambda_1 = \gamma_1$.

Functional form misspecification: Though estimating Eq. (5.10) is very straightforward, it relies on the linearity conditions Eqs. (5.5) and (5.6) to hold to be valid. One could make Eq. (5.10) more flexible by adding higher powers of X_i and their interactions with $D_i = \mathbb{1}(X_i \geq c)$, or limit estimation to a bandwidth around the cutoff. In Section 5.1.5, we consider a more robust version of this idea, using the idea of *non-parametric regression*. First, let us turn to parametric estimation in the fuzzy case.

5.1.4.2 Fuzzy RDD by 2SLS*

In the fuzzy case our estimand is, by Proposition 5.2:

$$\frac{\lim_{x \downarrow c} m(x) - \lim_{x \uparrow c} m(x)}{\lim_{x \downarrow c} d(x) - \lim_{x \uparrow c} d(x)} \tag{5.8}$$

where recall that $d(x) := \mathbb{E}[D_i | X_i = x]$.

Note now that for $x < c$, we no longer have $m(x) = \mathbb{E}[Y_i(1)|X_i = x]$, because $m(x)$ now mixes an average of $Y_i(0)$ for compliers and never-takers with $Y_i(1)$ among always-takers (by the LIE). Nevertheless, we might visually inspect a scatter plot of Y versus X and conclude that for $x < c$, $m(x)$ is indeed approximately linear, and similarly for $x \geq c$. In this case, we can estimate the numerator $\lim_{x \downarrow c} m(x) - \lim_{x \uparrow c} m(x)$ of (5.8) in the same way as in the last section. For reasons that will be clear below, let us use a new notation for Eq: (5.10):

$$Y_i = \rho_0 + \rho_1 \cdot X_i + \rho_2 \cdot \mathbb{1}(X_i \geq c) + \rho_3 \cdot X_i \cdot \mathbb{1}(X_i \geq c) + \epsilon_i \tag{5.9}$$

where $\hat{\rho}_2$ is estimated by OLS.

What about the denominator $\lim_{d \downarrow c} d(x) - \lim_{x \uparrow c} d(x)$ of (5.8)? If again, the relevant CEFs are linear, we can estimate the discontinuity with an interacted regression equation

$$D_i = \pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot \mathbb{1}(X_i \geq c) + \pi_3 \cdot X_i \cdot \mathbb{1}(X_i \geq c) + \nu_i \tag{5.10}$$

where our estimate of $\lim_{d \downarrow c} d(x) - \lim_{x \uparrow c} d(x)$ is $\hat{\pi}_2$, the coefficient on being over the threshold.

Altogether then, our estimate of the LATE would be $\hat{\rho}_2 / \hat{\pi}_2$, a ratio of two OLS regression coefficients. Does this look familiar, from our analysis of the 2SLS estimator? It should! Even though our identification result in the RDD is different from that of an IV (based on continuity rather than independence), it can in fact be estimated using the same 2SLS estimator. Define our "instrument" to be $Z_i = \mathbb{1}(X_i \geq c)$. Then, let $\hat{\beta}_{2sls}$ be the vector of 2SLS coefficients from the "structural equation":

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \pi_3 \cdot X_i \cdot D_i + \nu_i$$

and "first stage" equation:

$$D_i = \pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot Z_i + \pi_3 \cdot X_i \cdot Z_i + \nu_i$$

2SLS then treats Z_i and $Z_i \cdot X_i$ as instruments for the endogenous variables D_i and $D_i \cdot X_i$. The 2SLS estimate of β_2 then recovers the ratio of $\hat{\rho}_2$ to $\hat{\pi}_2$, which provides a consistent estimate of the LATE among compliers at the threshold:

$$\hat{\beta}_{2,2sls} \xrightarrow{p} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c, D_i^+ > D_i^-]$$

Standard errors for $\beta_{2,2sls}$ then follow from the typical 2SLS standard errors.

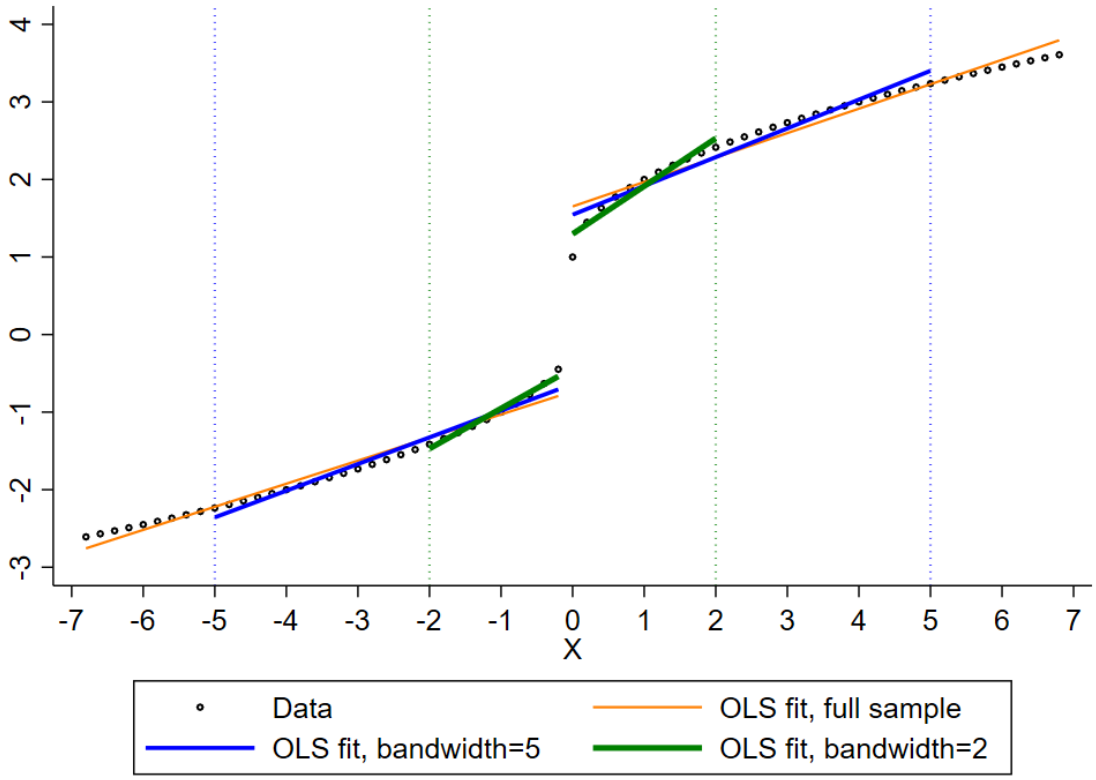
Note: The fuzzy RDD estimand has the same issues regarding functional form misspecification as in the sharp case. If $m(x)$ and $d(x)$ are not actually piecewise linear, $\beta_{2,2sls}$ will generally be inconsistent. The next Section provides an alternative method to estimation that can eliminate this risk, given a big enough sample.

5.1.5 Nonparametric estimation in the RDD*

How can we perform estimation for the RDD without assuming that the functions $m_0(x)$ and $m_1(x)$ are linear in x (and furthermore that $d_1(x)$ and $d_0(x)$ are also linear in x)?

One strategy would be to fit a polynomial function to the regression line on each side of the cutoff, by adding powers of the running variable X^2, X^3, etc. and their interactions to the OLS or 2SLS regression equation. This works, provided that the additional powers of X are sufficient to obtain a good fit. But trying to fit the regression globally, across the full range of X values in the dataset, kind of misses the point of RDD. All that really matters is that we capture the gap between $\lim_{x \uparrow c} m(x)$ and $\lim_{x \downarrow c} m(x)$ well–what matters for this is the fit of $m(x)$ *close* to c.

This leads naturally to the idea of fitting the RDD regression functions within a *bandwidth* around the cutoff, ignoring data that falls far from it. Given a bandwidth h, this is simple as dropping all data for which $|X_i - c| > h$ and estimating (5.9) or 2SLS (in the fuzzy case) within this restricted sample. If the fit of the regression equation (whether linear, quadratic, or whatever) is good on each side of the cutoff everywhere within that bandwidth, it will do a good job of quantifying the discontinuity in $m(x)$ at c. It does not need to fit well outside of that bandwidth.



The figure above shows a simulated dataset in which the regression function $m(x)$—depicted by the black dots labelled "Data"—is somewhat non-linear. The slope of $m(x)$ increases as we approach the threshold $c = 0$ from the left, and then decreases as x increases for positive x. The treatment effect $\tau(c)$ is equal to the gap between a curve that connects the black circles to the left of the cutoff and another curve that connects the black circles to the right of the cutoff. You can think of this as the gap between the rightmost black circle below $x = 0$ ($\lim_{x \uparrow 0} m(x) \approx -1$) and the leftmost black circle above $x = 0$ ($\lim_{x \downarrow 0} m(x) \approx 1$), so $\tau(c) \approx 1.5$.

When OLS is applied using the full sample, it results in the orange lines. Since they use data far from the cutoff, these underestimate the slopes of both $m_1(x)$ and $m_0(x)$ near $x = 0$, and lead to an overestimate of the treatment effect: the gap between the two orange lines at the cutoff is more like 2.5.

The (slightly thicker) blue lines reflect OLS estimation of the regression lines on either side of the cutoff restricted to a bandwidth of $h = 5$. Notice that these still underestimate the slopes of $m_0(x)$ and $m_1(x)$ close to the cutoff, but by less than the orange lines. The (still thicker) green lines use a smaller bandwidth of $h = 2$, and get even closer. Using the bandwidth of $h = 2$ results in an estimate of $\tau(c)$ that is only a little bit too high.

It would seem from the above that making the bandwidth still smaller would be better. The smaller the bandwidth, the better the linear regression predictions will approximate the values of the functions $m_1(x)$ and $m_0(x)$ close to the cutoff. But, we can't keep making h smaller and smaller without limit though, because by making h smaller we use less and less of the sample. This increases the sensitivity of our estimates to the random variability of each observation. And eventually, we'd simply run out of data!

Nonparametric regression

The idea of *non-parametric estimation* is based around the following tradeoff: as we make a statistical model more flexible (e.g. by assuming a regression function is approximately linear only locally rather than globally), we reduce bias but we also tend to increase the variance of our estimator. The optimal amount of flexibility to introduce into the model (in our case, the value of h) depends on balancing this tradeoff. If our parameter of interest is $\tau(c)$ and we denote by $\hat{\tau}_h(c)$ an estimate using a bandwidth of h, we could think of the optimal value of h as the one that minimizes the mean squared error (MSE) of the estimator, i.e. $\mathbb{E}[(\hat{\tau}_h(c) - \tau(c))^2]$.

Before discussing how we might choose this MSE-optimal value of h for the estimator $\hat{\tau}_h(c)$, let us step back and imagine the more general problem of estimating a regression function

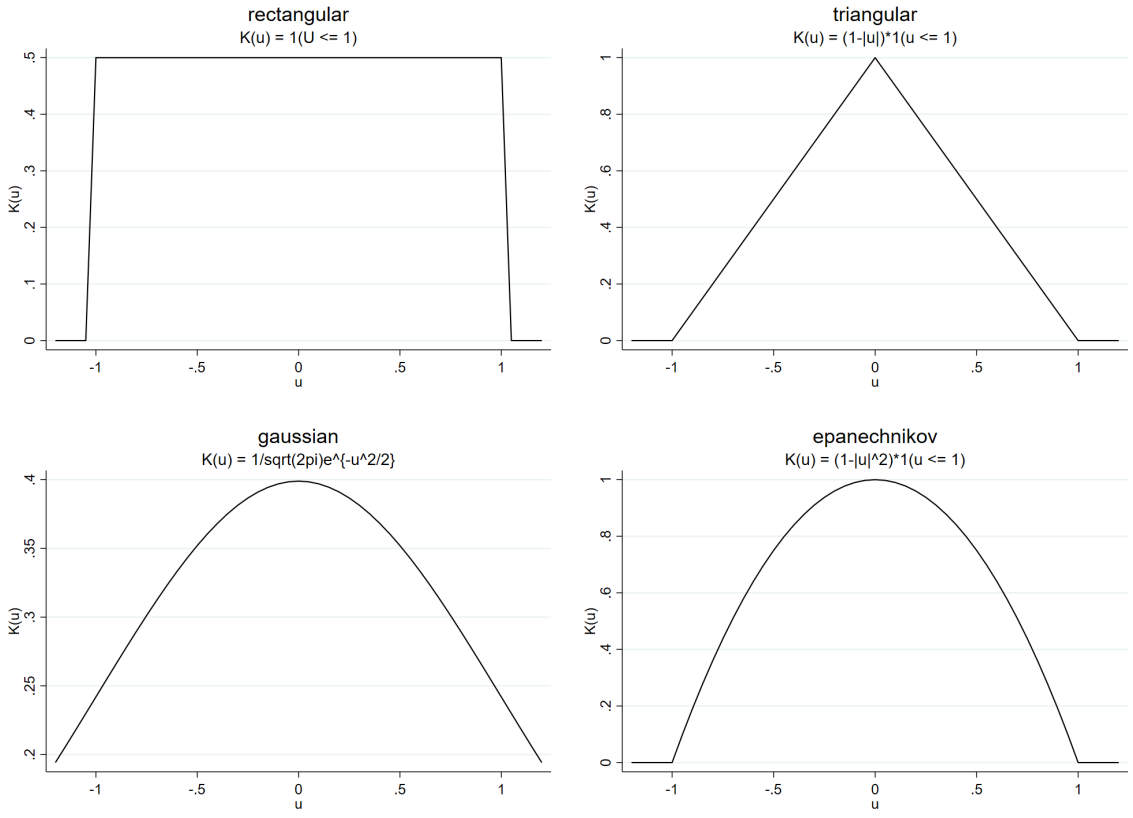
$$m(x) = \mathbb{E}[Y_i | X_i = x]$$

As we've seen, when $m(x)$ is linear in x, we can use OLS to deliver estimates of $m(x)$ at any point x. But what if we don't want to make this assumption, or $m(x)$ looks highly non-linear upon inspection of the scatterplot of Y and X? If X were discrete, we could simply compute the sample mean of Y_i at each value of X_i. When X is continuous, a natural approach would be to "bin" values of X, and simply average Y within each X.

The so-called *kernel regression* or *local-polynomial* estimator generalizes this idea of "binning" and averaging Y, and provides a foundation for the most popular non-parametric estimator in the regression discontinuity design. Suppose we're interested in constructing an estimator $\hat{m}(x_0)$ of $m(x)$ evaluated at some point $x = x_0$. Kernel regression generalizes the "bin-and-average" approach in two ways.

First, we introduce a function $K(u)$ referred to as a *kernel* function. $K(u)$ reflects the "weight" that an observation that is u bandwidths away from x_0 will receive. The "bin-and-average" approach corresponds to a so-called rectangular kernel, in which all observations for which $|X_i - x_o| \geq h$ receive equal weight, and any observation for which $|X_i - x_o| > h$ does not count at all. One undesirable feature of the rectangular kernel is that it is not continuous, which makes the estimator $\hat{m}(x_0)$ a discontinuous function of x_0.

Popular kernel functions



The triangular kernel, by contrast, down-weights observations that are farther from x_0, giving a weight that approaches zero as $|X_i - x_0|$ approaches h. Two other popular kernel functions, the Gaussian and Epanechnikov, do this in slightly different ways. The box above depicts these four choices of kernel $K(u)$. Important properties of the kernel are that it is positive, integrates to unity, has a finite second moment, and in most applications we want a kernel function to be symmetric about zero. The choice of kernel function often makes little difference in practice, but certain choices are optimal in certain settings.

Given a choice of $K(u)$ and a bandwidth h, the *Nadaraya-Watson estimator* $\hat{m}_{NW}(x_0)$ is

$$\hat{m}_{NW}(x_0) = \frac{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) \cdot Y_i}{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)}$$

Exercise: convince yourself that when using the rectangular kernel, the Nadaraya-Watson estimator evaluated at x_0 is equivalent to the "bin-and-average" approach for bin $[x_0 - h, x_0 + h]$.

Some algebra shows that $\hat{m}_{NW}(x_0)$ solves the following minimization problem

$$\hat{m}_{NW}(x_0) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \gamma_0)^2$$

$\hat{m}_{NW}(x_0)$ can thus be seen as a particular case of so-called "local polynomial estimators", which—like OLS—maximize the fit between a regression function and Y_i. The key difference is that local polynomial estimators weight the data using $K(\cdot/h)$, in order to only use data that is "close" to x_0. For example just as OLS solves:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \gamma_0 - \gamma_1 X_i)^2,$$

a local linear estimator solves

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) \cdot (Y_i - \gamma_0 - \gamma_1 X_i)^2$$

This idea generalizes to any order of polynomial in X_i, for example the local quadratic estimator would solve

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) \cdot (Y_i - \gamma_0 - \gamma_1 X_i - \gamma_2 X_i^2)^2$$

The Nadaraya-Watson estimator is the "local constant" estimator. Note that linear OLS corresponds to the local linear estimator if we let $h \to \infty$. Regardless of the order of the polynomial used, the constant $\hat{\beta}_0$ provides a non-parametric estimate of $m(x_0)$, the regression function evaluated at x_0. To estimate the full regression function $m(x)$, one estimates the local polynomial estimator at each value of x_0 (in practice, usually along a grid of many x_0 values).

The local polynomial regression estimator looks very much like OLS does, when written in terms of the data. The main difference is the introduction of a diagonal matrix $\mathbf{K}$ with value $K_{ii} = K\left(\frac{X_i - x_0}{h}\right)$. Then, the estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{K}\mathbf{Y})^{-1}$$

Implementing local polynomial estimators yourself for a given choice of h and K, you can even just use the typical OLS command, computing the K_{ii} and passing them to the regression function as weights.

Note: the idea of local-polynomial estimation can also be used to estimate a density function non-nonparametrically. This is referred to as *kernel density estimation* (KDE). Just as kernel regression generalizes the idea of averaging Y within bins of X, KDE generalizes the idea of a histogram, which counts observations within bins of X.

Nonparametric asymptotics

Given a choice of the kernel function K, how does one choose the bandwidth h? Recall that we want to choose the complexity of our model to balance bias and variance of the resulting estimator. First, note that for any estimator $\hat{m}(x)$:

$$\mathbb{E}[\hat{m}(x_0) - m(x_0))^2] = \mathbb{E}[(\hat{m}(x_0) - \mathbb{E}[\hat{m}(x_0)]) + (\mathbb{E}[\hat{m}(x_0)] - m(x_0)))^2] = Bias(\hat{m}(x_0))^2 + Var(\hat{m}(x_0))$$

where $Bias(\hat{m}(x_0)) := \mathbb{E}[\hat{m}(x_0)] - m(x_0)$. This *bias-variance decomposition* can be derived by observing that the cross-terms in the above evaluate to zero. The bias-variance decomposition lets us think carefully about how to choose h in a kernel regression. For any given sample size n, we can decrease bias by shrinking h, but at the cost of increasing variance.

To ensure the optimal balance of bias and variance, one wants to choose h in a way that does not let one term get too much larger than the other. This principle allows one to show that as the sample size increases, the optimal bandwidth h shrinks proportional to $n^{-1/5}$. In practice, one can estimate this optimal h via the data using one of several methods. One approach is known as *cross-validation*, which uses one part of the data to estimate $\hat{\beta}$ and the other to evaluate the MSE of the estimator, choosing h to optimize the out-of-sample MSE.

Note: local-polynomial regression is not the only method of non-parametrically estimating a regression function $m(x)$. Another popular method is called *series* regression, which rather than focusing on a single point x_0 attempts to fit a global function of increasing complexity, as the sample size increases.

Back to RDD

Using kernel-regression techniques, one can estimate each of the various limits in the fuzzy RDD estimand from Proposition 5.2:

$$\frac{\lim_{x \downarrow c} m(x) - \lim_{x \uparrow c} m(x)}{\lim_{x \downarrow c} d(x) - \lim_{x \uparrow c} d(x)} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c, D_i^+ > D_i^-]$$

where recall that in the case of a sharp RDD we only need to estimate the numerator of the above.

The main difference between RDD estimation and the types of non-parametric estimation problems we discussed above is that we seek left and right limits of regression functions $m(x)$ and $d(x)$, evaluated at $x_0 = c$. In practice this is accomplished by only using data from the right side of the cutoff when estimating $\lim_{x \downarrow c} m(x)$ (or analagously for $d(x)$), and only using data from the left side when estimating $\lim_{x \uparrow c} m(x)$. A good package in Stata or R for doing all of this automatically, including a data-driven choice of the bandwidth, is the `rdrobust` package.

5.1.6 Manipulation robust inference in the RDD*

5.1.7 Covariates in the RDD*

To come.

5.1.8 Quantile treatment effects in the RDD*

To come.

5.1.9 Regression discontinuity with multivalued or continuous treatments*

To come.

5.2 The regression kink design*

To come.

5.3 Using bunching for identification*

5.3.1 Bunching at a kink

To come.

5.3.2 Bunching at a notch

To come.

5.3.3 Bunching at zero

To come.

Chapter 6

Difference-in-differences

So far we've studied research designs that can be implemented using *cross-sectional* data, where observational units are indexed by i. What new opportunities for identification arise when we have data that follows such individuals i over multiple time periods? This takes us into the world of *panel data*, and opens up a new family of research design, including *difference- in-differences*, *event study*, and *fixed effects models*. This chapter is named for the first of these, which will be our canonical use of a time dimension for causal inference.

With panel data (as with repeated cross sections, see above), a single observation of an outcome variable can be written as Y_{it}, where i denotes an observational unit (such as an individual, firm, country, etc.) and t indexes a period or moment in time. "Time" here could really be any feature t of an observation that can change while i is fixed, but we'll speak as though t denotes time, as is usually the case with panel data.

Panel data vs. repeated cross sections. In *panel data*, we observe the same unit i for multiple time periods t, for example Y_{i0}, Y_{i1}, Y_{i2} are all in our dataset for one individual i, while $Y_{i'0}, Y_{i'1}, Y_{i'2}$ are in the dataset for some other individual i'.

A *repeated cross section* dataset is one in which different the set of individuals observed at one time t might be completed different then the set of individuals observed at another time t' This data structure is common in surveys, in which different individuals are sampled in different waves of the survey.

A panel dataset is called *balanced* all individuals are observed for the same set of time periods t.

6.1 Difference-in-differences with two time periods

Consider a setting in which outcomes Y_{it} are observed for the same individual at two time periods $t = 0$ and $t = 1$. A subset of individuals receive a binary treatment in 1, while no individuals are treated in period 0. Let $G_i \in \{0, 1\}$ indicate whether i is one of the individuals who is treated in the second period. We have two groups: those who are treated in the second period ($G_i = 1$), and those who are never treated ($G_i = 0$).

The classic example is Card and Krueger (1994), who looked at the effect of the minimum wage on employment over a period in 1992 in which New Jersey increased their minimum wage (from \$4.25 to \$5.05, a 20% increase). They collected data by surveying fast food establishments i in spring of 1992, which we'll call $t = 0$, and again in the fall of 1992, which we'll call $t = 1$. The minimum wage increased between the two periods, in April 1992.

As a comparison group, the researchers also surveyed fast food establishments in nearby eastern Pennsylvania. In Pennsylvania, the minimum wage remained constant at \$4.25 in both time periods. Thus we have $G_i = 1$ for each store i in NJ, and $G_i = 0$ for each store i in PA.

Recall from Chapter 1 that with a binary treatment D_i, we can eliminate selection bias if we have *random assignment*, which implies $D_i \perp Y_i(0)$. With panel data Y_{it} for our outcome, the analogous condition would be $G_i \perp Y_{it}(0)$ where we introduce time dependent potential outcomes $Y_{it}(0)$ and $Y_{it}(1)$.

These denote the outcome that would have occured for unit i in period t with and without treatment, respectively. In the minimum wage example, $Y_{it}(0)$ denotes the employment that store i would have in period t if the minimum wage were \$4.25, while $Y_{it}(1)$ denotes the employment that store i would have in period t if the minimum wage were higher, at \$5.05.

Since G_i is binary, the assumption $G_i \perp Y_{it}(0)$ would imply that $\mathbb{E}[Y_{it}(0)|G_i = 1] = \mathbb{E}[Y_{it}(0)|G_i = 0]$. This assumption would be very strong in the minimum wage example: it would say that NJ and PA would have had the same mean employment per establishment in either time period t, if the minimum wage were \$4.25. We can test this assumption directly when $t = 0$, since the minimum wage is \$4.25 in both states during the spring wave of the survey. Card and Krueger (1994) found that there were an average of 23.3 full-time equivalent workers (FTEs) in PA fast food restaurants, compared with just 20.4 in NJ. If $\mathbb{E}[Y_{it}(0)|G_i = 1] = \mathbb{E}[Y_{it}(0)|G_i = 0]$ can be verified not to hold in $t = 0$, should we have any confidence that it would hold when $t = 1$?

The *parallel trends* assumption instead says that

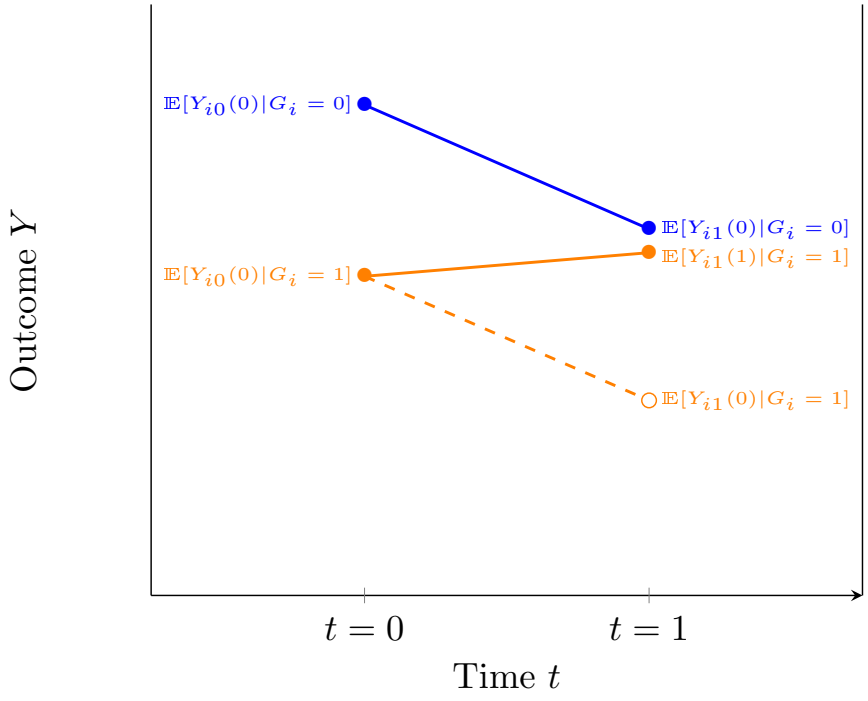
$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0] \tag{6.1}$$

i.e. the *change* in employment per restaurant between spring and fall 1992 would have been the same, on average, among restaurants in both NJ and PA had the minimum wage in both states been \$4.25. Notably Eq. (6.6) allows for PA restaurants i to have higher $Y_{it}(0)$ than NJ restaurants in both periods, as we indeed saw was the case for $t = 0$.

We call Eq. (6.6) the parallel trends assumption because it simply assumes that if it were not for the minimum wage increase in NJ, the two states would have exhibited parallel trajectories in their outcomes. These "parallel trajectories" are depicted in the figure below. The four solid circles indicate the observed means in the difference-in-differences setting: one for each combination of group G_i and time period t. In the Card and Krueger (1994) example, these values are estimated as:

- $\hat{\mathbb{E}}[Y_{i0}|G_i = 0] = \hat{\mathbb{E}}[Y_{i0}(0)|G_i = 0] = 23.3$, observed PA employment before the MW increase
- $\hat{\mathbb{E}}[Y_{i1}|G_i = 0] = \hat{\mathbb{E}}[Y_{i1}(0)|G_i = 0] = 21.2$, observed PA employment after the MW increase
- $\hat{\mathbb{E}}[Y_{i0}|G_i = 1] = \hat{\mathbb{E}}[Y_{i0}(0)|G_i = 1] = 20.4$, observed NJ employment before the MW increase
- $\hat{\mathbb{E}}[Y_{i1}|G_i = 1] = \hat{\mathbb{E}}[Y_{i1}(1)|G_i = 1] = 21.0$, observed NJ employment after the MW increase

Note that of these four values, only NJ employment per restaurant *after* the MW increase reflects the $Y_{it}(1)$ potential outcome, rather than the $Y_{it}(0)$.



The quantity $\mathbb{E}[Y_{i1}(0)|G_i = 1]$, which is a counterfactual quantity and not observed, is depicted as the open circle in the figure above. However, under the parallel trends assumption Eq. (6.6), we can *impute* its value, as

$$\begin{aligned}\mathbb{E}[Y_{i1}(0)|G_i = 1] &= \mathbb{E}[Y_{i0}(0)|G_i = 1] + \mathbb{E}[Y_{i1}(0)|G_i = 0] - \mathbb{E}[Y_{i0}(0)|G_i = 0] \\ &= \mathbb{E}[Y_{i0}|G_i = 1] + \mathbb{E}[Y_{i1}|G_i = 0] - \mathbb{E}[Y_{i0}|G_i = 0]\end{aligned}$$

In the minimum wage example, the parallel trends assumption implies that employment per restaurant in NJ in the fall of 1992 would have been 20.4+21.2-23.3=18.3 had New Jersey not increased its minimum wage in April of 1992.

Given this, we arrive at a simple estimate of the average treatment effect among restaurants in New Jersey, in the fall of 1992:

$$\begin{aligned}\mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|G_i = 1] &= \mathbb{E}[Y_{i1}(1)|G_i = 1] - \mathbb{E}[Y_{i1}(0)|G_i = 1] \\ &= \mathbb{E}[Y_{i1}|G_i = 1] - (\mathbb{E}[Y_{i0}|G_i = 1] + \mathbb{E}[Y_{i1}|G_i = 0] - \mathbb{E}[Y_{i0}|G_i = 0]) \\ &= \{\mathbb{E}[Y_{i1}|G_i = 1] - \mathbb{E}[Y_{i0}|G_i = 1]\} - \{\mathbb{E}[Y_{i1}|G_i = 0] - \mathbb{E}[Y_{i0}|G_i = 0]\} \quad (6.2)\end{aligned}$$

In the minimum wage example, this is $21.0 - 18.3 = 2.7$. The third line rewrites this as the difference between the change in NJ employment: $(21.0 - 20.4 = 0.6)$ and the change in PA employment: $(21.2 - 23.3 = -2.1)$, i.e. a difference between two differences. The result suggests that increasing the minimum wage in NJ did not decrease employment per restaurant in the fast-food industry. If anything, the estimate suggests an increase caused by the minimum wage change.

We call the LHS of Eq. (6.2) the ATT, or average treatment effect *on the treated*, because the treatment effect $Y_{i1}(1) - Y_{i1}(0)$ is averaged over individuals in the treated group $G_i = 1$ in the treatment period $t = 0$, hence it averages over the treated units in the population (NJ restaurants after the minimum wage increase).

6.1.1 Estimation in the two-period difference-in-differences model

A simple estimator of the ATT in the two period diff-in-diff model simply replaces the expectations in Eq. (6.2) by their sample counterparts, i.e.

$$\widehat{ATT} = \{\hat{\mathbb{E}}[Y_{i1}|G_i = 1] - \hat{\mathbb{E}}[Y_{i0}|G_i = 1]\} - \{\hat{\mathbb{E}}[Y_{i1}|G_i = 0] - \hat{\mathbb{E}}[Y_{i0}|G_i = 0]\} \quad (6.3)$$

A more popular way to estimate (6.2) is through a regression framework. This provides a standard error for the estimate, coming right out with the regression results.

$$Y_{it} = \alpha + \delta \cdot G_i + \gamma \cdot T_t + \beta \cdot T_t \cdot G_i + \epsilon_{it} \quad (6.4)$$

where for observation it, T_t is a dummy variable that is equal to one if $t = 1$, and zero if $t = 0$. The OLS estimator $\hat{\delta}$ is numerically identical to $\widehat{ATT}$.

Note: In class, I had the notation swapped where δ was the coefficient on $T_t \cdot G_t$, and β was the coefficient on G_i. I've switched it now to be more consistent with later notation when we move to the multi-period case, so I hope you don't get confused!

The logic of regression (6.4) can be seen as follows. First, the parallel trends assumption let's us write

$$Y_{it}(0) = \alpha + \delta \cdot G_i + \gamma \cdot T_t + \eta_{it}$$

where $\eta_{it} := Y_{it}(0) - \mathbb{E}[Y_{it}(0)|G_i, T_t]$ (the proof of this is left as an exercise). Then use that

$$Y_{it} = Y_{it}(0) + G_i \cdot T_t \cdot \{Y_{it}(1) - Y_{it}(0)\}$$

and we arrive at (6.4) with $\beta = ATT$, if we define an error term to be $\epsilon_{it} = \eta_{it} + Y_{it}(1) - Y_{it}(0) - ATT$.

Exercise: show that under the parallel trends assumption, we may write $\mathbb{E}[Y_{it}(0)|G_i] = \alpha + \delta \cdot G_i + \gamma \cdot T_t$ for some α, δ, γ.

Estimating the ATT via Eq. (6.2) allows one to incorporate covariates in a straightforward way. A common specification is

$$Y_{it} = X_i'\lambda + \delta \cdot G_i + \gamma \cdot T_t + \beta \cdot T_t \cdot G_i + \epsilon_{it} \tag{6.5}$$

where we let the constant α be included in the covariates term $X_i'\lambda$. Eq. (6.4) can motivated by a parallel trends assumption that conditions on X_i, i.e.

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1, X_i] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0, X_i] \tag{6.6}$$

coupled with an assumption that one of the expectations above is linear in X_i: $\mathbb{E}[Y_{it}(0)|G_i = 0, T_t = 0, X_i] = \alpha + X_i'\lambda$. See Abadie (2005) and Sant'Anna and Zhao (2020) for estimators that drop this linearity assumption. In the above I have included time-independent covariates X_i, which are fixed at $t = 0$ and can thus be seen as baseline characteristics of our individuals i. See Caetano et al. (2022) for a discussion of including time-varying covariates that might be themselves effected by treatment.

The OLS estimate of $\hat{\beta}$ regression Eq. (6.5) can be rewritten in the style of a so-called two way fixed effects (TWFE) regression:

$$Y_{it} = \alpha_{G_i} + \gamma_t + \beta \cdot D_{it} + \epsilon_{it} \tag{6.7}$$

The group-level fixed effect α_g replaces $\alpha + \delta \cdot G_i$ (which only varied at the group level). To clean up notation, we have rewritten $\gamma \cdot T_t$ as γ_t, and introduced the treatment indicator $D_{it} = T_t \cdot G_i$. We can take $\gamma_0 = 0$ without loss of generality, since we must omit one time period if we have not omitted one g from the group fixed effects α_g. Note that the the above regression is often written as $Y_{igt} = \alpha_g + \gamma_t + \beta \cdot D_{it} + \epsilon_{it}$, where we use the notation Y_{igt} to indicate that unit i is in treatment group g, i.e. $G_i = g$.

If one has panel data, in which the same individual is observed at different points in time, one could instead estimate TWFE estimator with *individual* fixed effects α_i:

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot D_{it} + \epsilon_{it} \tag{6.8}$$

Specification (6.8) is more flexible than that of (6.7) because the coefficients α_i do not need to be the same for all i sharing a group G_i. Therefore, it tends to be the default when one has panel data at the individual level. When one has a panel of aggregated individual-level data (say at the state or province level), then one can estimate (6.8) with these state/provincial identifiers playing the role of i.

6.2 Basic setup with multiple time periods

In this section we see how the basic difference-in-differences approach to causal inference generalizes beyond the two-period case. Suppose we now observe outcomes Y_{it} at $T + 1$ different time periods labeled $\{0, 1, \ldots T\}$.

With more than two periods to consider $T > 1$, we now have some new terminology to discuss. For instance, rather than estimating "the effect" of being treatment, we can talk about the effect of being treated "k periods ago", for some $k \in \{1, 2, \ldots T\}$.

Dynamic treatment effects consider the effect of being treated k periods ago. For instance, in a setup where some units receive treatment at $t = 1$, while others never receive treatment. Then with observations of Y_{it} from $t = 0$ to $t = T$, we can talk about the effect of being treated this period, one period ago, two periods ago, etc., all the way to T periods ago. These are called *dynamic treatment effects*. When dynamic treatment effects are indexed by how long ago treatment was received (what we've been calling k), we say that we are considering an *event-study* design.

Staggered adoption/event study design: starting with the above example, suppose that in addition to the never-treated group and the group that receives treatment in period $t = 1$, there is a third group that receives treatment in period $t = 2$, another group that receives treatment at $t = 3$, and so on. We call this a setting of *staggered adoption*. Staggered adoption is standard for the event-study design.

Absorbing treatment: We call the treatment *absorbing* if after unit i receives treatment, they remain "treated" for all periods following it (Callaway and Sant'Anna 2021 call this condition "irreversibility of treatment"). In more complicated settings, one might imagine that units become "untreated" again at some later time, possibly become treated a second time, etc. We will focus on an absorbing treatment

in this section, but generalizations exist to non-absorbing treatments as well.

In practice, researchers often use the terms "difference-in-differences design" and "event-study design" fairly interchangeably. If somebody says "event-study", they definately have a case in which there are more than two time periods in mind. However it's all one research design, in the sense that they both use the same identifying assumption: the parallel trends assumption.

6.2.1 Notation for the timing of treatment

As indicated above, suppose the outcome Y_{it} is observed for time periods $t = 0 \dots T$ for some finite $T > 1$. Consider an *absorbing* treatment D_{it}, meaning that if $D_{it} = 1$ for some i and t, then $D_{it'} = 1$ for all $t' > t$ as well.

Recall that in the two period difference-in-differences model, we let $G_i = 1$ indicate the units i that received treatment in period $t = 1$, and $G_i = 0$ indicate the units that never receive treatment. Generalizing this, we now let $G_i = 2$ for the units that receive treatment in period $t = 2$, $G_i = 3$ for the units that receive treatment in period $t = 3$, and so on. In general, we let G_i indicate the first period at which unit i receives treatment. However, instead of indicating the group that never receives treatment as $G_i = 0$ (as we did with the control group in a two-period difference-in-differences model), we denote these units as having $G_i = \infty$. The group $G_i = 0$ instead represent "always-treated" units, since $D_{it} = 1$ for all $t = 0, \dots T$. However, in most applications, there are no always-treated units because the data typically begins before any units are treated.

The above notation lets us write a simple and general treatment assignment formula:

$$D_{it} = \begin{cases} 1 & \text{if } t \geq G_i \\ 0 & \text{otherwise} \end{cases}$$

i.e. unit i has been treated in period t if $t \geq G_i$.

The above formula lets us define a binary treatment in a context with staggered treatment adoption, to think about the effect of *having been treated* at some point in the past. However, researchers are usually interested in a more detailed kind of treatment effect: what is the effect of having been treated k periods ago? This leads to a potential outcomes notation based on when a unit is first treated, which we introduce in the next section.

6.2.2 Potential outcomes based on treatment timing

Now we introduce potential outcomes to let us define dynamic treatment effects. To do so, we consider counterfactuals based on the *timing* of treatment. For a unit i that was for example first treated at $t = 2$, the relevant thought experiment is the following: what if i had instead first been treated at $t = 3$, or $t = 1$, or not at all?

Accordingly, let $Y_{it}(0)$ indicate the outcome for unit i at period t if they never receive treatment at any time before T. For any $g \geq 1$, let $Y_{it}(g)$ denote the outcome that unit i would receive at period t if they were first treated in period g. Observed outcomes are $Y_{it} = Y_{it}(0)$ if $G_i = \infty$ and $Y_{it} = Y_{it}(G_i)$ otherwise

With this notation, we can define our dynamic treatment effects as

$$\Delta_{it}(g) = Y_{it}(g) - Y_{it}(0)$$

which denotes the effect on period t outcomes of being treated at time period g relative to not being treated at all. We can define an "average treatment on the treated" type parameter for any combination of g and t:

$$ATT(g,t) = \mathbb{E}[Y_{it}(g) - Y_{it}(0) | G_i = g] = \mathbb{E}[\Delta_{it}(g) | G_i = g] \tag{6.9}$$

$ATT(g,t)$ measures the average effect of having first been treated in period g, rather than not at all, among those units that are actually first treated in period g.

Finally, note that we can write unit i's observed outcome at time t in terms of dynamic treatment effects, as:

$$Y_{it} = Y_{it}(0) + \sum_g \mathbb{1}(G_i = g) \{Y_{it}(k) - Y_{it}(0)\} = Y_{it}(0) + \sum_g \mathbb{1}(G_i = g) \cdot \Delta_{it}(g) \tag{6.10}$$

where the sum is over all the groups $g = 1, 2, \dots T$ aside from the never-treated.

6.2.3 Event time

When discussing dynamic treatment effects, it will be convenient to refer to think of the effects difference $t - G_i$ as *event time*. When event time is positive $t \geq G_i$, it measures the length of exposure of unit i to the treatment by period t. or example, if $G_i = 4$ for a given unit i, then period $t = 5$ represents event time k of $k = 1$. In period $t = 6$, this unit experiences event time of $k = 2$, etc. The period before i is treated represents event time $k = -1$. One might look for anticipatory effects or violations of paralell trends by considering negative values of event time. We call t *calendar time* to distinguish it from event time.

6.2.4 Parallel trends with multiple time periods

Just as in the two period case, the central identifying assumption in the multi-period difference-in-differences design is that individuals who differ in their treatment status would have followed parallel trends in the absence of treatment. We now have potentially several such groups to consider, but we can again define the parallel trends assumption making reference only to the untreated potential outcome $Y_{it}(0)$.

We can define a "strong version" of parallel trends that for identification of dynamic treatment effects as the following:

Definition 6.1. *We say that* ***parallel trends*** *holds when the evolution of untreated potential outcomes $Y_{it}(0)$ over time t follows the same trend on average for all groups defined by treatment timing, that is:*

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = g']$$

Sometimes it is more reasonable to assume that parallel trends holds conditional on some observed covariates X_i.

Definition 6.2. *We say that* ***conditional parallel trends*** *holds when the evolution of untreated potential outcomes $Y_{it}(0)$ over time t follows the same trend on average for all groups defined by treatment timing, that is:*

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = g, X_i] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = g', X_i]$$

Here I have used time invariant covariates X_i that are assumed to be unaffected (e.g. measured prior to) by treatment. For a generalization to time-varying X_{it} and a discussion of avoiding so-called *bad control* problems, see Caetano et al. (2022).

The second identifying assumption, in addition to the parallel-trends assumption, is the following:

Definition 6.3. *The* ***no-anticipation assumption*** *states that for some $\delta \geq 0$, $\mathbb{E}[Y_{it}(g)|G_i = g] = \mathbb{E}[Y_{it}(0)|G_i = g]$ for all $t < g - \delta$.*

This assumptions states that more than δ periods before treatment, there is no average effect of treatment. This is the same as saying that $ATT(g,t) = 0$ for all $t < g - \delta$. However, average treatment effects "before" treatment may be non-zero for $g - \delta \leq t < g$. In many applications, one however assumes that no-anticipation holds with $\delta = 0$, on the grounds that units do not react in an anticipatory way to treatment before it occurs.

Note that the no-anticipation assumption was made implicitly in the two-period difference-in-differences model, when we replaced $Y_{i0}(0)$ by Y_{i0} among the treated group. If making use of conditional parallel trends, the no-anticipation assumptions should be made conditional on X_i.

6.3 The two way fixed effects estimator and its pitfalls

Recall from Section 6.1.1 that in the two-period difference-in-differences model, we can use the so-called two way fixed effects (TWFE) estimator, which estimates β from the regression

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot D_{it} + \epsilon_{it} \tag{6.11}$$

by fixed effects OLS, where $D_{it} = T_t \cdot G_i$ is an indicator for receipt of treatment by unit i in time period t. In the mulit-period model, one might estimate this same equation, with the treatment indicator D_{it}

replaced by $\mathbb{1}(t \geq G_i)$ for the multi-period case. This is sometimes referred to as the "static" TWFE regression.

More frequently, authors focus on TWFE regressions for dynamic treatment effects, the vector of coefficients β in the regression

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\substack{k=L\ldots U \\ k \neq \delta - 1}} \beta_k \cdot D_{itk} + \epsilon_{it} \tag{6.12}$$

where for $t < U$, we define $D_{itk} := \mathbb{1}(G_i = t-k) = \mathbb{1}(t = G_i + k)$. That is, rather than considering the effect of treatment being received by time t: D_{it}, one estimates the effect of treatment being first received k periods ago, for various values of k ranging from some minimum lead $L \leq 0$ to some maximum lag $U > 0$.

Binning: The intended interpretation of the coefficient β_k is the effect of receiving treatment k periods ago. As we'll see in the next section, choosing a value $L > -T$ and/or $U < T$ amounts to assuming that these effects are zero for $k \leq L$ or for $k \geq T$. Provided that $L < -\delta$, one would expect that $\beta_k = 0$ for all $k < -\delta$. An assumption that is more plausible in many contexts is that treatment effects are constant in event time for $k \geq U$. In this case, one should estimate (6.12) with $D_{itU} := \mathbb{1}(t \geq G_i + k)$. See Schmidheiny and Siegloch (2023) for more details on this practice of "binning", and the eqivalence between TWFE models with binning and "distributed lag models" that put leads and lags of the treatment indicator D_{it} on the RHS.

Omitting base-periods: Note that specification (6.12) omits the period $k = \delta - 1$ from the treatment effect sum. Most commonly, researchers take $\delta = 0$ and thus β_{-1} is omitted, while β_k for $k < -1$ are used to assess the validity of the parallel trends assumption. However, Borusyak et al. (2022) point out that if L and U are set far enough from zero to include all possible treatment leads and lags (what they call a *fully dynamic specification*), then a perfect multicollinearity problem arises unless two coefficients β_k are dropped. We can already deduce that one β_k needs to be dropped, in order to avoid $\sum_k D_{itk} = 1$.

The reason that we still need to drop a *second* β_k is that one can always add a linear time trend to the γ and subtract it off by changing the time and event-time effects. First note that we can write the sum $\sum_k \beta_k \cdot D_{itk}$ in (6.12) as β_{t-G_i}. Then note that:

$$\begin{aligned}
\alpha_i + \gamma_t + \beta_{t-G_i} &= (\alpha_i - \lambda \cdot G_i) + (\gamma_t + \lambda \cdot t) + (\beta_{t-G_i} - \lambda \cdot (t - G_i)) \\
&= (\alpha_i - \lambda \cdot G_i) + (\gamma_t + \lambda \cdot t) + \sum_k (\beta_k - \lambda \cdot k) \cdot D_{itk}
\end{aligned}$$

One thus needs to be careful, because in a fully dynamic specification statistical software will often drop one of the β_k arbitrarily, and the numerical estimate of one's entire vector of β_k will shift accordingly. To avoid the problem: one needs to solve the underidentification problem by imposing some additional structure on the event-time effects β_k. One solution is to simply increase L by one, which amounts to setting $\beta_L = 0$. Binning also avoids this issue, as discussed by Schmidheiny and Siegloch (2023). However, binning does not solve the next issue we discuss, which is what TWFE estimates when treatment effects are heterogeneous.

Misspecification due to treatment effect heterogeneity: Goodman-Bacon (2021) shows that given standard parallel trends assumption like (6.1) for the multi-period case, the OLS estimate of β from Eq. (6.11) can be written as a linear combination of 2×2 diff-in-diff estimates. However, these 2×2 estimators sometimes compare treated units with already-treated units. Similar considerations apply to the event-study specification (6.12) as well, see Chaisemartin and D'Haultfœuille (2020) for details.

6.3.1 When TWFE works: homogenous treatment effects in event-time*

To develop some intuition for the TWFE estimator, let us see how it can be justified under the parallel trends assumption if dynamic treatment effects are homogenous in "event time". By homogenous treatment effects, we don't mean that the effect of being treated k periods ago is the same for all k (in this case we would not need to use multiple β_k), but instead that for each fixed k, the effect of being treated k periods ago is the same for all i.

Similar to Eq. (6.10) which expressed observed outcomes in terms of dynamic treatment effects in calendar time, one can instead express observed outcomes in terms of dynamic treatment effects in event time. Define

$$\beta_{ik} = Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(0) = \Delta_{i,G_i+k}(G_i)$$

to be $i's$ effect k periods after the receive treatment at G_i. Then, relabeling the sum in (6.10) to be over $k = t - g$ instead of g:

$$Y_{it} = Y_{it}(0) + \sum_k \mathbb{1}(G_i = t - k) \cdot \Delta_{it}(t - k) = Y_{it}(0) + \sum_k \mathbb{1}(G_i = t - k) \cdot \beta_{ik} \tag{6.13}$$

where the sum is over all $k \in -T, \ldots, 0, \ldots T$.

Homogenous treatment effects in event time says that

$$\beta_{ik} = \beta_k \text{ for all } i$$

Given homogenous treatment effects in event time, note that we can write Eq. (6.13) as

$$\begin{aligned} Y_{it} &= Y_{it}(0) + \sum_k \mathbb{1}(G_i = t - k) \cdot \beta_k \\ &= \underbrace{Y_{i0}(0)}_{\alpha_i} + \underbrace{\mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)]}_{\gamma_t} + \left\{ \sum_k \beta_k \cdot \mathbb{1}(G_i = t - k) \right\} + \epsilon_{it} \end{aligned} \tag{6.14}$$

where we let

$$\epsilon_{it} := Y_{it}(0) - Y_{i0}(0) - (\mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)])$$

Given the parallel trends assumption 6.1, we have for any g and t that

$$\mathbb{E}[\epsilon_{it}|G_i = g] = \mathbb{E}[Y_{it}(0)|G_i = g] - \mathbb{E}[Y_{i0}(0)|G_i = g] - (\mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)]) = 0$$

since $\mathbb{E}[Y_{it}(0)|G_i = g] - \mathbb{E}[Y_{i0}(0)|G_i = g] = \mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)]$.[1]
Eq. (6.14) matches the form of the TWFE regression (6.12), but unlike the TWFE regression (6.12), (6.14) sums over *all* possible event times k. This highlights the important role of binning treatment effects β_k outside of some window $L \ldots U$, as discussed in the last section.

Note: TWFE "works" more generally if instead of assuming perfectly homogenous treatment effects in event time, we assume that

$$\mathbb{E}[\beta_{ik}|G_i = g] = \beta_k \text{ for all } g \tag{6.15}$$

i.e. there is a version of "no selection on gains" in the sense that earlier and later treated groups do not have differential treatment effects. In this case, we can write

$$\begin{aligned} Y_{it} &= Y_{it}(0) + \sum_k \mathbb{1}(G_i = t - k) \cdot \beta_{ik} \\ &= \underbrace{Y_{i0}(0)}_{\alpha_i} + \underbrace{\mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)]}_{\gamma_t} + \left\{ \sum_k \beta_k \cdot \mathbb{1}(G_i = t - k) \right\} + \epsilon_{it} \end{aligned} \tag{6.16}$$

where we let

$$\epsilon_{it} := Y_{it}(0) - Y_{i0}(0) - (\mathbb{E}[Y_{it}(0)] - \mathbb{E}[Y_{i0}(0)]) + (\beta_{i,t-G_i} - \beta_{t-G_i})$$

where we still have that $\mathbb{E}[\epsilon_{it}|G_i = g]$ since the conditional expectation of the last term is

$$\mathbb{E}[\beta_{i,t-G_i} - \beta_{t-G_i}|G_i = g] = \mathbb{E}[\beta_{i,t-g}|G_i = g] - \beta_{t-g} = \beta_{t-g} - \beta_{t-g} = 0$$

by assumption of (6.15).

[1] To show that the parameters of Eq. (6.14) can be consistently estimated by fixed effects regression, $\mathbb{E}[\epsilon_{it}|G_i = g]$ is not obviously quite enough, since we also have the time and unit fixed effects in the model. A sufficient condition is *strict exogeneity* (see e.g. p620 of the Hansen textbook) which in our context says that $\mathbb{E}[\epsilon_{it} \cdot G_i]$ and $\mathbb{E}[\epsilon_{it} \cdot \gamma_s]$ for any $s = 1 \ldots T$. The first of these follows from what we've shown above, since $\mathbb{E}[\epsilon_{it} \cdot G_i] = \mathbb{E}\{G_i \cdot \mathbb{E}[\epsilon_{it}|G_i]\} = 0$, and the latter follows because γ_s is not random (it is just a number given the time-period of interest s), so $\mathbb{E}[\epsilon_{it} \cdot \gamma_s] = \gamma_s \cdot \mathbb{E}[\epsilon_{it}] = 0$.

6.4 What can go wrong with TWFE*

For simplicity, let us first consider the "static" specification (6.11) in which we aim a single coefficient β on D_{it}, an indicator for having received treatment by period t. This is the case considered by Goodman-Bacon (2021), and helps establish intuition for how TWFE can fail. Very similar considerations apply to the dynamic treatment-effect specification (6.12), as studied by Chaisemartin and D'Haultfœuille (2020).

I copy Eq. (6.11) here for quick reference:

$$Y_{it} = \alpha_i + \gamma_t + \beta \cdot D_{it} + \epsilon_{it}$$

Goodman-Bacon (2021) shows that the OLS estimate $\hat{\beta}$ can be written as a linear combination of a large number of simple two-period, two-group difference-in-differences comparisons. These comparisons always pick two groups g and g', where g' is treated later than g. Given $g < g'$, let

$$\bar{Y}_i^{post(g)} := \frac{1}{T-g} \sum_{i=g}^{T} Y_{it}$$

be the average observed outcome for unit i after time $t = g$ (until the end of the sample at $t = T$). Similarly, define

$$\bar{Y}_i^{pre(g)} := \frac{1}{g} \sum_{i=0}^{g-1} Y_{it} \qquad \bar{Y}_i^{mid(g,g')} := \frac{1}{g'-g} \sum_{i=g}^{g'-1} Y_{it}$$

be the average observed outcome for unit i *before* time $t = g$, and the average outcome for unit i across all time periods between g andg', respectively.

With this notation Goodman-Bacon (2021) shows that $\hat{\beta}$ can be written as a linear combination of terms taking the following three forms:

$$\hat{\beta}^{g;\text{never}} = (\hat{\mathbb{E}}[\bar{Y}_i^{post(g)}|G_i = g] - \hat{\mathbb{E}}[\bar{Y}_i^{pre(g)}|G_i = g]) - (\hat{\mathbb{E}}[\bar{Y}_i^{post(g)}|G_i = \infty] - \hat{\mathbb{E}}[\bar{Y}_i^{pre(g)}|G_i = \infty])$$

$$\hat{\beta}^{g,g';\text{mid/pre}} = (\hat{\mathbb{E}}[\bar{Y}_i^{mid(g,g')}|G_i = g] - \hat{\mathbb{E}}[\bar{Y}_i^{pre(g)}|G_i = g]) - (\hat{\mathbb{E}}[\bar{Y}_i^{mid(g,g')}|G_i = g'] - \hat{\mathbb{E}}[\bar{Y}_i^{pre(g)}|G_i = g'])$$

$$\hat{\beta}^{g',g;\text{post/mid}} = (\hat{\mathbb{E}}[\bar{Y}_i^{post(g)}|G_i = g'] - \hat{\mathbb{E}}[\bar{Y}_i^{mid(g,g')}|G_i = g']) - (\hat{\mathbb{E}}[\bar{Y}_i^{post(g)}|G_i = g] - \hat{\mathbb{E}}[\bar{Y}_i^{mid(g,g')}|G_i = g])$$

With T time periods, there are on the order of T^2 such terms for various g and g'. The coefficients on each term depend on the number of observations in each group, and share of time that each group spends in treatment. These coefficients are all (weakly) positive.

The quantity $\hat{\beta}^{g;\text{never}}$ is the most straightforward of the three defined above. It simply compares treatment group g during all of their treated periods to their untreated periods, and differences this with respect to the never-treated group $G_i = \infty$. The second type of term, $\hat{\beta}^{g,g';\text{ mid/pre}}$, compares the outcomes of the earlier treatment group g to those of the later treatment group g', differencing the mean outcomes in between periods g and g' (during which time the earlier treatment group is treated but the later group is not). The final variety of term, $\hat{\beta}^{g',g;\text{post/mid}}$, compares the later treated group to the earlier one, taking the difference in mean outcomes in the post period (when both groups are treated) to the middle period between g and g' in which the later group g' is not treated, but the earlier group g is.

The presence of this final type of term $\hat{\beta}^{g',g;\text{post/mid}}$ is where TWFE can lead you astray. Since D_{it} is not changing for group g during the periods between g and g', OLS wants to use it as a control group. But if the effect of treatment lasts more than one period (in event time), then the outcomes Y_{it} are still changing for these observations *due to treatment*. Since these prolonged treatment effects get subtracted in $\hat{\beta}^{g',g;\text{post/mid}}$, the estimand β treates some of the treatment effects of earlier-treated groups as if they were part of the common time trend, which it is attempting to eliminate from β by differencing.

As a result, if one decomposes the probability limit β of the TWFE estimate $\hat{\beta}$ into average treatment effect parameters $ATT(g,t)$, it turns out that some group/time combinations can receive *negative* weights. In principle, this means that even if all of the units receive a positive treatment effect (at all time horizons k), one might still end up with a negative value of β. Yikes!

Note: for simplicity, I have here assumed a balanced panel. If some of the observations between g and g' for unit i are for example missing, $\bar{Y}_i^{between(g,g')}$ would be defined as the average among the non-missing

observations, and unit i would receive correspondingly less weight in an overall average like.

While this section has illustrated the potential problems of TWFE in the case of the static specification (6.11), the same threat of negative weights persists when we consider the dynamic difference-in-differences specification (6.12). The basic reason is the same: when estimating the coefficient β_k, OLS will want to use already-treated groups $g < t-1$ as comparison groups when differencing outcomes between t and $t-1$ among units whose value of D_{itk} increases between $k-1$ and k (i.e. those for whome $G_i = t-k$). But with dynamic treatment effects that persist for many periods, these control units g might still be experiencing changes in their Y_{it} coming from the treatment, and these dynamic treatment effects will incorrectly get subtracted out in the estimator $\hat{\beta}_k$. This issue is analyzed in Chaisemartin and D'Haultfœuille (2020).

Note that if treatment effects are homogenous across units, negative weights are not a problem per-se. Chaisemartin and D'Haultfœuille (2020) show that the weights on group-time specific average treatment effect parameters add up to one. So, if they are all the same as one another, they'll add up to the right number. However heterogeneity should be expected in general, so the threat of negative weights is a serious one in principle.

For a nice illustration of the Goodman-Bacon (2021) result for the static TWFE specification, see https://andrewcbaker.netlify.app/2019/09/25/difference-in-differences-methodology/. For simulation illustrations in the case of estimating dynamic treatment effects via TWFE, see https://bcallaway11.github.io/did/articles/TWFE.html.

6.5 Constructing estimators that allow for general treatment effect heterogeneity*

In light of the problems for TWFE decsribed in the last section, I know outline one approach to identification and estimation in the multi-period difference-in-differences model that avoids these issues by carefully weighting over the two-group difference-in-differences estimands. This method was proposed by Callaway and Sant'Anna (2021), and is I think particularly illuminating because it is explicit about estimating over the $ATT(g,t)$ and then aggregating over them. However there are several alternatives that are also popular and might be preferable in certain circumstances, see e.g. Sun and Abraham (2021), Chaisemartin and D'Haultfœuille (2020), Borusyak et al. (2022), and Gardner (2022). The approach of Chaisemartin and D'Haultfœuille (2020) in particular does not assume an absorbing treatment, so may be useful in cases where treatment switches on and off or occurs in multiple spells.

6.5.1 Identification using never-treated units

Note that when there is a never-treated group, we can focus on the implication of Assumption 6.1 that

$$\mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,t-1}(0)|G_i = \infty] = \mathbb{E}[Y_{it} - Y_{i,t-1}|G_i = \infty]$$

Adding this up across all periods up to t:

$$\mathbb{E}[Y_{it}(0) - Y_{i0}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,0}(0)|G_i = \infty] = \mathbb{E}[Y_{it} - Y_{i0}|G_i = \infty]$$

for all g, where we take $g' = \infty$ to represent the never-treated group. The key consequence of the above is that, for any $g \geq 1$ and t:

$$\begin{aligned}
ATT(g,t) &= \mathbb{E}[Y_{it}(g) - Y_{it}(0)|G_i = g] = \mathbb{E}[Y_{it}(g)|G_i = g] - \mathbb{E}[Y_{it}(0)|G_i = g] \\
&= \mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i0}(0)|G_i = g] - \{\mathbb{E}[Y_{it}(0)|G_i = g] - \mathbb{E}[Y_{i0}(0)|G_i = g]\} \\
&= \mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i0}|G_i = g] - \{\mathbb{E}[Y_{it}(0)|G_i = \infty] - \mathbb{E}[Y_{i0}(0)|G_i = \infty]\} \\
&= (\mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i0}|G_i = g]) - (\mathbb{E}[Y_{it}|G_i = \infty] - \mathbb{E}[Y_{i0}|G_i = \infty])
\end{aligned}$$

where we've used that $Y_{it} = Y_{it}(G_i)$ and $Y_{i0}(0) = Y_{i0}$ by the no-anticipation assumption. Parallel trends has also been used, to replace the term in brackets.

Notice that each term on the RHS of the final equation above is identified by the observable data: a simple difference in differences that compares treatment group g to the never-treated group. Intuitively,

we can identify $ATT(g,t)$ by comparing the time evolution of the $G_i = g$ units between periods 0 and t with the evolution of the never-treated units between periods 0 and t.

Callaway and Sant'Anna (2021) show how the above result extends to the case of *conditional* parallel trends. In this case one must estimate a propensity score function $p_g(x)$ that yields the probability of being in treatment group $G_i = g$ given $X_i = x$.

6.5.2 Identification using not-yet-treated units

In some empirical applications, there are no never-treated units, and so we cannot use the result of the last section to identify the $ATT(g,t)$. Even if we do have such never-treated units, not relying on them for identificaiton might be desirable anyways, since the units that *never* receive treatment might be very different from all of the other units, who do receive treatment at some point.

Instead, we might seek to use as a control group for outcomes at t among those with $G_i = g$ other units having with a $G_i > t+\delta$, i.e. the not-yet treated units who should not be showing any anticipation effects yet at time t. Combined with the no-anticipation assumption, we will use the implication of parallel trends (Assumption 6.1) that for any $t \geq g - \delta$ and $s < g - \delta$:

$$\mathbb{E}[Y_{it}(0) - Y_{is}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{is}(0)|D_{i,t+\delta} = 0] = \mathbb{E}[Y_{it} - Y_{is}|D_{i,t+\delta} = 0]$$

where notice that the event $D_{i,t+\delta} = 0$ is the same as $G_i > t + \delta$.

To see that this is implied by parallel trends, note that

$$\begin{aligned}
\mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|D_{i,t+\delta} = 0] &= \mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|G_i > t + \delta] \\
&= \sum_{g' > t+\delta} P(G_i = g'|G_i > t + \delta) \cdot \mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|G_i = g'] \\
&= \mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|G_i = g] \cdot \underbrace{\left(\sum_{g' > t+\delta} P(G_i = g'|G_i > t + \delta) \right)}_{=1}
\end{aligned}$$

where we have used parallel trends in the last step, which adding over subsequent time periods between s and t yields $\mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,s}(0)|G_i = g']$.

Now choosing $s = g - \delta - 1$, the last period at which units treated at g are gauranteed to be unaffected by treatment given the no-anticipation assumption:

$$\mathbb{E}[Y_{it}(0) - Y_{i,g-\delta-1}(0)|G_i = g] = \mathbb{E}[Y_{it}(0) - Y_{i,g-\delta-1}(0)|D_{i,t+\delta} = 0] = \mathbb{E}[Y_{it} - Y_{i,g-\delta-1}|D_{i,t+\delta} = 0] \quad (6.17)$$

and re-arranging therefore:

$$\mathbb{E}[Y_{it}(0)|G_i = g] = \mathbb{E}[Y_{i,g-\delta-1}|G_i = g] + \mathbb{E}[Y_{it}|D_{i,t+\delta} = 0] - \mathbb{E}[Y_{i,g-\delta-1}|D_{i,t+\delta} = 0]$$

using that $Y_{i,g-\delta-1} = Y_{i,g-\delta-1}(0)$ for any unit first treated at g. Thus, for any $t \geq g - \delta$, $ATT(g,t)$ is identified as:

$$\begin{aligned}
ATT(g,t) &= \mathbb{E}[Y_{it}(g) - Y_{it}(0)|G_i = g] \\
&= (\mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i,g-\delta-1}|G_i = g]) - (\mathbb{E}[Y_{it}|D_{i,t+\delta} = 0] - \mathbb{E}[Y_{i,g-\delta-1}|D_{i,t+\delta} = 0])
\end{aligned}$$

where we have used (6.17) to replace the term in brackets.

Notice that the RHS of the last line above is a difference in differences that compares treatment group g to at time t to not-yet-treated units at time $g - \delta - 1$, the last period before the $G_i = g$ can begin to exhibit anticipation effects. Note that we can be sure that all the not-yet-treated units at time t ($D_{it} = 0$) are also free of anticipation effects at $g - \delta - 1$.

6.5.3 Aggregating the group-time specific average effects for the TWFE target parameters

The last two sections have shown that we can identify the parameter $ATT(g,t)$ for various choices of treatment group g and time period t. How should we summarize and report this potentially vary large set of treatment effect parameters?

The most common choice is to mimic the coefficients β_k that appear in the TWFE estimating equation, by constructing average treatment effects by length of exposure, or "event-time" k. In particular, when we have never-treated units $G_i = \infty$, we might seek to estimate

$$\begin{aligned}\beta_k &:= \mathbb{E}[\beta_{ik}|G_i \neq \infty] = \mathbb{E}[Y_{i,G_i+k}(G_i) - Y_{i,G_i+k}(0)|G_i \neq \infty] \\ &= \sum_{g \neq \infty} P(G_i = g|G_i \neq \infty) \cdot \mathbb{E}[Y_{i,g+k}(g) - Y_{i,g+k}(0)|G_i = g] \\ &= \sum_{g \neq \infty} P(G_i = g|G_i \neq \infty) \cdot ATT(g, g+k)\end{aligned}$$

Recall that when using never-treated units as controls, and when there are no covariates X, the $ATT(g,t)$ is equal to $(\mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i0}|G_i = g]) - (\mathbb{E}[Y_{it}|G_i = \infty] - \mathbb{E}[Y_{i0}|G_i = \infty])$. Thus, β_k can be writen as

$$\beta_k = \sum_{g \neq \infty} P(G_i = g|G_i \neq \infty) \cdot \{(\mathbb{E}[Y_{i,g+k}|G_i = g] - \mathbb{E}[Y_{i0}|G_i = g]) - (\mathbb{E}[Y_{i,g+k}|G_i = \infty] - \mathbb{E}[Y_{i0}|G_i = \infty])\}$$

Of course, for k large or small enough, there may be no units in the data for which $Y_{i,g+k}$ is actually observed in the data. Thus our target parameter β_k will instead need to aggregate only over g for which $g + k \leq T$, which is a stronger condition than $G_i \neq \infty$:

$$\begin{aligned}\beta_k &:= \mathbb{E}[\beta_{ik}|G_i + k \leq T] \\ &= \sum_{g \leq T-k} P(G_i = g|G_i + k \leq T) \cdot \{(\mathbb{E}[Y_{i,g+k}|G_i = g] - \mathbb{E}[Y_{i0}|G_i = g]) - (\mathbb{E}[Y_{i,g+k}|G_i = \infty] - \mathbb{E}[Y_{i0}|G_i = \infty])\}\end{aligned}$$

When instead using not-yet-treated units rather than never-treated units as controls, recall that $ATT(g,t) = (\mathbb{E}[Y_{it}|G_i = g] - \mathbb{E}[Y_{i,g-\delta-1}|G_i = g]) - (\mathbb{E}[Y_{it}|D_{it} = 0] - \mathbb{E}[Y_{i,g-\delta-1}|D_{it} = 0])$. Thus, we can instead identify the parameter β_k via

$$\begin{aligned}\beta_k &:= \mathbb{E}[\beta_{ik}|G_i + k \leq T] \\ &= \sum_{g \leq T-k} P(G_i = g|G_i + k \leq T) \cdot \{\mathbb{E}[Y_{i,g+k}|G_i = g] - \mathbb{E}[Y_{i,g-1}|G_i = g]) - (\mathbb{E}[Y_{i,g+k}|D_{it} = 0] - \mathbb{E}[Y_{i,g-1}|D_{it} = 0])\}\end{aligned}$$

where I've taken $\delta = 0$ for simplicity. If one wants to estimate β_k to assess pre-trends for some $L, L+1, \cdots -2$, one would instead use a $\delta < L$ as the comparison time period.

Callaway and Sant'Anna (2021) discuss how one can further limit the groups g that appear in the definition of β_k, with the goal of increasing the comparability of β_k and $\beta_{k'}$ for different values k' and k. To achieve this, on can "balance" the groups with respect to event-time to ensure that the same g appear in each β_k.

Callaway and Sant'Anna (2021) also discuss how the $ATT(g,t)$ parameters can be aggregated to provide a single overall summary measure of the effect of treatment, rather than separating it out by length of exposure k. This target parameter is analogous to the one researchers have in mind when estimatinga TWFE regression like (6.11), with a single coefficient on treatment D_{it}.

Estimating the above expressions for β_k is straightforward when there are no covariates. When one is instead leveraging a *conditional* parallel trends assumption with covariates X_i, one must estimate conditional expectations of the form $\mathbb{E}[Y_{it}|X_i = x, G_i = g]$ as well as group propensity scores $P(G_i = g|X_i = x)$ to construct the analogs of the above expressions. Callaway and Sant'Anna (2021) suggest a "doubly-robust" approach to these estimation tasks that helps to guard against misspecification of how these functions depend upon x, based upon Sant'Anna and Zhao (2020). See Callaway and Sant'Anna (2021) for details both for the case with and the case without covariates.

6.6 Assessing and relaxing the parallel trends assumption using pre-treatment observations

To come.

6.7 Difference-in-differences with a continuous treatment variable

To come.

Appendices

Appendix A

Probability

A.1 Probability spaces

Main idea: A *probability function* ascribes a number to each of a collection of *events*, where each event is a set of *outcomes*.

This section develops the mathematical notion of probability. Probability is a function that associates a number between zero and one to *events*. Events, in turn, are sets of *outcomes*. It's easiest to think of outcomes in the context of a process that could have multiple distinct results, like flipping a coin or randomly choosing a number from a phone book.

A.1.1 Outcomes and events

We begin with a set Ω of conceivable outcomes, which is referred to as the *sample space* or *outcome space*.

Examples: When flipping a coin, the sample space is $\Omega = \{H, T\}$, corresponding to "heads" or "tails", respectively. When rolling a six-sided die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. When drawing a card from a 52-card deck, the sample space can de denoted as a combination of a card-value and a suit, or $\{(n, s) : n \in \{A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}, s \in \{hearts, spades, diamonds, clubs\}\}$. When using a random number generator to draw any number between 0 and 1, the sample space is $\Omega = [0, 1]$.

We denote a generic element of the sample space as as $\omega \in \Omega$. What we call *events* are simply sets of such ω, i.e. subsets of Ω. But in general, not all subsets of Ω necessarily need to be events. Rather, we consider a collection of sets F, referred to as an *event space*.

Definition A.1. *An event space F is a collection of subsets $A \subseteq \Omega$.*

In all of the examples given above, the outcome space Ω has a finite number of elements. In such cases, it is typical to choose F to be the collection of *all* subsets of Ω. This collection is referred to as the *powerset* of Ω and is often denoted as 2^Ω. As an example, the powerset of the set $\{1, 2\}$ is $2^{\{1,2\}} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$. When we consider Ω that are uncountable sets (for example when Ω is a continuum), we'll need to restrict the event-space, as discussed below.

A.1.2 The probability of an event

A *probability function* P associates a positive real number to each event $A \in F$.

Definition A.2. *A probability function $P(\cdot)$ is a function from F to $\mathbb{R}$, satisfying the following properties:*

1. $P(A) \geq 0$ *for each* $A \in F$
2. $P(\Omega) = 1$
3. *If $A_1, A_2 \ldots$ is a countable collection of disjoint sets (i.e. $A_j \cap A_k = \emptyset$ for any $j \neq k$), then*

$$P\left(\bigcup_j A_j\right) = \sum_j P(A_j)$$

This formulation of probability is sometimes referred to as the *Kolmogorov axioms* of probability.

These axioms imply several intuitive properties of probability. For example, if A has a countable number of elements, then the third property in Definition A.2 implies that:

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

provided that $\{\omega\} \in F$ for each $\omega \in A$. In particular, this result implies that for a finite set A we can simply sum up the probability of each of the outcomes in A. For example, for a six-sided die $P(even) = P(\{2\}) + P(\{4\}) + P(\{6\})$.

A few other properties of probability functions are left as exercises. As practice, I'll include a proof of the familiar property that $P(A^c) = 1 - P(A)$. To see this, note that A and A^c are disjoint sets, and that $A \cup A^c = \Omega$. Thus, by the third property of Definition A.2 $P(\Omega) = P(A) + P(A^c)$. Then use the second property to obtain the result.

Exercise: Show that if $A \subseteq B$: $P(A) \leq P(B)$.

Exercise: Use the above to show that $P(A \cap B) \leq \min\{P(A), P(B)\}$.

Exercise: Derive the expression: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exercise: Derive the expression: $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. *Hint:* use $(A \cap B)^c = A^c \cup B^c$.

A.1.3 Which sets of outcomes get a probability?

In addition to the Kolmogorov axioms for the function P, we also place some requirements on the event space F. In particular, we require it to be a σ-algebra:

Definition A.3. *A σ-algebra on Ω is a collection F of subsets of Ω with the following properties:*

1. $\Omega \in F$
2. *If any $A \in F$, then $A^c \in F$, where A^c is the complement of A in Ω*
3. *If $A_1, A_2 \ldots$ are each in F, then $\bigcup_j A_j$ is also in F*

Recall that events $A \in F$ are those subsets of Ω that the function P must ascribe a probability (these sets A are called *measurable sets*). The first item above, that $\Omega \in F$, was already assumed by item 2. of Definition A.2: we can always associate a probability with the whole outcome space, and that probability is one. Item 2. of Definition A.3 says that if we are willing to give a probability to event A, then we should also be willing to give a probability to the event that A does not happen, i.e. A^c. The third property assures that given events A and B, we can always talk about the probability of A or B, which is $P(A \cup B)$.

Note that all of the properties of a σ-algebra tell us about things that must be in F, they guarantee that F is not to "small". The biggest collection of subsets of Ω is the set of all of its subsets: the powerset 2^Ω. The powerset of Ω is always a $\sigma-$algebra (exercise: check that it satisfies all three properties). However, using 2^Ω as the event space F can also be too *big* for certain applications. This is why it is necessary to introduce the idea of a σ-algebra.

Example: Consider as an outcome space the entire unit interval: $\Omega = [0, 1]$. It turns out that it is impossible to define a "uniform" probability function on this Ω, if we insist on using the whole powerset of $[0, 1]$ as our event space F. That is, there is no function $P(\cdot)$ satisfying Kolmogorov's axioms, and defined over all $A \in 2^{[0,1]}$, that satisfies our intuitive notion that moving a set around in the unit interval does not change its probability. See Proposition 1.2.6 of Rosenthal (2006) for details.

This example demonstrates that in some cases we may need to work with something smaller than 2^Ω. In particular, issues like the above arise when Ω is uncountably infinite, e.g. corresponding to a continuum of numbers. When Ω is finite or countable, it usually makes sense to consider the full powerset of Ω as our event space. When we are in the uncountable case (e.g. when Ω is a convex subset of the real line $[a, b]$), we typically appeal to the Borel σ-algebra:

Definition A.4. *The Borel σ-algebra $\mathcal{B}$ is the collection that consists of all intervals of the forms $[a, b]$, $(a, b]$, $[a, b)$, (a, b), and all other sets in $\mathbb{R}$ that are then implied by the definition of a σ-algebra.*

Exercise: Show that for any Ω, the collection $\{\emptyset, \Omega\}$ is a $\sigma-$algebra.

Exercise: Show that $\emptyset \in F$ if F is a $\sigma-$algebra.

Exercise: Show that $\sigma-$algebras are closed under countable intersections, that is $\bigcap_j A_j$ is in F if $A_1, A_2, \ldots$ are each in F.

A.1.4 Bringing it all together: a probability space

Once we have a sample space, event space, and probability function, we refer to them altogether as a probability space (sometimes called a *probability triple*).

Definition A.5. *An probability space is a triple (Ω, F, P) in which F is a σ-algebra defined on Ω, and P is a probability function defined on F.*

A.2 Random variables

Main idea: If we associate a number to each outcome in a probability space, we have what is called a *random variable*.

A.2.1 Definition

Most data we use in econometrics is quantitative in nature, so its natural to think of probability spaces in which the outcome space is composed of numbers. Many of the examples have this feature already, for example $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a six-sided die. But even when the ω do not have an immediate numeric interpretation, we can define a random variable by associating a number to each outcome ω:

Definition A.6. *Given a probability space (Ω, F, P), a random variable X is a function $X : \Omega \to \mathbb{R}$.*

Example: Suppose I randomly select a student in this class, which I represent by a probability space with $\Omega = \{\text{all students in this class}\}$, $F = 2^\Omega$, and $P(\{\omega\}) = 1/|\Omega|$ for each $\omega \in \Omega$. If we let $X(\omega)$ denote the height in inches of student ω.

A random variable X defined from a primitive probability space (Ω, F, P) allows us to define a new probability space in which the outcomes are real numbers. We can now define a new probability function P_X on sets of real numbers, using the original probability function P on Ω:

$$P_X(A) := P(\{\omega \in \Omega : X(\omega) \in A\}) \tag{A.1}$$

Technical note: observe that the above definition gives a way to associate a probability P_X with any set A of real numbers, provided that $\{\omega \in \Omega : X(\omega) \in A\} \in F$. To ensure this condition holds it is typical to restrict to sets A that belong to the Borel algebra $\mathcal{B}$ defined in Section A.1.3, and further insist that the function X is *measurable*. X being measurable is a technical condition that just means that for any $x \in \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \leq x\} \in F$. Our new probability space can now be denoted as $(\mathbb{R}, \mathcal{B}, P_X)$.

A *realization* of random variable X is the specific value $X(\omega)$ that it ends up taking, given ω. While X is a *function*, $X(\omega)$ is a *number*. Lowercase letters x are often used to denote numbers that are possible realizations: e.g. $x = X(\omega)$ for some $\omega \in \Omega$.

A.2.2 Notation

The notation of Equation (A.1) is pretty cumbersome to work with, so the convention is to simplify it in a few ways.

Let's start with an example. If we're interested in the probability that $X(\omega)$ is less than or equal to 5, we'll typically write this as: $P(X \leq 5)$, which can be interpreted as $P_X(A)$ where $A = (-\infty, 5]$, or equivalently: $P(\{\omega \in \Omega : X(\omega) \leq 5\})$. What's changed in this notation? Let's go through step-by-step:

- First, we haven't bothered with the subscript X on P_X like in Equation (A.1) because it's clear from what's inside the parentheses that we're talking about random variable X.
- Second, inside the function P we're using the language of *conditions* rather than sets. That is, rather than writing out the set $A = (-\infty, 5]$ of values we're interested in, we just write this as a condition: "≤ 5".
- Third, we've made ω implicit and written X rather than $X(\omega)$. However, you often see ω left in. For example, we might write $P(X(\omega) = x)$ for the probability that X takes a value of x. In the context of Equation (A.1), this maps onto $P_X(\{x\})$, or equivalently $P(\{\omega \in \Omega : X(\omega) = x)\})$.

Given that we're using the language of conditions, we often write "and" inside probabilities, for example: $P(X \leq 5 \text{ and } X \geq 2)$. The "and" operation translates into intersection in the language of sets: $P(\{\omega \in \Omega : X(\omega) \leq 5\} \cap \{\omega \in \Omega : X(\omega) \geq 2\})$. Similarly, "or" translates into the union of sets: $P(X \leq 5 \text{ or } X \geq 2) = P(\{\omega \in \Omega : X(\omega) \leq 5\} \cup \{\omega \in \Omega : X(\omega) \geq 2\})$.

Note: We may have multiple random variables, e.g. X could be a randomly chosen state's minimum wage, while Y their unemployment rate. Mathematically, these two random variables correspond to functions $X(\cdot)$ and $Y(\cdot)$ applied to a common underlying outcome space Ω, which in this case corresponds to the set of US states. Probabilities like $P(X \leq \$10 \text{ and } Y \leq 5\%)$ are interpreted as $P(\{\omega \in \Omega : X(\omega) \leq \$10 \text{ and } Y(\omega) \leq 5\%\})$. If $P(\{\omega\}) = 1/50$ for all ω, then this probability is in turn equal to the number of states that have a minimum wage less than or equal \$10 and an unemployment rate less than or equal 5%, divided by 50.

A.3 The distribution of a random variable

Main idea: The *cumulative distribution function* (CDF) provides a concise and convenient way to represent the probability function of a random variable or of multiple random variables. From the CDF we can define everything else we use to work with specific types of random variables, for example *probability density functions* and *probability mass functions*.

A.3.1 Central concept: the cumulative distribution function

We can summarize the probability function over values of a random variable X through the so-called *cumulative distribution function* or CDF of X.

Definition A.7. *The cumulative distribution function of X is the function $F_X(x) := P(X \leq x)$.*

Note that $F_X(x)$ is a function from $\mathbb{R}$ to the unit interval $[0, 1]$, that is $F_X(x)$ is defined for all $x \in \mathbb{R}$ and $F_X(x)$ is always between zero and one. The following properties can be proven to hold for any random variable X:

- $F_X(x)$ is a weakly increasing function, that is $F_X(x') \geq F_X(x)$ if $x' > x$
- $\lim_{x \downarrow -\infty} F_X(x) = 0$
- $\lim_{x \uparrow \infty} F_X(x) = 1$
- $F_X(x)$ is right-continuous, i.e. $F_X(x) = \lim_{\epsilon \downarrow 0} F_X(x + \epsilon)$

Note on notation: when the context is clear, we often denote a CDF as $F(x)$ rather than $F_X(x)$. However, when we have multiple random variables like X and Y, we may need the notation $F_X(x)$ and $F_Y(y)$ to be clear about which variable we are referring to. When using the notation $F(x)$ for a CDF, keep in mind that this is not the same "F" as we used to denote the event space of a generic probability triple (Ω, F, P).

From the CDF, we can derive anything we'll need to know about a single random variable. When we have multiple random variables, the *joint-CDF* tells us everything we need to know about them.

Definition A.8. *The joint-CDF of two random variables X and Y is the function*

$$F_{XY}(x, y) := P(X \leq x \text{ and } Y \leq y)$$

We'll come back to the joint-CDF of two (or more) random variables in Section A.5.

Although the CDF $F(x)$ of a random variable is a function of a single variable x, we can use it to recover the probability that X lies in a *set*. For example, consider the set $(a, b]$, that is all numbers between a and b, including b itself.

Proposition A.1. *For any numbers a and b such that $b \geq a$, $P(X \in (a, b]) = F(b) - F(a)$*

Proof. Given that $P(A) = 1 - P(A^c)$, (see Section A.1.2), we have that:

$$P(X \in (a, b]) = P(a < X \leq b) = 1 - P(X \leq a \text{ or } X > b)$$

Using the third property of a probability function, we have that $P(X \leq a \text{ or } X > b) = P(X \leq a) + P(X > b)$, since the sets $\{x \in \mathbb{R} : x \leq a\}$ and $\{x \in \mathbb{R} : x > b\}$ are disjoint. Thus:

$$P(X \in (a, b]) = 1 - \{P(X \leq a) + P(X > b)\} = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

where I've used that $P(X \leq b) = 1 - P(X > b)$. □

More generally, we can from the function $F(x)$ compute the probability that $X \in A$ for any *Borel-measurable* set, that is a set A that belongs to the Borel $\sigma-$algebra. Sets that are simple intervals on the real line like $(a, b]$ are the leading example of such sets. Computing the probability associated with more complicated sets that aren't intervals is also possible using the CDF. The next section develops two functions that can be derived from the CDF, and are sometimes easier to work with for such computations.

A.3.2 Probability mass and density functions

Let X be a random variable with CDF $F(x)$. We often refer to the whole function F as the *distribution* of X. It always tells us everything we need to know about X. But there are two important special cases in which we can represent the distribution of X in an alternative way that is often more convenient.

A.3.2.1 Case 1: Discrete random variables and the probability mass function

Call $\mathcal{X}$ a *discrete set* if $\mathcal{X}$ contains a finite number of elements, or a countably infinte number of elements (e.g. $\mathcal{X} = \mathbb{N}$, the set of all integers).

Definition A.9. *A discrete random variable X is a random variable such that $P(X \in \mathcal{X}) = 1$ for some discrete set $\mathcal{X}$.*

Example: If X is the number returned by rolling a die, then X is a discrete random variable because $P(X \in \{1, 2, 3, 4, 5, 6\}) = 1$.

For any random variable, we call the smallest set $\mathcal{X}$ for which $P(X \in \mathcal{X})$ the *support* of X. A discrete random variable has as its support a discrete set.

When X is a discrete random variable, its CDF ends up looking like a staircase: flat everywhere except at each x in its support, where it "jumps" up by an amount $P(X = x)$. For example, for a six-sided die:

Note: The open/closed dots at e.g. $x = 1$ indicate the $F(1)$ is equal to $1/6$, and not 0 (although it is equal to 0 for x arbitrarily close but to the left of 1). We see from this graph why CDFs are right-continuous but not necessarily left-continuous.

At each point in its support $\{1, 2, 3, 4, 5, 6\}$, the CDF for the die jumps by $P(X = x)$, or $1/6$. This is a general feature of discrete random variables. Thus, rather than use the CDF function $F(x)$ to represent the distribution of X, we can just keep track of where it jumps and by how much. To do this, we use *probability mass function* or *p.m.f.* of X

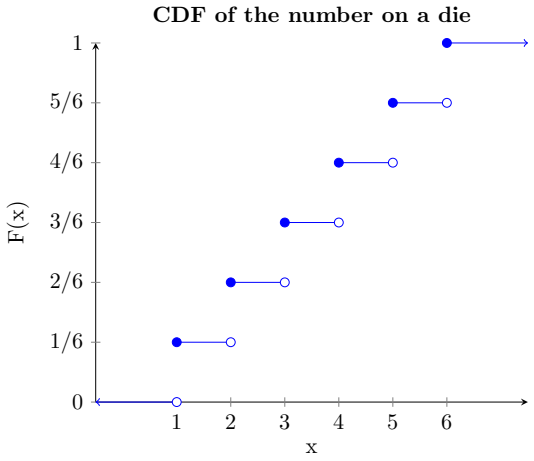


Figure A.1: The CDF of the number returned by a fair six-sided die.

Definition A.10. *The probability mass function of a random variable X is the function $\pi(x) = P(X = x)$*

For a discrete random variable, we can express the p.m.f. alternatively as a sequence, rather than a function. Label the points in the support of X as $\{x_1, x_2, x_3, \dots\}$, in increasing order so that $x_1 < x_2 < x_3 \dots$. Let x_j denote the j^{th} value in this sequence. For any j, let $\pi_j = \pi(x_j) = P(X = x_j)$.

The sequence of probabilities $\{\pi_1, \pi_2, \pi_3, \dots\}$ coupled with the sequence of support points $\{x_1, x_2, x_3, \dots\}$ carries exactly the same information as the full CDF.

Obtaining the p.m.f. from the CDF: For a given support point x_j: $\pi_j = F(x_j) - F(x_{j-1})$, and for any x: $\pi(x) = \lim_{\epsilon \downarrow 0} F(x) - F(x - \epsilon)$. Note that $\pi(x) = 0$ for any x that is not a support point, and F is continuous $\{x_1, x_2, x_3, \dots\}$.

Obtaining the CDF from the p.m.f (only possible for a discrete random variable): $F(x) = \sum_{j:x_j \leq x} \pi_j$.

Note that from this last expression, we can see that since $\lim_{x \to \infty} F(x) = 1$, we must have that $\sum_j \pi_j = 1$ – probability mass functions sum to one when the sum is taken across all support points j.

A.3.2.2 Case 2: Continuous random variables and the probability density function

For random variables that are not discrete, knowing the probability mass function isn't sufficient to recover the whole CDF. Often $P(X = x) = 0$ for all x, so the p.m.f does not even really tell us anything useful about X's distribution.

An important class of random variables that are not discrete are random variables for whom the CDF is differentiable for all x. When it is, we can define the *probability density function* or p.d.f. of X.

Definition A.11. *The probability density function of a random variable X having a differentiable CDF $F(x)$, is $f(x) = \frac{d}{dx} F(x)$.*

We will refer to random variables that have a density function $f(x)$ as *continuous* random variables (another phrasing is that X is *continuously distributed*). Recall that for a function to be differentiable, it must be continuous; thus, the CDF of a continuous random variable must be continuous, lacking any jumps like those that characterize the CDF of a discrete random variable.

Note: you may see in various texts a few different notions of "continuity" of a random variable. For the purposes of this class, a continuous random variable is a random variable with a continuous CDF, which is basically equivalent to it being differentiable everywhere in its support. We won't worry about the distinction between these two things: e.g. random variables with CDFs that are continuous but non-differentiable.

For a continuous random variable we can use the p.d.f rather than the CDF to calculate anything we need to know. For example the probability that X lies in any interval $[a, b]$ can be obtained by integrating over the density function:

$$P(X \in [a, b]) = \int_a^b f(x)dx \tag{A.2}$$

Intuitively, this gives us the area under the curve $f(x)$ between points a and b, as depicted in Figure A.2. Note that $\int_a^b f(x)dx = F(b) - F(a)$, because the CDF is the anti-derivative of the p.d.f.

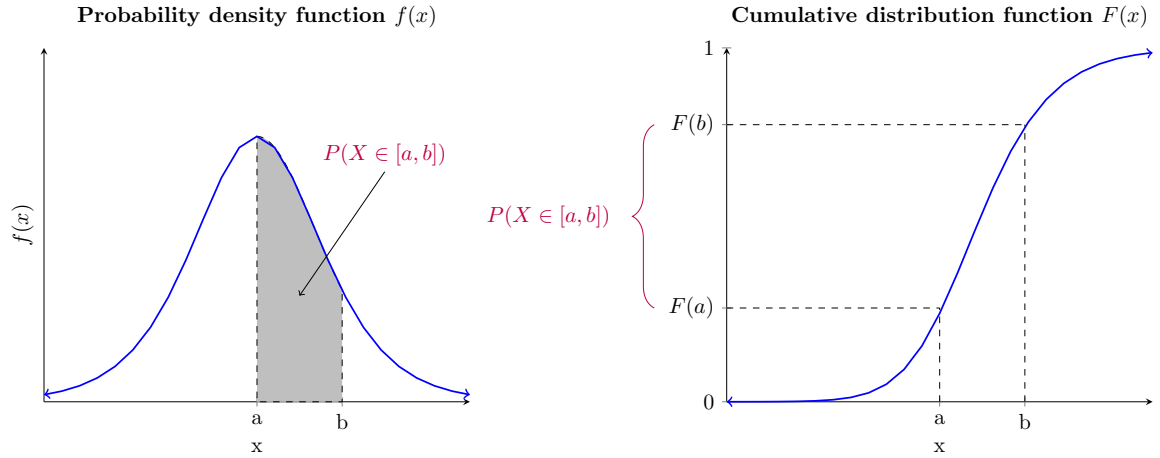


Figure A.2: The left panel depicts an example of the p.d.f. $f(x)$ of a random variable X. The probability that $a \leq X \leq b$ is given by the area under the $f(x)$ curve between $x = a$ and $x = b$. $P(a \leq X \leq b)$ is also equal to $F(b) - F(a)$, the difference in the CDF of X evaluated at $x = b$ and at $x = a$, as depicted in the right panel.

While the probability mass function $\pi(x)$ gives us the probability that X equals x exactly, the p.d.f does not tell us the probability that $X = x$ (in fact for any x: $P(X = x) = 0$ for a continuous random variable!).

Rather $f(x)$ can be interpreted as telling us the probability that X is close to x, in the following sense. Consider a point x and some small $\epsilon > 0$. Recall the definition of $f(x)$ as the derivative of $F(x)$:

$$f(x) = \frac{d}{dx}F(x) = \lim_{\epsilon \to 0} \frac{F(x + \epsilon) - F(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(X \in (x, \epsilon])}{\epsilon}$$

where we've used Proposition A.1 to replace $F(x + \epsilon) - F(x)$ with $P(X \in (x, \epsilon])$. Thus $f(x)$ is limit of the ratio of the probability that X lies in a small interval that beings at x, and the width ϵ of that interval. Note also that for small ϵ: $F(x + \epsilon) \approx F(x) + f(x) \cdot \epsilon$, which is called the first-order *Taylor approximation* to $F(x + \epsilon)$ around x.

Let us end this section with a few properties of a probably density function:

- From Eq. (A.2), we see that the density must integrate to one, when the integral is taken over the whole real line, i.e. $\int_{-\infty}^{\infty} f(x)dx = 1$.
- since $F(x)$ is increasing and $f(x)$ is its derivative, $f(x)$ is *positive* everywhere: $f(x) \geq 0$.

A.3.2.3 Case 3 (everything else): mixed distributions

Although most familiar examples of random variables are either discrete or continuous, a given random variable X need not be either. However, a powerful result known as the *Lebesque decomposition theorem* shows that we can combine the two tools we've just developed: the p.m.f. and the p.d.f., to work with any random variable.

Definition A.12. *Given two random variables X and Y with CDFs F_X and F_Y, a third random variable Z is called a mixture of X and Y if it has a CDF that for some $p \in (0, 1)$ satisfies $F_Z(t) = p \cdot F_X(t) + (1 - p) \cdot F_Y(t)$, for all t.*

The Lebesgue decomposition theorem says that a generic random variable X can be seen as a "mixture" of a discrete random variable and a continuous one, that is

$$F(x) = p \cdot F_{discrete} + (1-p) \cdot F_{continuous} \tag{A.3}$$

for some $p \in (0,1)$, where $F_{discrete}$ admits of a probability mass function, and $F_{continuous}$ admits of a probability density function (i.e. is differentiable everywhere). The support points of $F_{discrete}$ are often referred to as mass points of F.

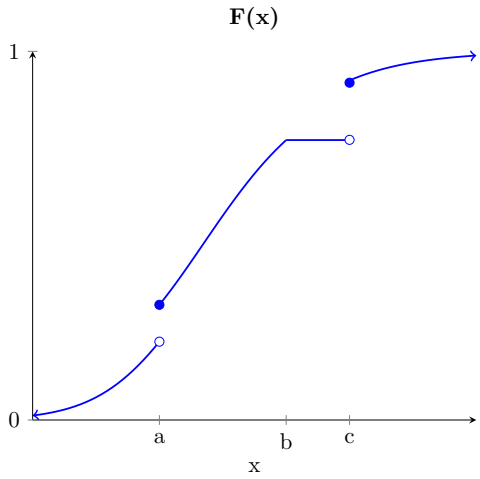


Figure A.3: An example of the CDF of a mixed random variable. This example has mass points at a and c, where the CDF jumps discretely. It is continuous everywhere else, and is differentiable everwhere except $\{a, b, c\}$.

There are some technical aspects to stating the Lebesque decomposition theorem formally, which we won't explore here. Rather, it's easiest to think of this result visually: a generic CDF is any increasing function bounded between 0 and 1 (which is also right-continuous). The jumps in $F(x)$ define the discrete part of X (note that it can only jump up, and not down, since F is increasing). The function $F(x)$ will be differentiable almost everywhere else, defining it's continuous part.[1]

Note for the interested: to explicitly generate decomposition (A.3), first collect the locations x_j and sizes y_j of each of the jumps $j = 1, 2, \ldots$ in $F(x)$. Then $\pi(x_j) = \sum_j y_j$, and $\pi_j = y_j/p$ yields a well-defined p.m.f. function. This characterizes $F_{discrete}$. For any remaining point where $F(x)$ is differentiable, we define a density $f_{continuous}(x) = \frac{1}{1-p}\frac{d}{dx}F(x)$, which characterizes $F_{continuous}$. Note that there may be points at which $F(x)$ doesn't jump, but also isn't differentiable, such as point b in Figure A.3. We can safely ignore such points, since they are isolated and have probability zero, e.g. $P(X = \{b\}) = 0$.

A.3.3 Marginal and joint distributions

Recall that when we have two random variables X and Y, we have defined the joint CDF $F_{XY}(x, y) = P(X \leq x, Y \leq y)$ as well as the individual CDFs: $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$. The functions F_X and F_Y are often referred to as the *marginal distributions* of X and Y.

The following relationships hold between marginal and joint distributions:

- $F_X(x) = F_{XY}(x, \infty) = P(X \leq x, Y \leq \infty) = P(X \leq x)$. Similarly, $F_Y(y) = F_{XY}(\infty, y)$.
- If Y is discrete: $P(X = x) = \sum_j P(X = x \text{ and } Y = y_j)$ where y_j are the support points of Y
- If X and Y are both continuously distributed: $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$, where the joint density $f_{XY}(x, y)$ is the derivative of the joint CDF with respect to both x and y: $f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y)$.

[1] "Almost everywhere" here has a technical meaning. Any monotonic function is guaranteed to be differentiable everywhere except at isolated points: see Lebesque's theorem for the differentiability of a monotone function.

Intuitively, we can obtain the marginal distribution of X from the joint distribution by summing or integrating over all values of Y, and we can similarly derive the marginal distribution of Y from the joint distribution of X and Y by summing/integrating over values of X.

The above results all follow from a fundamental identity for probabilities called the *law of total probability*:

Proposition (law of total probability): Consider a countable collection of events $A_1, A_2, \ldots$ that partition the sample space (this means that the A_j are disjoint and that $\bigcup_j A_j = \Omega$). Then for any event B: $P(B) = \sum_j P(B \cap A_j)$.

Proof. The proof is good practice, so I include it here. Since any event $B \subseteq \Omega$, $B = B \cap \Omega$ and thus $P(B) = P(B \cap \Omega)$. Now, since $\bigcup_j A_j = \Omega$, we have that $P(B) = P\left(B \cap (\bigcup_j A_j)\right)$. Observe that $B \cap (\bigcup_j A_j) = \bigcup_j (B \cap A_j)$, and that the events $(B \cap A_j)$ are disjoint for different values of j (since each is a subset of A_j). Thus, $P(B) = \sum_j P(B \cap A_j)$, proving the result. □

We can use the ideas of marginal and joint distributions to define the notion of *independence* between two random variables:

Definition A.13. *We say that random variables X and Y are independent if $F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$ for all x and y.*

When X and Y are independent, we denote this fact as $X \perp\!\!\!\perp Y$. When they are not, we say $X \not\perp\!\!\!\perp Y$.

A.3.4 Functions of a random variable

An important property of random variables is that we can apply a function to a random variable, and this results in a new random variable. For example, if we start with a random variable X, and have a function $g : \mathbb{R} \to \mathbb{R}$, then $g(X)$ is also a random variable. For example, $X + 1$ defines a new random variable that is one larger than X for all i.

The reason that we can do this is simple: the original random variable was defined from a function X defined on an underlying outcome space Ω. Evaluating $g(X(\omega))$ for any $\omega \in \Omega$ yields a new function, the so-called composition of g with X (this is often denoted as $g \circ f$).

Technical note: Recall that the function $X(\omega)$ that defines the original random variable X must be a *measurable* function. For the above logic to go through, the function $g(\cdot)$ applied to X must also be measurable, so that the function $g \circ f = g(X(\cdot))$ is also measurable. A sufficient condition for a function to be measurable is that it is piece-wise continuous, which is a very weak condition.

To work with a random variable $Y = g(X)$, we need to know it's CDF, which is:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

This RHS expression can always be evaluated using the CDF of X. However, there are two important special cases in which deriving the distribution of Y from that of X is particularly easy:

1. If X has a discrete distribution with support points $x_1, x_2, \ldots$ and p.m.f. $\pi_1, \pi_2, \ldots$, then Y has the same p.m.f. $\pi_1, \pi_2, \ldots$ but at new support points $g(x_1), g(x_2), \ldots$.
    - Example: if X is a random variable that takes value 0 with probability p and 1 with probability $1 - p$, then the random variable $Y = X + 1$ is a random variable that takes value 1 with probability p and 2 with probability $1 - p$.
2. (homework problem) If X has a continuous distribution with density $f_X(x)$, and if the function $g(x)$ is strictly increasing and differentiable with derivavtive g', then Y has a density $f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}$ where g^{-1} is the inverse function of g.
    - Example: if $g(x) = \log(x)$, then $f_Y(y) = f_X(e^y) \cdot e^y$, since $g^{-1}(y) = e^y$ and $g'(x) = 1/x$.

Just as a function applied to a random variable defines a new random variable, functions applied to *multiple* random variables also yield a new random variable. For example, if X and Y are each random variables, then $Z = g(X, Y)$ is also a random variable, where $g(x, y)$ is now a function that takes two arguments. Some examples would be the random variables $X + Y$, $X \cdot Y$, or $\min\{X, Y\}$. When taking a functions of two random variables, $Z = g(X, Y)$, we need the full *joint distribution* of X and Y to derive the CDF of Z. Knowing the two functions $F_X(x)$ and $F_Y(y)$ is generally not enough, rather we need to know the function $F_{XY}(x, y)$ (see Definition A.8). This will come up later in the course.

A.4 The expected value of a random variable

Main idea: The *expected value* of a random variable is a measure of its average value across realizations. In the special case of a continuous random variable, its value can be obtained by an integral involving the density function. In the special case of a discrete random variable, its value can be obtained by a sum involving the probability mass function.

The expected value (a.k.a. *expectation value*, or simply *expectation*) of a random variable is a measure of its average value over all possible realizations. The expectation of X is denoted $\mathbb{E}[X]$.

To motivate how $\mathbb{E}[X]$ will be defined, think of task of computing the average of a list of numbers. For example, the average of the numbers 1, 2, 2, and 4 is $(1 + 2 + 2 + 4)/4 = 2$. Notice that the number 2 occurred twice in the series, so we added 2 to the sum two times. We could thus have written the averaging calculation as $\frac{1}{4}(1 \cdot 1 + 2 \cdot 2 + 4 \cdot 1)$, where each number is mulitplied by the number of times it occurs in the list. The general formula could be written

$$\text{average of a list of numbers} = \sum_j (j^{th} \text{ distinct number }) \cdot \underbrace{\frac{\#\text{ times } j^{th} \text{ distinct number occurs in the list}}{\text{length of the list}}}_{w_j}$$

where notice that "weight" w_j on the j^{th} distinct number sums to one over all j, i.e. $\sum_j w_j = 1$.

The definition of $\mathbb{E}[X]$ for a discrete random variable is exactly analogous to this formula, where we average over the values x_j that X can take, and use as "weights" the probabilities π_j:

$$\mathbb{E}[X] = \sum_j x_j \cdot \pi_j \tag{A.4}$$

where $x_1, x_2, \ldots$ are the distinct support points of the random variable and π_j is it's p.m.f. Note that the π_j sum to one, as we saw in Section A.3.2.

In the case of a continuous random variable, the analogous expression to Eq. (A.4) replaces the sum with an integral, and the probability $\pi_j = \pi(x_j)$ is replaced by $f(x)dx$:

$$\mathbb{E}[X] = \int x \cdot f(x) dx \tag{A.5}$$

The quantity $f(x) \cdot dx$ can be interpreted as the probability that X lies in an interval $[x, x + dx]$ having a very small width dx, as discussed in Section A.3.2.

A.4.1 General definition*

We now give a general definition of the expectation of a random variable X, and see that Equations (A.4) and (A.5) emerge as simple special cases of it when X is discrete or continuous, respectively.

Definition A.14. *The expectation of a random variable X having CDF $F(x)$ is $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot dF(x)$, where we define the integral $\int_{-\infty}^{\infty} x \cdot dF(x)$ as*

$$\int_{-\infty}^{\infty} x \cdot dF(x) := \lim_{a \to -\infty, b \to \infty} \lim_{N \to \infty} \sum_{n=1}^{N} \left\{ a + n \cdot \frac{b-a}{N} \right\} \cdot \left\{ F\left(a + n \cdot \frac{b-a}{N} \right) - F\left(a + (n-1) \cdot \frac{b-a}{N} \right) \right\}$$

The quantity $\int_{-\infty}^{\infty} x \cdot dF(x)$ in an example of a *Riemann–Stieltjes integral*, in which we "integrate'with respect to the function" $F(x)$ rather than with respect to the variable x. Let's try to unpack this long expression, with the aid of the color-coding above.

First, let's fix values of a, b, N and consider the quantity appearing inside all of the limits. For given $b > a$, imagine cutting the interval $[a, b]$ into N regions of equal size, so that they each have width $\frac{b-a}{N}$. The n^{th} such region extends from the value $a + (n-1) \cdot \frac{b-a}{N}$ to the value $a + n \cdot \frac{b-a}{N}$. Note the following:

- $F\left(a + n \cdot \frac{b-a}{N}\right) - F\left(a + (n-1) \cdot \frac{b-a}{N}\right)$ yields $P(X \in \text{ region } n)$.
- $\left\{a + n \cdot \frac{b-a}{N}\right\}$ is the location of (the right end of) region n.
- $\lim_{N \to \infty}$ takes the sum to an integral, and the a, b limit covers full support of X.

Thus, we can interpret $\mathbb{E}[X]$ as an integral of the function x over the whole real line, in which each value of x is multiplied by the probability that X is very close to x, essentially $F(x + dx) - F(x)$.

Discrete case: Now let's see how Definition A.14 yields Eq. (A.4) in the special case that X is a discrete random variable. Let $x_1, x_2 \ldots$ be the support points of X. Notice that for large enough N, only one x_j can be between $a + \frac{n-1}{N}(b-a)$ and $a + \frac{n}{N}(b-a)$. Thus: $F\left(a + \frac{n}{N}(b-a)\right) - F\left(a + \frac{n-1}{N}(b-a)\right) = \pi_j$ if x_j lies in the n^{th} region. If on the other hand no x_j lies in the n^{th} region, this quantity is equal to zero. We arrive at one term for each value x_j, and $\mathbb{E}[X] = \sum_j x_j \cdot \pi_j$.

Continuous case: When X is a continuous random variable with density $f(x)$, we can recover Eq. (A.4) by noticing that for large N:

$$F\left(a + \frac{n}{N}(b-a)\right) - F\left(a + \frac{n-1}{N}(b-a)\right) \approx f\left(a + \frac{n}{N}(b-a)\right) \cdot \frac{b-a}{N}$$

Substituting in this approximation delivers the familiar formula that $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$.

Exercise: Consider a so-called *Bernoulli random variable* X that takes a value 1 with probability p and 0 with probability $1 - p$. Show that $\mathbb{E}[X] = p$.

Exercise: Consider a *uniform* $[0, 1]$ *random variable*, that is a continuous random variable with density $f(x) = x$ for all $0 \leq x \leq 1$, and $f(x) = 0$ everywhere else. Show that $\mathbb{E}[X] = 1/2$.

A key property of the expectation operator that is very useful is that it is *linear*. It's actually "linear" in a few distinct senses:

1. *Linearity with respect to functions of a single variable:* $\mathbb{E}[a + b \cdot X] = a + b \cdot \mathbb{E}[X]$
2. *Linearity over sums of random variables:* $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
3. *Linearity with respect to mixtures:* if X, Y and Z are random variables such that $F_Z(t) = p \cdot F_X(t) + (1-p) \cdot F_Y(t)$, then $\mathbb{E}[Z] = p \cdot \mathbb{E}[X] + (1-p) \cdot \mathbb{E}[Y]$.

Note that because of Property 2, we can compute the expectation value of the random variable $X + Y$ knowing only the CDFs $F_X(x)$ and $F_Y(y)$, without needing the full joint-CDF $F_{XY}(x, y)$ of X and Y. This is a very special property of the expectation, which doesn't hold for most of the things we might want to know about the random variable $X + Y$ (for example $P(X + Y \leq t)$).

Property 3. gives us a nice way to evaluate the expectation value of a random variable that is neither discrete nor continuous. Recalling decomposition (A.3) of a general mixed random variable, let $f^c(x)$ be the density of the continuous part $F_{continuous}$ and let x_j^d and π_j^d denote the support points and associated probabilities according to the discrete part $F_{discrete}$. Then:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot dF(x) = p \cdot \left\{ \sum_j x_j^d \cdot \pi_j^d \right\} + (1-p) \cdot \int_{-\infty}^{\infty} x \cdot f^c(x) \cdot dx$$

A.4.2 Application: variance

From the expectation operator, we can also define the *variance* of a random variable, which measures how "dispersed" it is. We'll see that the variance plays an important role in asymptotic theory.

Definition A.15. *The variance of X is the expected value of the random variable $(X - \mathbb{E}[X])^2$, i.e. $Var(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$.*

The variance of X can be interpreted as the average value of the squared distance between X and its expectation $\mathbb{E}[X]$. Note that $Var(X) \geq 0$ for any random variable, with $Var(X) = 0$ only when X takes one value with probability one (i.e. X is a so-called *degenerate random variable*).

Exercise: Use the linearity of the expectation operator to prove the following (very useful) alternative expression for the variance: $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Exercise: Show that for a Bernoulli random variable (defined above), the variance is equal to $p \cdot (1 - p)$.

A.5 Conditional distributions and expectation

In this section we develop a final fundamental tool that we will use to analyze random variables: the idea of *conditional* distributions and *conditional* expectations.

Main idea: *Conditioning* on an event allows us to examine a restricted probability space in which that event is true (but other things are still random). When this idea is applied to random variables, we can define *conditional distributions* that we can work with in all of normal ways.

A.5.1 Conditional probabilities

We begin with a concept that applies to all probability spaces, not just to random variables.

Definition A.16. *Given an event B such that $P(B) > 0$, the conditional probability of event A given B is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where recall that the intersection of two events $A \cap B$ can be interpreted as the event that *both* of events A and B occur. This definition is often referred to as *Bayes' rule*. You can think of Bayes' rule as a way to define a probability function using B as the whole outcome space: yielding a way to talk about the probability that $\omega \in A$, *given* that $\omega \in B$.

Extension: Given events A, B and C, we can also define the probability of A given B *and* C as $P(A|B \cap C) = P(A \cap B \cap C)/P(B \cap C)$, and so on for any number of events.

Exercise: We call events A and B *independent* if $P(A \cap B) = P(A) \cdot P(B)$. Suppose that $P(B) > 0$. Show that A and B are independent if and only if $P(A|B) = P(A)$.

A.5.2 Conditional distributions

Consider now two random variables X and Y.

Definition A.17. *The conditional CDF of Y given $X = x$ is*

$$F_{Y|X=x}(y) := P(Y \leq y|X = x) := \lim_{\epsilon \downarrow 0} P(Y \leq y|X \in [x, x + \epsilon])$$

where the conditional probability appearing in the RHS is defined by Definition A.16.

Notation: The conditional CDF will sometimes also be denoted as $F_{Y|X}(y|X)$.

We define $P(Y \leq y|X = x)$ using $X \in [x, x + \epsilon]$ as our conditioning event B, and then taking the limit, because the probability of $X = x$ may be zero, e.g. for a continuously distributed X. *Note:* The Hansen

book uses $x \in [x - \epsilon, x + \epsilon]$ instead of $x \in [x, x + \epsilon]$, but the two definitions are equivalent.

Given the general Definition A.17, we can consider each of our two typical special cases:

- When $P(X = x) > 0$ (e.g. for a discrete random variable with a support point at x), Definition A.17 reduces to the simpler expression $P(Y \leq y | X = x) = \frac{P(Y \leq y \text{ and } X = x)}{P(X = x)}$. We can interpret $F_{Y|X=x}(y)$ as the CDF among the sub-population of i for which $X = x$.
- If on the other hand $f_X(x) = \frac{d}{dx} F_X(x)$ exists (e.g., for a continuous random variable), then Definition A.17 simplifies to $P(Y \leq y | X = x) = \frac{\frac{d}{dx} P(Y \leq y, X \leq x)}{f_X(x)}$. We can interpret $F_{Y|X=x}(y)$ as the CDF among the sub-population of i for which X is "very close" to x.

Exercise: derive each of these two expressions from Definition A.17 . For the discrete case, you may find useful the "quotient rule" that $\lim_{t \to 0} \frac{g(t)}{h(t)} = \frac{\lim_{t \to 0} g(t)}{\lim_{t \to 0} h(t)}$ when both limits exist and $\lim_{t \to 0} h(t) \neq 0$. For the continuous case, try dividing both the numerator and the denominator of $P(Y \leq y | X \in [x, x + \epsilon])$ by ϵ before taking the limit.

Exercise: Show that if X and Y are independent then $F_{Y|X=x}(y) = F_Y(y)$ and $F_{X|Y=y}(x) = F_X(x)$ for all x and y. Note: it's actually an if-and-only-if, but proving the other direction is more difficult.

A.5.3 Conditional expectation (and variance)

Consider a fixed value of x, and view the conditional CDF $F_{Y|X=x}(y)$ as a function of y. This function satisfies the four properties of a CDF mentioned in Section A.3.1: it is weakly increasing, right-continuous, and ranges from zero to one.

Thus, we can define the expectation over this distribution in exactly the same way as we would for $\mathbb{E}[Y]$ based on Definition A.14, except that use $F_{Y|X=x}(y)$ as the CDF rather than it's "unconditional" analog $F(y)$. We can write this using the general notation of Definition A.14 as:

$$\mathbb{E}[Y | X = x] = \int_{-\infty}^{\infty} y \cdot dF_{Y|X=x}(y)$$

We can unpack this expression depending on what type of random variable Y is:

- If Y is continuous: $\mathbb{E}[Y | X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X=x}(y) \cdot dy$, where $f_{Y|X=x}(y) = \frac{d}{dy} F_{Y|X=x}(y)$.
- If Y is discrete: $\mathbb{E}[Y | X = x] = \sum_j y_j \cdot \pi_{j|X=x}$, where $\pi_{j|X=x} = \lim_{\epsilon \downarrow 0} \left\{ F_{Y|X=x}(y_j) - F_{Y|X=x}(y_j - \epsilon) \right\}$.

Observe that the conditional expectation $\mathbb{E}[Y | X = x]$ depends on x only, as we've averaged over various values of Y. Accordingly, we can define a function that evaluates $\mathbb{E}[Y | X = x]$ over different values of x:

Definition A.18. *The conditional expectation function (CEF) of Y given X is $m(x) := \mathbb{E}[Y | X = x]$.*

We can also use the CEF to define a new random variable, denoted $\mathbb{E}[Y | X]$.

Definition A.19. $\mathbb{E}[Y | X] = m(X)$, *where* $m(x) := \mathbb{E}[Y | X = x]$.

For example, if X is discrete, then $\mathbb{E}[Y | X]$ takes value $m(x_j) = \mathbb{E}[Y | X = x_j]$ with probability π_j.

The so-called *law of iterated expectations* shows that the expectation value of $\mathbb{E}[Y | X]$ recovers the (unconditional) expectation of Y:

Proposition (law of iterated expectations): $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y | X]\right]$

Proof. We prove it for the case in which both X and Y are continuous random variables. The other

cases are analagous so I leave them as an exercise.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[Y|X]\right] &= \int_{x\in\mathbb{R}:f_X(x)>0} f_X(x)\cdot\mathbb{E}[Y|X=x]\cdot dx \\
&= \int_{x\in\mathbb{R}:f_X(x)>0} f_X(x)\cdot\left\{\int_{y\in\mathbb{R}} y\cdot f_{Y|X}(y|x)\cdot dy\right\}\cdot dx \\
&= \int_{x\in\mathbb{R}:f_X(x)>0} \cancel{f_X(x)}\cdot\left\{\int_{y\in\mathbb{R}} y\cdot\frac{f_{XY}(x,y)}{\cancel{f_X(x)}}\cdot dy\right\}\cdot dx \\
&= \int_{y\in\mathbb{R}} y\cdot\underbrace{\left\{\int_{x\in\mathbb{R}:f_X(x)>0} f_{XY}(x,y)\cdot dx\right\}}_{=f_Y(y)\cdot dy} = \int_{y\in\mathbb{R}} y\cdot f_Y(y)\cdot dy = \mathbb{E}[Y]
\end{aligned}$$

□

The law of iterated expectations is useful because in many settings the quantity $\mathbb{E}[Y|X=x]$ is easier to work with than $[Y]$ is directly.

Example: Suppose that Y is individual i's height and X is an indicator for whether they are a child or an adult. Then the law of iterated expectations tells us that the average height in the population can be obtained by averaging together the mean height among children with the mean height among adults. Suppose that 75% of the population are adults. Then the law of iterated expectations reads as:

$$\mathbb{E}[height] = .75\cdot\mathbb{E}[height|adult] + .25\cdot\mathbb{E}[height|child]$$

Proposition (CEF minimizes mean squared prediction error): Suppose we're interested in constructing a function $g(\cdot)$ with the goal of using $g(X)$ as a prediction of Y. We can show that $m(x) := \mathbb{E}[Y|X=x]$ is the best such function, in the sense that for each value of x

$$m(x) = \text{argmin}_g\mathbb{E}[Y-g(X))^2]$$

Proof. Here I'll use the general notation so we don't need to make any assumptions about what type of random variable X is (discrete, continuous, etc.):

$$\begin{aligned}
\mathbb{E}[(Y-g(X))^2] &= \mathbb{E}\left\{\mathbb{E}[(Y-g(X))^2|X]\right\} = \int\mathbb{E}[(Y-g(X))^2|X=x]\cdot dF(x) \\
&= \int\mathbb{E}[(Y-g(x)^2|X=x]\cdot dF(x) = \int\mathbb{E}[Y^2-2Yg(x)+g(x)^2|X=x]\cdot dF(x) \\
&= \int\left\{\mathbb{E}[Y^2|X=x]-2g(x)\mathbb{E}[Y|X=x]g(x)+g(x)^2\right\}\cdot dF(x)
\end{aligned}$$

For each value of x, the quantity in brackets is minimized by $g(x)=\mathbb{E}[Y|X=x]$. To see this, note that the quantity $\mathbb{E}[Y^2|X=x]-2g\mathbb{E}[Y|X=x]g+g^2$ is a convex function of g, and the first-order condition for minimizing it is satisfied when $g=\mathbb{E}[Y|X=x]$. □

We can also define a *conditional variance* function $Var(Y|X=x)=\mathbb{E}[(Y-\mathbb{E}[Y|X=x])^2|X=x]$ from the conditional distribution $F_{Y|X=x}$. An analog to the law of iterated expectations exists for the conditional variance, which is sometimes called the *law of total variance*.

Proposition (law of total variance): $Var(Y)=\mathbb{E}[Var(Y|X)]+Var(\mathbb{E}[Y|X])$.

Example: Recall the height example from the law of iterated expectations. The law of total variance reveals that the variance of heights in the population overall is *greater* than what we would get by just averaging the variances of each subgrup. That is:

$$Var(height) > .75\cdot Var(height|adult) + .25\cdot Var(height|child)$$

The reason is that $Var(height)$ involves making comparisons directly between the heights of children and adults, which are not captured in $Var(Y|X=x)$ for either value of x. The law of total variance

tells us exactly what correction we would need to make, which is to add the second term $Var(\mathbb{E}[Y|X])$. Remarkably, the correction required just depends on the *average* height within each group $\mathbb{E}[Y|X = x]$, as well as the proportion of adults vs. children: $P(X = x)$.

A.6 Random vectors and random matrices

Main idea: *Random vectors* are vectors in which each component is a random variable, and *random matrices* are matrices where each entry is a random variable. These concepts allow us to define the expectation, variance, and covariance between random vectors, which gives us a compact notation to discuss many random variables at the same time.

A.6.1 Definition

Rather than coming up with new letters X, Y, Z for multiple random variables, sometimes a more compact notation is to think of a single "random vector" containing all three.

Definition A.20. *A random vector X is a vector in which each component is a random variable, e.g.*

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

where X_1, X_2, etc. are each random variables.

Note the following:

- A realization $\mathbf{x}$ of random vector X is a point in $\mathbb{R}^k$, i.e. $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$:

$$P(X = \mathbf{x}) = P(X_1 = x_1 \text{ and } X_2 = x_2 \ldots \text{ and } \ldots X_k = x_k)$$

- For a random vector X, the function F_X denotes the joint-CDF of the random variables $X_1, X_2, \ldots X_k$:

$$F_X(\mathbf{x}) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2 \ldots \text{ and } \ldots X_k \leq x_k)$$

- The expectation of a random vector X is simply the vector of expectations of each of its components, i.e.

$$\mathbb{E}[X] = \left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \right] := \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_k] \end{pmatrix}$$

- The law of iterated expecations $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$ still holds when X is a random vector, rather than a random variable.

.

Definition A.21. *An $n \times k$ random matrix $\mathbf{X}$ is a matrix in which each component is a random variable, e.g.*

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}$$

where X_{lm}, is a random variable for each entry lm.

Just as with a random variable, we define the expectation of a random matrix as a matrix composed of the expectation of each of it's components, i.e.

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_{11}] & \mathbb{E}[X_{12}] & \dots & \mathbb{E}[X_{1k}] \\ \mathbb{E}[X_{21}] & \mathbb{E}[X_{22}] & \dots & \mathbb{E}[X_{2k}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_{n1}] & \mathbb{E}[X_{n2}] & \dots & \mathbb{E}[X_{nk}] \end{pmatrix}$$

This allows us to generalize the notion of variance to random vectors.

Definition A.22. *The variance of a random vector X is* $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']$

where use the notation that for a vector $\mathbf{x}$: $\mathbf{x}'$ indicates its transpose $(x_1, x_2, \dots x_k)$. Note that for vectors $\mathbf{x} = (x_1 \dots x_n)'$ and $\mathbf{y} = (y_1 \dots y_k)$, $\mathbf{xy}'$ is an $n \times k$ matrix, where the lm component of $\mathbf{xy}'$ is $x_l \cdot y_m$. We will also use $'$ to denote the matrix transpose, i.e. $[X']_{lm} = X_{ml}$.

Note that when X is a random vector rather than a random variable, $Var(X)$ is often referred to as the "variance-covariance matrix" of X. We'll use the variance-covariance matrix a lot, because it plays an important role in studying parametric distributions like the multivariate normal distribution, and in asymptotic theory.

To understand the name, let us first define the *covariance* between random vectors X and Y:

Definition A.23. *The covariance of random vectors X and Y is* $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])']$

Note the following properties of covariance:

- For random vector X: $Var(X) = Cov(X, X)$
- When X and Y are scalars (i.e. single random variables), $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- For scalar X and Y, and numbers a, b: $Cov(X, a + bY) = b \cdot Cov(X, Y)$
- For a random vector X, the components of the matrix $Var(X)$ are scalar variance and covariances, hence its name:

  $$Var(X) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \dots & Var(X_k, X_k) \end{pmatrix}$$

  A consequence of this expression is that $Var(X)$ is a *symmetric* matrix: $[Var(X)]_{lm} = [Var(X)]_{ml}$, because $Cov(X_l, X_m) = Cov(X_m, X_l)$.
- When X and Y are scalars, we can define the *correlation coefficient* ρ_{XY} as $\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ (note that all quantities involved here are scalars). ρ_{XY} is always a number between -1 and $+1$ (homework problem).

Exercise: Show that $Cov(X, Y) = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y]'$

A.6.2 Conditional distributions with random vectors

A.6.2.1 Conditioning on a random vector

In Section A.5 we defined the conditional distribution of one random variable Y given another random variable X. This idea extends naturally to conditioning a random variable Y on multiple random variables at the same time, e.g. $F_{Y|X=x,Z=z}(y)$. Random vectors give us a nice notation for this:

Definition A.24. *With X a random vector, the conditional CDF of random variable Y given $X = \mathbf{x}$ is*

$$F_{Y|X}(y|\mathbf{x}) = \lim_{\substack{\epsilon_1 \downarrow 0 \\ \epsilon_2 \downarrow 0 \\ \dots \\ \epsilon_k \downarrow 0}} P(Y \leq y | X_1 \in [x_1, x_1 + \epsilon_1], X_2 \in [x_2, x_2 + \epsilon_2] \dots X_k \in [x_k, x_k + \epsilon_k])$$

where $\mathbf{x} = (x_1, x_2, \dots x_k)'$.

We can always use the above definition, even if the components of X can be a mix of continuous and discrete random variables.

For any given value of $\mathbf{x}$, $F_{Y|X=\mathbf{x}}(y) = F_{Y|X}(y|\mathbf{x})$ yields a proper CDF function for y, which means we can continue to define the conditional expectation as $\mathbb{E}[Y|X=\mathbf{x}] = \int_{\infty}^{\infty} y \cdot dF_{Y|X=\mathbf{x}}(y)$, where the meaning of this integral is as given in Definition A.14. The conditional variance of Y given $X=\mathbf{x}$ can also be defined in the typical way from the conditional distribution $F_{Y|X=\mathbf{x}}(y)$.

The law of iterated expectations caries over unchanged when X is a random vector. That is: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$, regardless of whether X has continuous or discretely distributed components, or a mix of the two. The law of total variance caries over too (see below).

Understanding and estimating the object $\mathbb{E}[Y|X=\mathbf{x}]$ from data, where X can be a vector, will be one of our main interests in this course, motivating the use of regression analysis. Take a deep breath, we made it!

A.6.2.2 The conditional distribution of a random vector

This section can be skipped for now, but later in the course we'll need to talk about joint-distribution of a random vector, conditional on the value of one or more other random variables.

When *both* X and Y are random vectors, we can talk about the conditional distribution of Y given X by defining a *conditional joint-CDF* of all the components of Y, conditional on $X=\mathbf{x}$.

Definition A.25. *With X and Y random vectors, the conditional CDF of Y given $X=\mathbf{x}$ is*

$$F_{Y|X}(\mathbf{y}|\mathbf{x}) = \lim_{\substack{\epsilon_1 \downarrow 0 \\ \epsilon_2 \downarrow 0 \\ \cdots \\ \epsilon_k \downarrow 0}} P(Y_1 \leq y_1, Y_2 \leq y_2, \ldots | X_1 \in [x_1, x_1+\epsilon_1], X_2 \in [x_2, x_2+\epsilon_2] \ldots X_k \in [x_k, x_k+\epsilon_k])$$

where $\mathbf{x} = (x_1, x_2, \ldots x_k)'$.

An important application of the concept of a conditional joint-distribution is the idea of conditional independence.

Definition A.26 (conditional independence). *We say that X and Y are independent conditional on Z, denoted $(X \perp\!\!\!\perp Y)|Z$, if for any value z of Z: $F_{XY|Z=z}(x,y) = F_{X|Z=z}(x) \cdot F_{Y|Z=z}(y)$ for all x, y.*

This definition can be understood by using Definition A.25 to define interpret $F_{XY|Z=z}(x,y)$ as the joint-CDF of a random vector composed of X and Y, conditional on the random vector Z. In this definition X and Y could be random variables or can each be random vectors themselves!

As another application of Definition A.25, the *law of total covariance* provides an analog of the law of iterated expectations for covariance (and hence, as a special case, for variance):

Proposition A.2. *For random vectors X, Y and Z: $Cov(X,Y) = \mathbb{E}[Cov(X,Y|Z)] + Cov(\mathbb{E}[X|Z], \mathbb{E}[Y|Z])$]*

Note that as a special case we have the *law of total variance*, that: $\mathbb{E}[Var(Y|X)] + Var(\mathbb{E}[Y|X])$.

Appendix B

Asymptotic theory

In Appendix A, we've developed the idea of a random vector, which has a probability distribution that can be characterized by the joint-CDF of all of its components. This allows us to then define concepts like expectation, conditional distributions, and the conditional expectation function.

In this section we use probability to study the properties of a *sample* of data. Consider for example a collection of observations regarding 5000 young working women in the 1970s-80s.[1] This chapter develops tools that let us address the following question: what are we learning about the *population* of young working women in this time period, given our sample? In doing so we move from the theory of *probability* to the theory of *statistics*, which studies what we can learn about probability distributions from data.

B.1 The idea of a random sample

The simplest (and most common) framework in which to develop statistical results is using the notion of a *independent and identically distributed* (i.i.d.) sample.

Definition B.1. *A collection of random vectors* $\{X_1, X_2, \dots X_n\}$ *are called independent and identically distributed (i.i.d.) if* $X_i \perp\!\!\!\perp X_j$ *for* $i \neq j$ *and each* X_i *has the same marginal distribution as the others.*

When a collection of random vectors are independent of one another, as with an i.i.d. collection knowing the CDF F for each member X_i of the collection is sufficient to recover the full joint-distribution of the collection. For example, let $n = 2$ and suppose both X_1 and X_2 are *i.i.d* random variables (rather than vectors). Then $P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) = F(x_1) \cdot F(x_2)$, where $F(\cdot)$ is the marginal CDF function of each of the X_i. With an i.i.d. collection, we only need to know the CDF F that applies to each of the X_i, in order to know anything about the collection.

The *i.i.d.* model is typically used to describe *simple random sampling*. Simple random sampling occurs when individuals are selected at random from some underlying population I, and a set of variables $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$ are recorded for each sampled individual i. Imagine for example a telephone survey, in which enumerators have a long list I of potential individuals to contact. They use a random number generator to choose an i at random from this list, contact them, and record responses to a set of k questions. This process is then repeated n times.

Note: With a finite population I, we must allow sampling "with replacement" for the i.i.d. model to hold strictly. If individual i is removed from the list after being contacted, then the random vectors X_i may no longer be independent. For example, suppose we are randomly selecting U.S. states and recording the population of each one. Suppose California (the post populous state) has 40 million and Georgia has 11. Then for example $P(X_2 = 40m | X_1 = 40m) \neq P(X_2 = 40m | X_1 < 40m)$, since the first probability is zero and the second is 1/49. This means that X_1 and X_2 are not independent. Simple random sampling is often referred to as *random sampling* for short, or as *i.i.d sampling*.

We'll use the terms *dataset* or *sample* to refer to an $n \times k$ matrix $\mathbf{X}$ that records characteristics $X_i = (X_{1i}, X_{2i}, \dots X_{ki})$ for each of n observational units (such as individuals) i. Data is not always generated

[1] A dataset fitting this description can be easily loaded into Stata using the command `wenuse nlswork`, which comes from the U.S. Bureau of Labor Statistics' National Longitudinal Survey.

by simple random sampling, but when it is, we can imagine $\mathbf{X}$ as being formed by randomly choosing rows from a much larger matrix that records X_i for all individuals in the population, depicted in Figure B.1. The actual data we see in $\mathbf{X}$ is a realization of the collection of random variables $\{X_1, X_2, \dots X_n\}$.

$$\mathbf{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \begin{pmatrix} (X_{11}, X_{21}, \dots X_{k1}) \\ (X_{12}, X_{22}, \dots X_{k2}) \\ \vdots \\ (X_{1n}, X_{2n}, \dots X_{kn}) \end{pmatrix}$$

The randomness of $\mathbf{X}$ comes from the random-sampling: we could have drawn a different set of individuals from the population, in which case we would have seen a different dataset $\mathbf{X}$.

Notation: Note that the entries of the sample matrix $\mathbf{X}$ are denoted X_{ji}, where i index rows (individual observations) and j index columns (variables/characteristics). This is backwards from the way we often denote entries M_{ij} of a matrix $\mathbf{M}$, where the row i comes before the column j. This is a consequence of two conventions interacting: that rows of $\mathbf{X}$ index individuals (just like when you open the dataset in `R`), but that X_{ji} indexes characteristic j of individual i (equivalently, characteristic j of the individual sampled in row i).

Sample $\mathbf{X}$

row **i**	$\omega_\mathbf{i}$	**age**$_i$	**married**$_i$	**college**$_i$
1	1	25	0	0
2	4	37	1	1
3	5	54	0	1

Population I

individual **i**	**age**$_i$	**married**$_i$	**college**$_i$
1	25	0	0
2	74	1	1
3	8	0	0
4	37	1	1
5	54	0	1

Figure B.1: An example of simple random sampling, in which $n = 3$ and $N = 5$. Each row of the dataset on the left is a realization of random vector $X = (age, married, college)$, which chooses a row at random from the population matrix on the right. We can conceptualize this sampling process as a probability space with outcomes $\omega = (\omega_1, \omega_2, \omega_3)$, where ω_i yields the index of the randomly selected individual in I. The random vectors $X_i = X_i(\omega_i)$ and $X_j = X_j(\omega_j)$ are independent for $i \neq j$, but the random variables within a row are generally not independent, e.g. age_i and $college_i$ are positively correlated.

Note that most sampling processes in the real world occur without replacement: the same individual cannot show up twice in the data. Given the note above, this suggests that these sampling processes are not *i.i.d.*, strictly speaking. However, when the size N of the underlying population is large, such samples can still be well-approximated as being *i.i.d.*. Intuitively, that's because when N is much larger than n (often denoted as $N >> n$), the chance that you would draw the same individual twice is very low. We thus typically assume *i.i.d.*, with the idea that N is suitably large to not worry about sampling with vs. without replacement.

The following are some alternative methods of generating data, aside from simple random sampling:

- *Stratified random sampling*: the population is divided into groups, and then simple random sampling occurs within each group (e.g. I run my sampling algorithm separately for men and women, so that I can ensure equal representation of each).
- *Clustered random sampling*: after defining groups, we randomly select some of the groups. Then all individuals from those groups are included in the sample (e.g. I interview everybody in a household, after choosing households at random)
- *Panel data*: suppose we have observations over multiple time-periods t for each individual i, where the individuals i are drawn as a simple random sample. Then if we arrange all of i's data onto one row, we can imagine $\mathbf{X}$ as reflecting an *i.i.d.* sample. But with rows correponding to (i, t) pairs, the rows are no longer independent (in general)

- *Observing the whole popluation*: this would be the case e.g. with state-level data from all 50 U.S. states. This situation occurs increasingly frequently with individual-level data now as well, e.g. administrative data on all tax-filers in a country.

These alternative sampling methods tend to violate the *i.i.d* assumption. However, methods exist to deal with each of them.

Let us end this section with a last bit of jargon. When X_i for $i = 1 \dots n$ denotes a collection of *i.i.d* random vectors, we'll refer to the distribution F that describes the marginal distribution of each X_i as the *population distribution*. The population distribution is the distribution we get when we randomly select any individual from the population. Features of the population distribution are the ones that you naturally think of when you think about summarizing a population. For example, if I is a finite population, then

$$\mathbb{E}_F[X_i] = \frac{1}{N} \sum_{i \in I} X_i$$

where we use the notation $\mathbb{E}_F$ to make explicit that the expectation is with respect to the CDF F. Normally, we won't write F explicitly. The population mean is simply the mean of X_i among everybody in I.

Another piece of terminology will be useful as we discuss samples and their population counterparts:

Definition B.2. *A **statistic** or **estimator** is any function of the sample* $\mathbf{X} = (X_1', X_2', \dots, X_n')'$.

A generic estimator or statistic will apply some function $g(\mathbf{X}) = g(X_1, X_2, \dots X_n)$ to the collection of random vectors that constitute the sample. An example is the so-called *sample mean* $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$, which simply adds together X_i for across the sample and divides by the number of observations n. $\bar{X}_n$ is an example of a statistic. Since each of the X_i is a random variable/vector, it follows that $\bar{X}_n$ is itself a random variable/vector. This is true of statistics in general: they are random.

The reason that we also refer to statistics as "estimators" is that statistics often attempt to estimate a population quantity of some kind from data. For example, we'll see in the next Chapter that for large n, we are justified in thinking that $\bar{X}_n \approx \mu$. It is therefore reasonable to use $\bar{X}_n$ as an estimate of μ. Note that $\bar{X}_n$ is random, while μ is just a fixed number. Thus we have to be careful in what we mean by saying that $\bar{X}_n \approx \mu$, which is the topic of the next chapter.

Notation: Often estimators are depicted with a "hat" on them, e.g. $\hat{\theta} = g(\mathbf{X})$. We'll use this notation to denote a generic estimator.

A useful property of *i.i.d.* random vectors that I'll mention here is the following:

Proposition B.1. *If* $\{X_1, X_2, \dots X_n\}$ *are i.i.d random vectors, then* $\{h(X_1), h(X_2), \dots h(X_n)\}$ *are also i.i.d for any (measurable) function* h.

An implication of Proposition B.1 is that if we have an *i.i.d.* sample X_i, we can from it construct an *i.i.d.* sample of e.g. X_i^2.

B.2 The law of large numbers

Consider an *i.i.d.* sample $\{X_1, \dots X_n\}$ of some random variable X_i. The *sample average* of X_i in our data simply takes the arithmetic mean across these n observations:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$$

The law of large numbers (LLN) states the deep and useful fact that for very large n, it becomes very unlikely that $\bar{X}_n$ is very far from $\mu = \mathbf{E}[X_i]$, the "population mean" of X_i.

Theorem 4 (law of large numbers). *If* X_i *are i.i.d random variables and* $\mathbb{E}[X_i]$ *is finite, then for any* $\epsilon > 0$*:*

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Note: The LLN is stated above for a random variable, but the result generalizes easily to random vectors. In that case, $\lim_{n\to\infty} P(||\bar{\mathbf{X}}_n - \mu||_2 > \epsilon) = 0$ where $||\cdot||_2$ denotes the Euclidean norm, i.e.: $|\bar{\mathbf{X}}_n - \mu| = (|\bar{\mathbf{X}}_n - \mu|)'(|\bar{\mathbf{X}}_n - \mu|)$, where $\bar{\mathbf{X}}_n$ is a vector of sample means for each component of X_i, and similarly for μ.

Note: the version of the law of large numbers above is called the *weak* law of large numbers. There exists another version called the strong LLN.

Let us now prove the LLN. We will do so using a tool called *Chebyshev's inequality*. This proof assumes that $Var(X_i)$ is finite, but the LLN holds even if $Var(X_i) = \infty$. Chebyshev's inequality allows us to use the variance of a random variable to put an upper bound on the probability that the random variable is far from its mean. In particular, for any random variable Z with finite mean and variance:

$$P(|Z - \mathbb{E}[Z]| \geq \epsilon) \leq \frac{Var(Z)}{\epsilon^2}$$

To see that this holds, use the law of iterated expectations to write out the variance as

$$\begin{aligned} Var(Z) = \mathbb{E}\left[Z - \mathbb{E}[Z])^2\right] &= P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 \geq \epsilon^2\right] \\ &\quad + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot \mathbb{E}\left[(Z - \mathbb{E}[Z])^2 | (Z - \mathbb{E}[Z])^2 < \epsilon^2\right] \\ &\geq P(|Z - \mathbb{E}[Z]| \geq \epsilon) \cdot \epsilon^2 + P(|Z - \mathbb{E}[Z]| < \epsilon) \cdot 0, \end{aligned}$$

noting that $|Z - \mathbb{E}[Z]| \geq \epsilon$ iff $(Z - \mathbb{E}[Z])^2 \geq \epsilon^2$.

Now, we will show that as $n \to \infty$, $Var(\bar{X}_n) \to 0$. This along with Chebyshev's inequality implies the LLN, by letting $Z = \bar{X}_n$.
To see that $Var(\bar{X}_n) \overset{n}{\to} 0$, note first that

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

The first equality is simply the definition of $\bar{X}_n$, while the second uses linearity of the expectation operator. Now consider

$$\begin{aligned} Var(\bar{X}_n) = \mathbb{E}\left[(\bar{X}_n - \mathbb{E}[\bar{X}_n])^2\right] = \mathbb{E}\left[(\bar{X}_n - \mu)^2\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] \\ = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu)(X_j - \mu)\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(X_i - \mu)(X_j - \mu)\right] \\ = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)\right]\left[(X_i - \mu)\right] &= \frac{1}{n^2} \cdot nVar(X_i) = \frac{Var(X_i)}{n} \end{aligned}$$

where the first equality in the third line follows because when $i \neq j$, $X_i \perp\!\!\!\perp X_j$ implies that $\mathbb{E}[(X_i - \mu)] \cdot \mathbb{E}[(X_j - \mu)] = 0 \cdot 0$. Thus, the only terms that remain are when $j = i$.

Another way to see that $Var(\bar{X}_n) = \frac{Var(X_i)}{n}$ is to notice that when Y and Z are independent, $Var(Y + Z) = Var(Y) + Var(Z)$. Thus:

$$Var\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n+\right) = n \cdot Var\left(\frac{1}{n}X_i\right) = n \cdot \frac{1}{n^2} \cdot Var(X_i) = \frac{Var(X_i)}{n}$$

B.3 Asymptotic sequences

The law of large numbers provides a way to justify the claim that when n is large, $\bar{X}_n$ will be close to μ with high probability. The approximation $\bar{X}_n \approx \mu$ lies at the heart of our claims to be learning about an underlying population when we have a large sample.

In the next section, we'll see that there is more than one way to develop a large-n approximation to the distribution of a random variable. To talk about such approximations, it is useful to introduce the idea of a *sequence* of random variables Z_n, where $n = 1, 2, \ldots \infty$. For example, we can consider the sample mean $\bar{X}_n$—which is a random variable for any given n—across various possible sample sizes n.

B.3.1 The general problem

The primary motivation for considering such *asymptotic sequences* of random variables Z_n is when Z_n represents a statistic $\hat{\theta}$—something that depends upon my data (see Definition B.2). Since $\hat{\theta}$ is random (it depends on the sample that I drew), I'd like to know something about its distribution. For example, how likely is it that my sample mean is far from the population mean?

Definition B.3. *The **sampling distribution** of an statistic $\hat{\theta}$ is its CDF: $F_{\hat{\theta}}(t) = P(\hat{\theta} \leq t)$.*

When our statistic is computed as $\hat{\theta} = g(X_1, X_2, \dots X_n)$ from an *i.i.d* sample of X_i, $F_{\hat{\theta}}$ depends upon three things: the function g, the population distribution of X_i, and the sample size n.

Knowing the sampling distribution of a statistic is typically a hard problem. We know g and n, but in a research setting we don't generally know the CDF F that describes the underlying population. However, if we view $\hat{\theta}$ as a point along a sequence of random variables Z_n, it is often possible to say something about the limiting behavior of F_{Z_n} as $n \to \infty$. *Asymptotic theory* is a set of tools for describing this limiting behavior. The law of large numbers is one such tool. If we believe that are actual sample size n is large enough that $F_{Z_n} \approx F_{Z_\infty}$, then tools like the LLN can be extremely useful. For the sample mean for example, we might, on the basis of the LLN, be prepared to believe that $\bar{X}_n$ is close to μ with very high probability.

Conceptually, we can think of what we're doing as follows. Suppose our sample size is $n = 10,576$, and we calculate a statistic $\hat{\theta} = g(X_1, X_2, \dots X_{10,576})$ from our sample. Now imagine applying the same function g to various samples of size $1, 2, \dots$ and so on, and defining a sequence Z_1, Z_2, Z_n of the corresponding values. Each Z along this sequence is itself a random variable: let $F_{Z_1}, F_{Z_2}, \dots$ be their corresponding CDFs. Our statistic $\hat{\theta}$ can be seen as a specific point along this sequence: $\hat{\theta} = Z_{10,576}$ (circled in red in Figure B.2). Since we don't know $F_{Z_{10,576}}$, but we can say something about F_{Z_∞}, we use the latter as an approximation for the former. Figure B.2 depicts this logic.
Of course, the above technique only works if we can say something definite about F_∞. The law of large numbers says that we can when our statistic is the sample mean. In Section B.5, we'll see that the *central limit theorem* provides even more information about the limiting distribution of the sample mean: that it will become approximately normal, regardless of F.

> *Note:* The logic of Figure B.2 is the "classical" approach to approximating the sampling distribution of $\hat{\theta}$, but it is certainly not the only one. An increasingly popular alternative involves *bootstrap* methods. These methods still appeal to n being "large enough", but they do so in a different way. They also require computing power, because bootstrapping involves resampling new datasets from our original dataset $\mathbf{X}$. This has become increasingly feasible, and boostrap-based methods have become increasingly popular.

B.3.2 Example: LLN and the sample mean

Let's go through the logic of Figure B.2 in more detail in the case of the the law of large numbers. The LLN tells us that when we let the sample mean $\bar{X}_n$ define our asymptotic sequence Z_n, the resulting distributions F_{Z_n} eventually cluster all of their probability mass around the point μ, the sample mean. Figure B.3 illustrates this point, through a simulation in `R`. I drew 1,000 *i.i.d* samples of size n of a random variable X_i for which $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$, representing a coin flip. Then, I plot a histogram of $\bar{X}_n$ across the 1,000 samples. This process is repeated for $n = 2$, $n = 10$, $n = 100$ and $n = 1,000$. You can think of this as illustrating Figure B.2 for the specific population distribution F that describes a coin-flip. With $n = 2$, we see that we have a 50% chance of getting $\bar{X}_n$ of 0.5, which is the true "population mean" of X_i: $\mu = \mathbb{E}[X_i] = 0.5$. Then 25% of the time we get $\bar{X}_n = 0$ (two flips of tails), and 25% of the time we get $\bar{X}_n = 1$ (two flips of heads). Thus, the distribution of $\bar{X}_n$ is not very well concentrated around $\mu = 0.5$.

The red vertical lines in Figure B.3 illustrate the law of large numbers in action. They mark the points 0.45 and 0.55, which represent a $\epsilon = .05$ in Theorem 4. We can see that by the time $n = 100$, $P(|\bar{X}_n - 1/2| > 0.05)$ starts to become reasonably small; roughly 1/3 of the mass of $\bar{X}_n$ is outside of $[0.45, 0.55]$. When $n = 1000$, there is an imperceptible chance of obtaining an $\bar{X}_n$ outside of the vertical red lines. If we continued this process for larger and larger n, we would see the mass of $\bar{X}_n$ continue to cluster closer and closer to $\mu = 1/2$. Regardless of how small a ϵ we choose, we can always find an n that fits as much of the mass as we want inside the corresponding red lines.

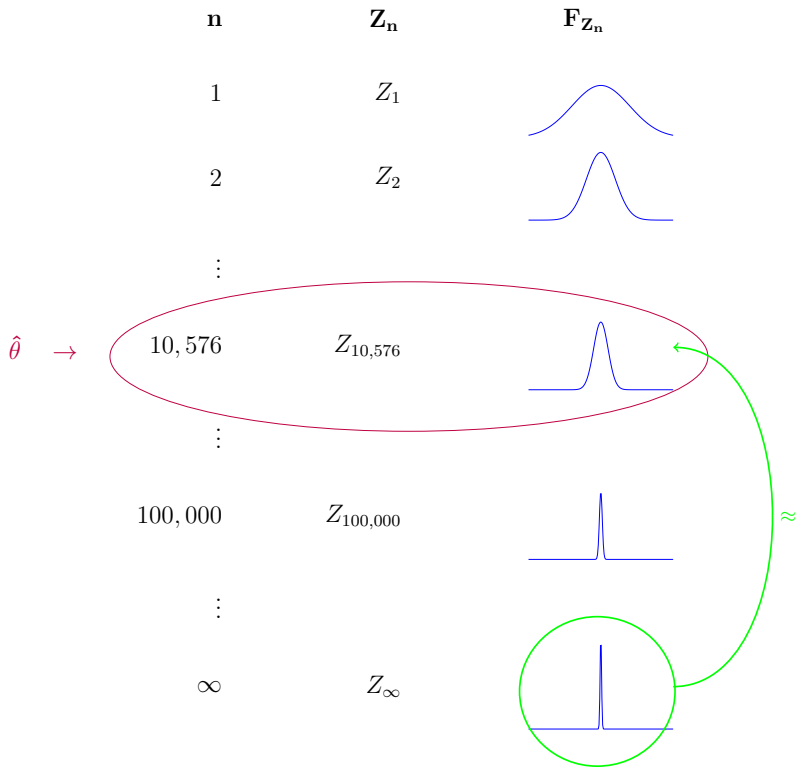


Figure B.2: We are interested in the sampling distribution of some statistic $\hat{\theta}$, computed on our sample of $10,576$ observations. This is in general hard to compute. As a tool, we imagine a sequence of random variables $Z_1, Z_2, \ldots$ in which $\hat{\theta} = Z_{10,576}$. Asymptotic theory allows us to derive properties of F_{Z_∞}, the limiting distribution of Z_n as $n \to \infty$ (circled in green). Then we use F_{Z_∞} as an approximation to $F_{Z_{10,576}}$, which we justify by n being "large". The above figure depicts a situation in which $Z_n = \bar{X}_n$, so that the distribution of Z_n narrows to a point as $n \to \infty$ (by the LLN).

Note that the law of large numbers does *not* say that $P(|\bar{X}_n - \mu| > \epsilon)$ will necessarily monotonically decrease with n, for each n. For example, we can see that for $\epsilon = .05$, we have that $P(|\bar{X}_1 - \mu| > \epsilon)$ is 0.5 and $P(|\bar{X}_2 - \mu| > \epsilon)$ is about 0.25. All that the LLN says is that $P(|\bar{X}_1 - \mu| > \epsilon)$ will get (arbitrarily) small with n, for any value of ϵ.
The following is the R code I used to generate this figure, if you'd like to copy-paste it and experiment:

```
numsims<-1000
par(mfrow=c(2,2), main="Title")
for (n in c(2,10,100,1000)){
        results<-data.frame(simulation_num=integer(), sample_mean=
            double())
        for (x in 1:numsims) {
                thissample<-sample(c(0, 1), size = n, replace=TRUE
                    )
                samplemean<-mean(thissample)
                results[x,] = c(x,samplemean)
        }

        h<-hist(results$sample_mean, plot=FALSE, breaks = seq(from
            =0, to=1, by=.01))
        h$density = h$density/100
        plot(h, freq=FALSE, main=paste0("Distribution of sample
            means, n=",n," coin flips"), xlab="Sample mean", ylab=
            "Proportion of samples", col="green")
        abline(v=c(.45,.55), col=c("red", "red"))
}
```

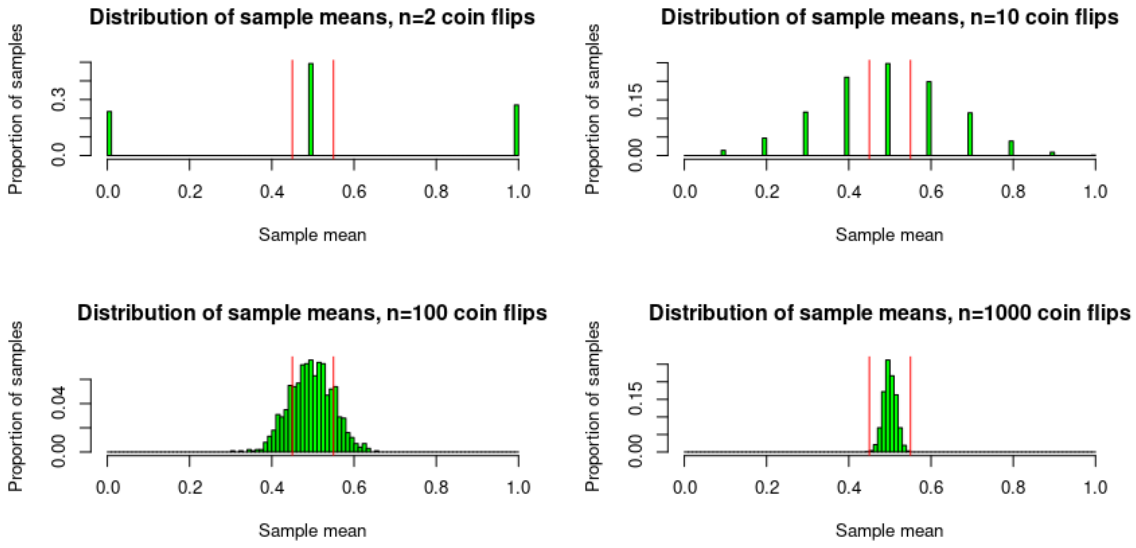


Figure B.3: Distributions along the sequence $\bar{X}_n$ for a set of n i.i.d. coin flips. Red liness illustrate the mass of the distribution $\bar{X}_n$ that is more than .05 away from 1/2.

B.4 Convergence in probability and convergence in distribution

Given a sequence of random variables or random vectors $Z_1, Z_2, \ldots$, let us now define two notions of convergence of the sequence Z_n. The first is *convergence in probability*:

Definition B.4. *We say that Z_n converges in probability to Z if for any $\epsilon > 0$:*

$$\lim_{n\to\infty} P(||Z_n - Z|| > \epsilon) = 0$$

In this definition, Z_n can be a random variable/vector. When Z_n is a random variable, then the notation $||Z_n - Z||$ jut refers to the absolute value of the difference: $|Z_n - Z|$. When Z_n is a vector, we can take $||Z_n - Z||$ to be the Euclidean norm of the difference (see Proposition B.3 for an example).

We will often talk about Z_n converging in probability to a *constant* c. This does not require a second definition because a constant is simply an example of a random variable that has degenerate distribution $P(Z = c) = 1$. Thus we say that Z_n converges in probability to a constant c if $\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$ for all $\epsilon > 0$.

Notation: When Z_n converges in probability to Z, we write this as $Z_n \xrightarrow{p} Z$, or alternatively $plim(Z_n) = Z$. We say that Z is the *probability limit* of the sequence Z_n. We use the same notation when Z is a constant.

The law of large numbers, for example, says that $\bar{X}_n \xrightarrow{n} \mu$, the sample mean converges in probability to the "population mean", or expectation, of X_i.

Exercise: This problem gives an example of a sequence that converges in probability to another random variable, rather than to a constant. Let $Z_n = Z + \bar{X}_n$, where Z is a random variable and $\bar{X}_n$ is the sample mean of i.i.d. random variables X_i having zero mean and finite variance. Suppose furthermore that Z and $\bar{X}_n$ are independent. Show that $plim(Z_n) = Z$.

Our second notion of convergence of a sequence of random vectors is *convergence in distribution*. Consider first a sequence of scalar random variables:

Definition B.5. *We say that a random variable Z_n converges in distribution to Z if, for any z such that the CDF $F_Z(z) = P(Z \leq z)$ of Z is continuous at z:*

$$\lim_{n\to\infty} P(Z_n \leq z) = F_Z(z)$$

Notation: When Z_n converges in distribution to Z, we write this as $Z_n \xrightarrow{d} Z$. As with convergence in probability, Z can be a random vector or a constant.

Note: The requirement that we only consider z where $F_Z(z)$ is continuous is a technical condition, which we can often ignore because we'll be thinking about continuously distributed Z. In general, we can construct examples in which $\lim_{n\to\infty} P(Z_n \leq z)$ is not right-continuous (and is thus not a valid CDF), but the valid CDF function $F_Z(z)$ nevertheless captures the limiting distribution of Z_n. In these cases we still want to say that $Z_n \xrightarrow{d} Z$.

The definition given above for convergence in distribution takes Z_n to be a random (scalar) variable to emphasize the idea, but the concept extends naturally to sequences of random vectors. We say that a sequence of random vectors Z_n converges in distribution to Z if for all $\mathbf{z}$ at which the joint CDF of the components of Z $F_Z(\mathbf{z})$ does not have a discontinuity, the limit of the CDF of Z_n evaluated at that point as $n \to \infty$ is $F_Z(\mathbf{z})$.

Convergence in distribution essentially says that the CDF of Z_n point-wise converges to the CDF of Z. By "point-wise", we mean that this occurs for each value z. When $Z_n \xrightarrow{d} Z$, we often refer to Z as the "large-sample" or "asymptotic" distribution of Z_n.

We close this section by investigating the relationship between convergence in probability and convergence in distribution. Convergence in distribution is a weaker notion of convergence (and is in fact often called "weak" convergence), in the sense that it is implied by convergence in probability.

Proposition B.2. *If $Z_n \xrightarrow{p} Z$, then $Z_n \xrightarrow{d} Z$. In the special case that Z is a degenerate random variable taking value of c, then $Z_n \xrightarrow{d} c$ also implies $Z_n \xrightarrow{p} c$. Thus when Z is degenerate, convergence in distribution and probability are equivalent to one another.*

One manifestation of the fact that convergence in probability is stronger than convergence in distribution is that with the former, covergence of elements of a random vector implies convergence of the whole random vector:

Proposition B.3. *If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{p} \begin{pmatrix} X \\ Y \end{pmatrix}$.*

Proof. Since for any $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$ and $\lim_{n\to\infty} P(|Y_n - Y| > \epsilon) = 0$ holds for any $\epsilon > 0$, let's consider a value $\epsilon/\sqrt{2}$. Let $Z_n := (X_n, Y_n)'$ and $Z := (X, Y)'$. Since $||Z_n - Z|| = \sqrt{(X_n - X)^2 + (Y_n - Y)^2}$ being larger than ϵ is the same as $(Z_n - Z)^2$ being larger than ϵ^2, and since at least one of $(X_n - X)^2$ or $(Y_n - Y)^2$ must then be larger than half of ϵ^2, we have:

$$\begin{aligned} P(||Z_n - Z|| > \epsilon) &= P((X_n - X)^2 + (Y_n - Y)^2 > \epsilon^2) \leq P((X_n - X)^2 > \epsilon^2/2 \text{ or } (Y_n - Y)^2 > \epsilon^2/2) \\ &\leq P((X_n - X)^2 > \epsilon^2/2) + P((Y_n - Y)^2 > \epsilon^2/2) \end{aligned}$$

we have that

$$\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = \lim_{n\to\infty} P(|X_n - X| > \epsilon/\sqrt{2}) + \lim_{n\to\infty} P(|Y_n - Y| > \epsilon/\sqrt{2}) = 0 + 0 = 0$$

□

Meanwhile, the same is not true of convergence in distribution: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does not in general imply that $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$. However, one important special case in which it does is when X or Y is a degenerate random variable. This is useful for example in proving Slutsky's Theorem in Section B.6.

The next section will introduce the most famous and useful instance of convergence in distribution: the *central limit theorem* (CLT). After introducing the CLT, we will return in Section B.6 to some further properties of convergence in probability and convergence in distribution, that will be useful in the analysis of large samples.

Optional: There is an even stronger notion of convergence than convergence in probability, referred to as *almost-sure convergence*. We say that Z_n converges almost surely to Z, or, $Z_n \overset{a.s.}{\to} Z$, if

$$P\left(\lim_{n\to\infty} Z_n = Z\right) = 1$$

To make sense of this expression we have to place a probability distribution over entire sequences $\{Z_n\}$ (something we didn't need to do for convergence in probability or convergence in distribution). That is, we imagine a probability space in which each outcome ω yields to a realization of all of the random variables: Z, Z_1, Z_2, Z_3, and so on. Then, the above expression says that $P\left(\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}\right) = 1$. In words: the probability of getting a sequence of Z_n that does not converge to Z with n is zero.

Almost sure convergence is stronger than convergence in probability, i.e. $Z_n \overset{a.s.}{\to} Z$ implies that $Z_n \overset{p}{\to} Z$ (which of course in turn implies that $Z_n \overset{d}{\to} Z$). The *strong law of large numbers* states that the sample mean in fact converges almost surely to the population mean, that is $\bar{X}_n \overset{a.s.}{\to} \mu$.

B.5 The central limit theorem

The central limit theorem (CLT) tells us that if we construct from the sample mean $\bar{X}_n$ the a random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)$, then the sequence Z_n converges in distribution to that of a normal random variable.

Theorem 5 (central limit theorem). *If X_i are i.i.d random vectors and $\mathbb{E}[X_i' X_i] < \infty$, then*

$$\sqrt{n}(\bar{X}_n - \mu) \overset{d}{\to} N(\mathbf{0}, \Sigma)$$

where $\Sigma = Var(X_i)$, $\mu = \mathbb{E}[X_i]$, and $\mathbf{0}$ is a vector of zeros for each component of X_i.

The central limit theorem is quite remarkable. It says that *whatever* the distribution of X_i is, the limiting distribution of $\bar{X}_n$ (recentered by μ and rescaled by $\sqrt{n}$) will be a normal distribution. This striking result will pave the way for us to perform inference on the expectation of a random variable, without knowing its full distribution.

Why the CLT is useful:
The practical value of the CLT is that it delivers an approximation to the distribution of $\bar{X}_n$. For large n, we know that $\sqrt{n}(\bar{X}_n - \mu)$ has approximately the distribution $N(0, \Sigma)$. Using properties of the normal distribution, we can re-arrange this to say that $\bar{X}_n \sim N(\mu, \Sigma/n)$, approximately. To get a good guess of the distribution of $\bar{X}_n$, we only need to have estimates of μ and Σ, which is much easier than estimating the full CDF of X_i from data.

Example: Suppose for simplicity that we had reason to believe that $\Sigma = 1$, i.e. we have a random variable X_i with a variance of 1. However, we don't know μ. We do know then, by the CLT, that for large n, $\bar{X}_n$ is approximately normally distributed around μ with a variance of $1/n$. This is extremely useful, because we can now evaluate candidate values of μ, based on how unlikely we would be to see a value of $\bar{X}_n$ like the one that we calculate, if that value of μ was true. Suppose for example that $n = 100$, and in our sample we observed that $\bar{X}_n = 0.31$. You want to evaluate the possibility that $\mu = 0$. Well, if this were the true value of μ, then given the asymptotic approximation that $\bar{X}_n \sim N(0, 1/n)$ (or equivalently, that $10 \cdot \bar{X}_n \sim N(0, 1)$), we'd only expect to see a value of $\bar{X}_n$ as large as 0.3 once in about 1000 samples. We might thus be willing to rule $\mu = 0$ out as a possibility. This is an example of a *hypothesis test*, which will be covered in Section C.4.1.

Illustrating the CLT:
Figures B.3 and B.4 illustrate the CLT in action. Recall that in this example X_i has a two-point distribution $P(X_i = 0) = 1/2$ and $P(X_i = 1) = 1/2$). The distribution of $\bar{X}_n$ becomes closer and closer to a normal distribution centered around $\mu = 1/2$ as n gets large. To the eye, the distribution of $\bar{X}_n$ definately does not look normal for $n = 2$ or for $n = 10$ in Figure B.3. But by the time we have $n = 100$, it starts to take on the bell-curve shape. We see the variance Σ/n falling as we compare $n = 100$ and

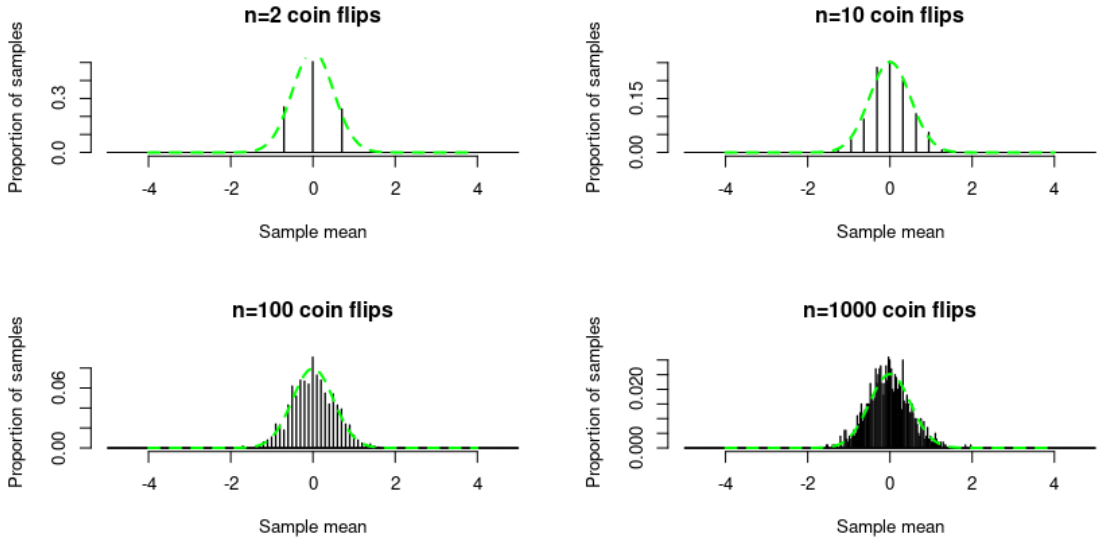


Figure B.4: The same simulation as in Figure B.3, except now we plot the distribution of $\sqrt{n}(\bar{X}_n - 1/2)$ rather than of $\bar{X}_n$. The CLT tells us that $\sqrt{n}(\bar{X}_n - 1/2) \xrightarrow{d} N(0, 1/4)$, since $1/4$ is the variance of X_i. Green dashed lines depict what is predicted by the distribution $N(0, 1/4)$, which we can see becomes close to what we see for larger values of n.

$n = 1000$: the latter has a variance about $1/10$ as large. In Figure B.4, we plot the distribution of $\sqrt{n}(\bar{X}_n - 1/2)$ overlaid with its limiting distribution.

Thought experiments like this simulation experiment are useful for getting intuition about the CLT. Accorindly, you often hear descriptions of the CLT along the lines of: "the sample mean becomes normal as the sample gets bigger and bigger". This isn't wrong, but can be a little misleading. A given real-world sample never gets bigger: it always has a single finite size n! Similarly, the sample size n never "goes to infinity"–though we can get pretty close by simulating a sequence of samples on a computer! Imagining an infinite sequence of samples having means $\bar{X}_1$, $\bar{X}_2$, and so on, is just a useful abstraction.

The following proof of the CLT is not necessary for you to know, but you may find it interesting, and being able to follow it is a good study device.

Proof of the CLT:

We'll consider a proof for the univariate case, which can be extended to random vectors using the Cramér-Wold theorem introduced in Section B.6. The proof here will use the concept of a *moment generating function*:

$$M_X(t) := \mathbb{E}[e^{t \cdot X_i}] = 1 + t \cdot \mathbb{E}[X_i] + \frac{t^2}{2} \cdot \mathbb{E}[X_i^2] + \frac{t^3}{3!} \cdot \mathbb{E}[X_i^3] + \dots \tag{B.1}$$

where the second equality uses the Taylor expansion of e^{tx}. This will be a useful expression for the moment generating function $M_X(t)$. Note that $M_X(t)$ is a (non-random) function of t: the randomness in X_i has been averaged out.

A useful result (that we will not prove here) is that if two random variables X and Y have the same moment generating function $M_X(t) = M_Y(t)$ for all t, then they have the same distribution. Our goal will be to show that whatever the distribution of X_i, the moment generating function of $\sqrt{n}(\bar{X}_n - \mu)$ converges to that of a normal random variable with variance $\sigma^2 = Var(X_i)$.

Let us divide out the variance to rewrite the CLT (in the univariate case) as $\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0, 1)$.

The moment generating function of the standard normal distribution is:

$$M_Z(t) := \frac{1}{\sqrt{2\pi}} \int dx \cdot e^{tx} \cdot e^{-\frac{x^2}{2}} = e^{-\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int dx \cdot \cdot e^{-\frac{(x-t)^2}{2}} = e^{-\frac{t^2}{2}}$$

where we've used that $(x-t)^2 = x^2 - 2tx + t^2$ and that the final integral is over the density of a normal random variable with mean t and variance 1.

Now for the magic part. We'll show that whatever the distribution of X_i is, and hence whatever the moment generating function of X_i, the moment generating function of

$$\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}} \frac{X_1 - \mu}{\sigma} + \frac{1}{\sqrt{n}} \frac{X_2 - \mu}{\sigma} + \cdots + \frac{1}{\sqrt{n}} \frac{X_n - \mu}{\sigma}$$

will end up being $e^{-\frac{t^2}{2}}$!

First, note that when Y and Z are independent of one another, the moment generating function of $Y + Z$ is equal to the product of each of their moment generating functions, i.e. $\mathbb{E}[e^{t(X_i+Z_i)}] = \mathbb{E}[e^{tY_i} \cdot e^{tZ_i}] = \mathbb{E}[e^{tY_i}\mathbb{E}[e^{tZ_i}]$. Applying this to the above expression, we have that:

$$M_{\sqrt{n} \cdot \frac{\bar{X}_n - \mu}{\sigma}}(t) = M_{\frac{1}{\sqrt{n}} \frac{X-\mu}{\sigma}}(t) \cdot M_{\frac{1}{\sqrt{n}} \frac{X-\mu}{\sigma}}(t) \cdot \cdots \cdot M_{\frac{1}{\sqrt{n}} \frac{X-\mu}{\sigma}}(t) = \left(\frac{1}{\sqrt{n}} M_{\frac{X-\mu}{\sigma}}(t) \right)^n$$

Note that for any random variable Y, $M_{\frac{1}{\sqrt{n}} \cdot Y}(t) = M_Y(t/\sqrt{n})$. Therefore, we wish to show that

$$\lim_{n \to \infty} \left(M_{\frac{X-\mu}{\sigma}}(t/\sqrt{n}) \right)^n = e^{-\frac{t^2}{2}}$$

for any t.

Applying the Taylor series expansion of the moment generating function in Equation B.1, we have that:

$$M_{\frac{X-\mu}{\sigma}}(t/\sqrt{n}) = 1 + \frac{t}{\sqrt{n}} \cdot \mathbb{E}\left[\frac{X_i - \mu}{\sigma} \right] + \frac{t^2}{2n} \cdot \mathbb{E}\left[\left(\frac{X_i - \mu}{\sigma} \right)^2 \right] + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right)$$

where by the Taylor theorem $\lim_{n \to \infty} g\left(\frac{t}{\sqrt{n}} \right) = 0$. Note that $\mathbb{E}\left[\frac{X_i - \mu}{\sigma} \right] = 0$ and $\mathbb{E}\left[\left(\frac{X_i - \mu}{\sigma} \right)^2 \right] = 1$, and thus we wish to show that

$$\lim_{n \to \infty} \left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right) \right)^n = e^{-\frac{t^2}{2}}$$

Recall the identity that $\lim_{n \to \infty} (1 + x/n)^n = e^x$. If we can ignore the g term then we are done. To show that the g term indeed does not contribute in the limit, consider taking the natural logarithm of both sides of the above equation (since the log is continuous function, it preserves limits):

$$\begin{aligned}
\lim_{n \to \infty} \ln \left\{ \left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right) \right)^n \right\} &= \lim_{n \to \infty} n \cdot \ln \left(1 + \frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right) \right) \\
&= \lim_{n \to \infty} n \cdot \left(\frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right) \right) = -\frac{t^2}{2} + t^2 \cdot \lim_{n \to \infty} \cdot g\left(\frac{t}{\sqrt{n}} \right) \\
&= -\frac{t^2}{2}
\end{aligned}$$

where we've used the Taylor theorem for the natural logarithm: $\ln(1 + z) = z + z \cdot h(z)$ where $\lim_{z \to 0} h(z) = 0$, and we have that $\lim_{n \to 0} \left(\frac{t^2}{2n} + \frac{t^2}{n} \cdot g\left(\frac{t}{\sqrt{n}} \right) \right) = 0$.

B.6 Properties of convergence of random variables

This section presents several results that are useful in the analysis of large samples. We will make heavy use of them, for example, when we study the asymptotic properties of the linear regression estimator.

B.6.1 The continuous mapping theorem

The *continuous mapping theorem* (CMT) states that the notions of convergence in probability and convergence in distribution are preserved when we apply a continuous function to each random vector in a sequence Z_n, that is:

Theorem 6 (continuous mapping theorem). *Consider a sequence Z_n of random vectors and a continuous function h. Then:*

- *if $Z_n \xrightarrow{p} Z$, then $h(Z_n) \xrightarrow{p} h(Z)$*
- *if $Z_n \xrightarrow{d} Z$, then $h(Z_n) \xrightarrow{d} h(Z)$*

Example: By the large of large numbers and the CMT: $\left(\bar{X}_n + 5\right) \xrightarrow{p} (\mu + 5)$, where $\mu = \mathbb{E}[X_i]$.

Example: Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$. Then by the CLT and CMT: $Z_n^2 = n\left(\bar{X}_n - \mu\right)^2 \xrightarrow{d} \chi_1^2$, where χ_1^2 is the chi-squared distribution with one degree of freedom (this is the distribution of a standard normal $N(0,1)$ random variable squared).

Note: The assumption that h is (globally) continuous can be weakened, which is often important in applications.

- When Z is a contant (call it c), then the convergence in probability part of the CMT only requires that $h(z)$ be continuous at c, rather than everywhere.
- The convergence in distribution part of the CMT can be extended to cases in which h has a set of points $z \in \mathcal{D}$ at which it is discontinuous, provided that $P(Z \in \mathcal{D}) = 0$. This is useful when combined with the CLT, for which Z is continuously distributed. Hence applying an arbitrary function h to $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ allows us to use the CMT provided that h has only a discrete set of points of discontinuity.

A set of useful/common applications of the CMT are summarized by the so-called *Slutsky's Theorem*:

Theorem 7 (Slutsky's Theorem). *Suppose $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$ with c a constant. Then:*

- $Z_n + Y_n \xrightarrow{d} Z + c$
- $Z_n \cdot Y_n \xrightarrow{d} cZ$
- $Z_n / Y_n \xrightarrow{d} Z/c$ *if* $c \neq 0$.

To see how these results follow from Theorem 6, note that since c is a constant, $Z_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{p} c$ is equivalent to

$$\begin{pmatrix} Z_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z \\ c \end{pmatrix}$$

(see discussion following Proposition B.3). Then we can apply the CMT to the sequence $\begin{pmatrix} Z_n \\ Y_n \end{pmatrix}$, with the following continuous functions h, respectively:

- $h(Z, Y) = Z + Y$
- $h(Z, Y) = Z \cdot Y$
- $Z_n / Y_n = h(Z, Y) = Z/Y$

B.6.2 The delta method

Note that when combined with the CLT, the continuous mapping theorem allows us to talk about the asymptotic distribution of $h\left(\sqrt{n}(\bar{X}_n - \mu)\right)$ for a continuous function h. What is often more useful is to talk about the asymptotic distribution of $\sqrt{n}(h(\bar{X}_n) - h(\mu))$. That is, when we apply a function h to our sample mean, how does the limiting distribution of $h(\bar{X}_n)$ look as it converges around $h(\mu)$? (Exercise: which result allows us to know that $h(\bar{X}_n)$ *does* converge around $h(\mu)$?)

The delta method gives us a tool to address exactly this question:

Theorem 8 (the delta method). *If $\sqrt{n}(Z_n - \mu) \xrightarrow{d} \xi$ for some random vector ξ, then if $h(z)$ is continuously differentiable in a neighborhood of $z = \mu$:*

$$\sqrt{n}(h(Z_n) - h(\mu)) \xrightarrow{d} \nabla h(\mu)'\xi$$

where $\nabla h(z) = (\frac{d}{d_1}h(z), \frac{d}{d_2}h(z), \dots)'$ is a vector of the derivatives of h with respect to each component of Z.

Consider now what this implies in the case of the CLT:

Corollary 1. *If X_i are i.i.d random vectors, $h(x)$ is a function that is continuously differentiable at $x = \mu$, and $\mathbb{E}[X_i' X_i] < \infty$, then*

$$\sqrt{n}(h(\bar{X}_n) - h(\mu)) \xrightarrow{d} N(0, \nabla h(\mu)'\Sigma\nabla h(\mu))$$

where $\Sigma = Var(X_i)$ and $\mu = \mathbb{E}[X_i]$.

Proof. Beginning from Theorem 8, we only need to show that for a random variable $Z \sim N(0, \Sigma)$, $h(\mu)'Z \sim N(0, \nabla h(\mu)'\Sigma\nabla h(\mu))$. We can see this in two steps. First of all, since a linear combination of normal random variables is also normal, we know that $\mathbf{a}'Z$ is normal for any normally-distributed $k-$component random vector and $k-$component vector $\mathbf{a}$. We thus need only to work out the mean and variance of $h(\mu)'Z$ to characterize its full distribution. By linearity of the expectation, $\mathbb{E}[h(\mu)'Z] = 0$, since each component of Z has mean zero. You also showed in HW 3 that the variance of $\mathbf{a}'Z$ is $\mathbf{a}'Z\mathbf{a}$. Substituting $\mathbf{a} = h(\mu)$ completes the proof. □

The most important special case of the corollary above is when X_i is a random variable. In this case, we don't need any matrix multiplication and we have that:

$$\sqrt{n}\left(h(\bar{X}_n) - h(\mu)\right) \xrightarrow{d} N\left(0, \left(\frac{d}{dx}h(\mu)\right)^2 \cdot \sigma^2\right)$$

Note that if the function h is very sensitive to the value of x near μ, i.e. $\frac{d}{dx}h(\mu)$ has a large magnitude, then the asymptotic variance of $h(\bar{X}_n)$ will be large, since the funciton h blows up the variance of X_i by a factor $\left(\frac{d}{dx}h(\mu)\right)^2$.

B.6.3 The Cramér–Wold theorem*

The following theorem, referred to as the Cramér–Wold theorem or the Cramér–Wold "device", is another tool in asymptotic analysis. We won't find it as useful as CMT or delta method, but it's worth seeing so I mention it here:

Theorem 9 (the Cramér–Wold device). *If Z_n is a sequence of random vectors having k components, then $Z_n \xrightarrow{d} Z$ if and only if $\mathbf{a}'Z_n \xrightarrow{d} \mathbf{a}'Z$ for all (non-random) $k-$component vectors $\mathbf{a}$.*

One very important application of the Cramér–Wold device is in extending the central limit theorem to random vectors. In Section B.5, we only proved the CLT for a random variable. The following exercise asks you to derive the multivariate CLT from the univariate CLT.

Exercise: Use the Cramér–Wold device to show that if Theorem 5 applies to random variables X_i, then it applies to a random vector $X_i = (X_{1i}, X_{2i}, \dots X_{ki})'$ as well (assume that any necessary moments exist).

B.7 Limit theorems for distribution functions*

While the law of large numbers might appear to be somewhat limited, in that it only talks about the mean, it is surprisingly versatile. For example, it implies that sample probabilities converge to their population counterparts. Suppose we have an *i.i.d.* collection of X_i and are interested in $F(x)$, the population CDF of X_i evaluated at some specific x. Then we can define $Z_i = \mathbb{1}(X_i \leq x)$, a random variable that takes a value of 1 if $X_i \leq x$, and zero otherwise. Since the collection $\{Z_1, Z_2, \dots Z_n\}$ is *i.i.d*, and has the finite mean:

$$\mathbb{E}[Z_i] = \mathbb{E}[\mathbb{1}(X_i \leq x)] = P(X_i \leq x) = F(x)$$

the law of large numbers implies that the sample mean of Z_i converges in probability to $F(x)$. The sample mean of Z_i is simply

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x) = \frac{\textit{number of } i \text{ for which } X_i \leq x}{\textit{number of } i \text{ in sample}},$$

the proportion of the sample for which $X_i \leq x$. When considering this quantity across all x, we call the resulting function the *empirical CDF* of X_i, denoted as $F_n(x)$.

Thus, for each x the empirical CDF evaluated at x converges in probability to the population CDF evaluated at x, i.e. $F_n(x) \xrightarrow{p} F(x)$. This result can be strengthened in two ways (which are not implied by the weak law of large numbers). Consider the error in $F_n(x)$ as an approximation of $F(x)$, $|F_n(x) - F(x)|$ as a function of x. This may be larger or smaller depending on x. The *Glivenko-Cantelli theorem* states that even the largest error, over all x, converges to zero, and furthermore that this convergence is almost sure convergence (see box at the end of Section B.4), rather than convergence in probability:

Theorem 10 (Glivenko-Cantelli theorem). *If X_i are i.i.d, then:*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

We won't use Theorem 10 in this class, but it can be useful for proving properties of asymptotic sequences that involve quantities that cannot be written as a function of $\bar{X}_n$.

Appendix C

Statistical decision problems

This chapter presents a formal view of the goals of using statistics for econometrics. It starts with the question: what is it that we would like to learn? Once we've defined our "parameter of interest", we can separate much of econometrics into three parts: identification, estimation and inference.

I will not attempt at a thorough or rigorous treatment of many of the concepts this chapter touches upon. Rather, I hope it can present a unified way to think about several concepts you have probably seen in one form or another in previous courses, and serve either as a reference or a starting point to exploring terms in econometrics as you come across them in your own research.

C.1 Step one: defining a parameter of interest

Why do we use statistics? A short answer is that we want to learn things about the world, and data is the lens with which we investigate some population within it. A more careful answer, which is well-aligned with the specific approach that econometrics takes to using statistics, is that there are specific features θ of the world that we care about.

We can distinguish between three types of *parameter of interest*, θ.

First type (model parameters): Think back to the idea of a parametric statistical model, introduced in Chapter **??**. Suppose we observe *i.i.d* data X_i, where the distribution of X_i is thought to belong to a parametric family $F(\cdot;\theta)$ for some $\theta \in \Theta$. For example, we might be willing to assume that X_i is a normally distributed random variable, with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. In this case, $\theta = (\mu, \sigma)$, and in the absence of any further assumptions about θ: $\Theta = \mathbb{R} \times \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of all positive real numbers (the variance cannot be negative). In this context, it is natural to take the full vector of model parameters θ to be our parameter of interest (of course, we might only be interested in e.g. μ, in which case μ alone is our parameter of interest, and similarly with σ).

Second type (features of observed variables in the population): We don't need parametric statistical models to talk about parameters of interest, however. If we have *i.i.d.* data drawn from any population distribution F, we might think of some aspect of F that we'd like to know. For example, we might be interested in $\mathbb{E}[X_i]$, but don't want to assume that X_i is normally distributed, as in the last example. Then, our parameter of interest is $\theta = \mathbb{E}[X_i]$. Another parameter of interest might be the median of F, the point x at which $F(x) = 1/2$. In this case, $\theta = \inf\{x : F(x) \geq 1/2\}$ (this general definition allows for a non-continuous F, in which case there may be no x such that $F(x) = 1/2$ exactly).

Third type (quantities that depend on unobservables): One of the exciting and difficult things about applied econometrics is that often our parameters of interest do not depend solely on the distribution F of the vector of variables X that we observe in our data. Rather, θ often depends also on the distribution of some other variables U that are *not* observed. This situation most often arises when discussing causality, for example when our parameter of interest summarizes the causal effect of a policy. Talking about causality involves some new notation and concepts, so we'll defer further discussion to Chapter 1. As a simpler example of a situation that involves unobservables, let us consider a different important practical problem: measurement error.

Suppose our parameter of interest is $\theta = \mathbb{E}[Z_i]$, the average value of some random variable Z_i. However, our data was not recorded perfectly, and instead of an *i.i.d* sample of Z_i, we observe an *i.i.d* sample of $X_i = Z_i + U_i$, where U_i represents unobserved "measurement error". In this case, our parameter of interest can be written as $\mathbb{E}[X_i - U_i]$, which depends both upon the distribution of X and the distribution of U.

C.2 Identification

Once we have a parameter of interest in mind, a good starting point is often to ask the question: "could I determine the value of θ if I had access to the population distribution F underlying my data?".

If the answer is no, then no amount of statistical wizardry will allow you to learn the value of θ. If the answer is yes, then we say that θ is *identified*.

Definition C.1. *Given a statistical model $\mathcal{F}$ for (X, U), we say that θ is **identified** when there is a unique value θ_0 of θ compatible with F_X, the population CDF of observable variables X.*

Often identification is described as saying that if we observed an "infinite" sample, we could determine the value of θ. The reason for this is that by the law of large numbers, we can learn the entire population distribution of X from an *i.i.d* sample X_i, as the sample size goes to infinity (see discussion in Section B.7). Of course, we never observe an infinitely large dataset, but defining identification in terms of what we *could* know if we did cleanly separates problems of research design from the statistical problem of having too small a sample.

Whenever our parameter of interest is defined directly from the population distribution F_X of observables (e.g. $\theta = \mathbb{E}[X_i]$), it will be identified. Thus, parameters of the second type are always identified. This logic often applies to parameters of the first type as well, except in cases when $F(\cdot; \theta)$ doesn't always change with θ (see example below). Questions of identification usually arise in the third case, when our parameter of interest θ depends on the distribution of unobservables: for example when we're interested in causality, have measurement error, or have "simultaneous equations".

Example: Suppose X_i are *i.i.d* draws from $N(\mu, \sigma^2)$. Then the parameters μ and σ are *identified*, because each pair (μ, σ) gives rise to a different CDF F_X of X_i.

Example: Suppose X_i are *i.i.d* draws from $N(\min\{\theta, 5\}, \sigma^2)$. Then θ is not identified, because different values of θ (e.g. $\theta = 6$ vs. $\theta = 7$), do not give rise to a different CDF F of X_i.

Example: In the measurement error example, suppose that we're willing to assume that $\mathbb{E}[U_i] = 0$, that the measurement error averages out to zero (e.g. there are equal chances of getting positive and negative errors of the same magnitude). Then $\theta = \mathbb{E}[Z_i]$ is identified, since now $\mathbb{E}[Z_i] = \mathbb{E}[X_i]$. This example underscores the role of $\mathcal{F}$ in Definition C.1. Whether or not θ is identified often depends on what assumptions we are willing to make, which restrict the set $\mathcal{F}$ of possible joint-distributions for (X, U).

Below I discuss some additional issues related to identification, which may relate to terms you've heard floating around about identification:

Parametric vs. non-parametric identification: When $\mathcal{F}$ is a non-parametric statistical model, in the sense described in Section **??**, we say that θ is *non-parametrically identified*. We have non-parametric identification when we do not need to specify a parametric functional form for the distribution of observables or unobservables. Sometimes we only have parametric identification but not non-parametric identification. Suppose, in the measurement error example, our parameter of interest is full distribution function F_Z of Z_i, and are willing to assume that $U_i \perp\!\!\!\perp Z_i$. Then F_Z is identified if we are willing to specify the exact form of F_U, e.g. $U_i \sim N(0, 1)$, through a technique known as *deconvolution*. However, F_z is not non-parametrically identified.

Partial vs. point identification: Sometimes knowing F_X is not enough to pin down the value of θ, but it is enough to determine a *set* of values that θ might take. For example, we may be able to

determine upper and lower bounds for θ. In such cases we often say that θ is *partially identified*. This can be contrasted with Definition C.1, which describes *point identification*.

Identification of a parametric model: Suppose we have an *i.i.d* sample of observables X_i and a parametric statistical model for (X_i, U_i), in the language of Section **??**. Then we might say the model is identified, when the full vector θ of model parameters are identified in the sense of Definition C.1:

Definition C.2 (full identification of a model). *Given a statistical model $\mathcal{F}$ for (X, U), we say that the model is **identified** when when the set $\{\theta \in \Theta : F_X(\cdot) = F_X(\cdot, \theta)\}$ is a singleton, where F_X is the CDF of X.*

Definition C.1 says that there is a unique value $\theta_0 \in \Theta$ such that $F_X(\cdot, \theta_0\}$ is equivalent to the population distribution of observables X_i. This situation arises often in econometrics in the context of so-called *structural* models in which the entire model can be characterized by a finite set of model parameters.

C.3 Estimation

If our parameter of interest θ is identified, then we can move on to our next question: how can we estimate it?

In this section, we treat the task of estimating θ as a decision problem. In the next section, we'll take the same approach to testing hypotheses about θ. This way of thinking about estimation and inference is called *statistical decision theory*.

Let's think about the task of estimating θ as a problem of choosing an optimal strategy in a particular game, which we play along with "nature". Nature goes first, giving us a sample $\mathbf{X}$, the distribution of which we denote abstractly as P (this is equivalent to the joint-CDF of all of the components of $\mathbf{X}$). Our goal is to think about how to form $\hat{\theta} = g(\mathbf{X})$ as a function of the data $\mathbf{X}$. How should we proceed?

Recall that in game theory, a *strategy* is a complete profile of what we would do, given whatever the other players do. In this context, we a strategy is not a particular numerical estimate of θ, but the function g. For example, if our estimator is the sample mean, then $g(\mathbf{X}) = g(X_1, X_2, \ldots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, which will depend upon the particular values of X_i occur in our sample.

As in game theory, our best-response to the actions of nature will depend upon our preferences (a.k.a. our utility function). In statistical decision theory this takes the form of a "loss function": $L(\hat{\theta}, \theta_0)$, where θ_0 is the true value of θ. For the most part, we consider the so-called quadratic loss function:

$$L(\hat{\theta}, \theta_0) = ||\hat{\theta} - \theta_0||_2^2 := (\hat{\theta} - \theta_0)'(\hat{\theta} - \theta_0)$$

When θ is a scalar, then this is just the square of the difference between our estimator $\hat{\theta}$ and the true value θ_0.

However, remember that $\hat{\theta} = g(\mathbf{X})$ is a random variable/vector, which depends on our randomly drawn dataset $\mathbf{X}$. Thus to pick a strategy g, we need to define our preferences over "lotteries", again–as in standard game theory. In line with expected utility theory, the convention here is to take our optimal action g to be the minimizer of *expected* loss: $\mathbb{E}[L(\hat{\theta}, \theta_0)]$ where the expectation is over the distribution of $\mathbf{X}$. The *risk* function $R_g(\theta)$ of estimator g views the expected loss as a function of the true value of θ. It is common to write this as $\mathbb{E}_\theta[L(\hat{\theta}, \theta)]$, where the notation $\mathbb{E}_\theta$ makes it clear that the distribution of $\mathcal{X}$ must depend in some way on the value of θ. This is motivated by cases in which we have *i.i.d.* data from a parametric statistical model where θ indexes the population distribution of X_i. Then the distribution of $\mathbf{X}$ depends on just two things: n and the true value of θ.

When we use the quadratic loss function, the optimal estimator g would be

$$g^* := \underset{g}{\text{argmin}}\ \mathbb{E}[||g(\mathbf{X}) - \theta_0)||_2^2] \tag{C.1}$$

However, solving this problem is not easy, because we generally don't know the distribution of $\mathbf{X}$ ex-ante. However, statisticians have developed various strategies to try to keep $\mathbb{E}[||g(\mathbf{X}) - \theta_0)||_2^2]$ small. These

strategies are best understood as ways to navigate the so-called *bias-variance tradeoff*. The following proposition shows that expected quadratic loss can be decomposed into two terms: one capturing the square of the "bias" of the estimator, and the other capturing its variance.

For simplicity, we state this result in the special case that θ is a scalar. We'll also just write $\hat{\theta}$ rather than $g(\mathbf{X})$, to keep the notation simple.

Proposition C.1 (the bias-variance decomposition).

$$\underbrace{\mathbb{E}[(\hat{\theta}-\theta_0)^2]}_{\text{expected loss}} = \underbrace{\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]))^2]}_{\text{variance of } \hat{\theta}} + \left(\underbrace{\mathbb{E}[\hat{\theta}]-\theta_0}_{\text{bias of } \hat{\theta}}\right)^2 \tag{C.2}$$

Proof. Add and subract $\mathbb{E}[\hat{\theta}]$ to obtain:

$$\mathbb{E}[\Big\{(\hat{\theta}-\mathbb{E}[\hat{\theta}])+(\mathbb{E}[\hat{\theta}]-\theta_0))\Big\}^2] = \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]))^2] + 2\cdot\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta_0)] + \Big(\mathbb{E}[\hat{\theta}]-\theta_0\Big)^2$$

Now observe that the middle term is zero, because

$$\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta_0)] = (\mathbb{E}[\hat{\theta}]-\theta_0)\cdot\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])] = (\mathbb{E}[\hat{\theta}]-\theta_0)\cdot\Big(\cancel{\mathbb{E}[\hat{\theta}]}-\cancel{\mathbb{E}[\hat{\theta}]}\Big) = 0$$

since $(\mathbb{E}[\hat{\theta}]-\theta_0)$ is just a non-random number. □

Equation (C.2) is described as a bias-variance *tradeoff* because often strategies to decrease bias come at the expense of increasing variance, and vice-versa. Suppose for example that we just pick $g(\cdot)=5$, estimating θ to be 5, regardless of what sample we see. This estimator will have zero variance! But we can expect the bias $5-\theta_0$ to be quite large. On the other hand, extremely flexible estimation methods are often good at minimizing bias, but doing so may increase variance. The field of *non-parametric* estimation chooses estimators to explicitly navigate this tradeoff.

C.3.1 Desirable properties of an estimator

This section investigates some desirable properties of an estimator, in light of the bias-variance tradeoff. It is not meant to deliver a detailed account of these properties, but simply to serve as a reference for what the associated terms mean. Please see the course textbooks for more details.

C.3.1.1 Consistency

The first thing that we might ask of our estimator is that it be *consistent*. What we mean by that is that

$$\hat{\theta} \xrightarrow{p} \theta_0$$

regardless of the value of θ_0. Consistency means that as n goes to infinity, the entire expected loss in Equation (C.2) converges to zero.

C.3.1.2 Rate of convergence

Consider the sample mean $\bar{X}_n$ viewed as an estimator of the population mean $\mu=\mathbb{E}[X_i]$. We know by the LLN that $\bar{X}_n$ is a consistent estimator of μ, and we furthermore know by the CLT that

$$n^p(\bar{X}_n-\mu) \xrightarrow{d} N(0, Var(X_i))$$

if we set $p=1/2$. Note that if we set the power p on n to be any larger than $1/2$, then the LHS would blow up, rather than converging in distribution to anything (like a normal distribution). On the other hand, if we had set $p<1/2$, then $n^p(\bar{X}_n-\mu)$ will simply converge in probability to zero. $1/2$ is "Goldilocks" level of p in which we get a non-degenerate asymptotic distribution for $n^p(\bar{X}_n-\mu)$.

In general, when we have a consistent estimator $\hat{\theta}$, we call the maximum value of p such that $n^p(\hat{\theta}-\theta_0)$ converges in distribution to something (technically, to some distribution that is "bounded in probability") the *rate of convergence* of $\hat{\theta}$. The rate of convergence of the sample mean is $1/2$, and we often say that it

is $\sqrt{n}-$*consistent*. $\sqrt{n}$-consistency is a desirable property, which is shared by many common estimators. However, some estimators have a slower rate of convergence. For example, suppose we'd like to estimate the density $\theta = f(\mathbf{x})$ of a $d-$dimensional random vector X_i at some point $\mathbf{x}$, and we'd like to make this estimation *non-parametric*—that is, not based on assuming a parametric model for $f(\mathbf{x})$.

We can do so using the so-called kernel density estimator $\hat{f}_K(\mathbf{x})$, which has a rate of convergence no better than $p = \frac{2}{d+4}$. When $d = 1$, for example, we can only blow up $(\hat{f}_K(\mathbf{x}) - f(\mathbf{x}))$ by a factor of $n^{2/5}$ and get an asymptotic distribution. In practice, this means that we need a *larger* sample n for asymptotic arguments to provide good approximations to the sampling distribution of $\hat{f}_K(\mathbf{x})$. This becomes a real problem as d starts to increase: for example, the rate of convergence $\hat{f}_K(\mathbf{x})$ when $d = 5$ is just 2/9. This problem is often referred to as the *curse of dimensionality*, and is why we need very large samples—and/or even more clever techniques—to do non-parametric estimation with many covariates.

C.3.1.3 Unbiasedness

If an estimator is *unbiased* if it manages to make the second term in Equation (C.2) zero, that is:

$$\mathbb{E}[\hat{\theta}] = \theta_0$$

Unbiasedness has a nice interpretation: we know that $\hat{\theta} \neq \theta_0$ in general, but we know that $\hat{\theta}$ will be right *on average*, over different realizations of our dataset.

An example of an unbiased estimator is the sample mean, when our parameter of interest θ is the population mean $\mathbb{E}[X_i]$. In Section B.2, we showed indeed that $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_i]$, regardless of n or the true value of $\mathbb{E}[X_i]$ (so long as it exists).

Note that an estimator can be consistent without being unbiased. For example, the estimator $\hat{\theta} = \frac{n+1}{n}\bar{X}_n$ is biased as an estimator for $\theta_0 = \mathbb{E}[X_i]$, because

$$\mathbb{E}[\hat{\theta}] - \theta_0 = \frac{n+1}{n} \cdot \theta_0 - \theta_0 = \frac{\theta_0}{n} \neq 0$$

unless $\theta_0 = 0$. However, this $\hat{\theta}$ is consistent. This implies that as n approaches infinity, both its bias and its variance converge to zero. If an estimator has an asymptotic bias (that is, a bias that doesn't go away with n), then it cannot be consistent.

C.3.1.4 Efficiency

Econometricians often speak of an estimator as being efficient. Loosely speaking, this typically means that $\hat{\theta}$ minimizes mean squared error (C.2) among some class of estimators.

For example, we might consider the class of unbiased estimators, and ask whether a given $\hat{\theta}$ minimizes Eq. (C.2). Since the bias term is zero for all estimators in this class, the efficient estimator will be the one that minimizes variance.

In the context of parametric models, the *Cramer-Rao lower-bound* establishes the smallest variance that an unbiased estimator can possibly have (even when θ is a vector, though the definition of "smallest" here requires qualification). The maximum likelihood estimator, discussed in the next section, achieves this bound: it is thus efficient, whenever it happens to be unbiased (which is not guaranteed in general).

A related notion is *asymptotic efficiency*, which says that an estimator is efficient as $n \to \infty$.

C.4 Statistical Inference*

In Section C.3, our goal was to deliver a *point-estimate* $\hat{\theta}$ of our parameter of interest. That is, we want a number that yields something close to the true value θ_0 of θ.

Sometimes we can settle for a less ambitious goal, which is to ask not what the exact value of θ_0 is, but rather we want to know whether or not θ_0 belongs to some *set* of values. I will discuss two approaches of this type: i) *hypothesis testing*, in which we want to test whether $\theta_0 \in \Theta_0$ for some fixed set Θ_0; and ii) *interval estimation*, in which we want to construct a set $\hat{\Theta}$ that has some desirable relationship to θ_0 (for example contains θ_0 with high probability)

C.4.1 Hypothesis testing

Beginning with some overall space of admissible values Θ (e.g. the real numbers), let us carve the space into two sets: Θ_0 and Θ_1, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. We call our hypothesis that $\theta_0 \in \Theta_0$ the *null-hypothesis*:

$$\text{(Null hypothesis)} \quad H_0 : \theta_0 \in \Theta_0 \qquad\qquad \text{(Alternative hypothesis)} \quad H_1 : \theta_0 \in \Theta_1$$

Note that provided that our model $\theta_0 \in \Theta$ is correctly specified, either the null hypothesis H_0 or the alternative hypothesis H_1 holds.

Continuing of the approach of statistical decision theory, we may think of our action space as now as consisting of two actions $d \in \{a, r\}$, either accept (a) or reject (r) the null-hypothesis H_0. This can be contrasted with estimation, in which our action space was to pick a specific value in Θ to serve as an estimate for θ.

In this context, a *strategy* is a mapping from the possible datasets $\mathbf{X}$ that we might see to an action $\{a, r\}$. This function $d(\mathbf{X})$ is referred to as a decision rule, or a *test*. To think about what kind of a test might be optimal, we again need to specify our preferences, or a loss function, over actions. Compared with estimation, in which our loss function took the form $L(\hat{\theta}, \theta)$, it now takes the form $L(d, \theta_0)$: how happy would we be with our decision $d \in \{a, r\}$, if we learned the true value of θ was θ_0?

Compared with estimation—where the quadratic loss function is very standard—in testing it is less obvious what our cost function could be. One thing is clear however, we'd prefer not to be *wrong*: we don't want to reject the null hypothesis (often referred to as *failing to accept* the null) when in fact $\theta\Theta_0$, and we also don't want to accept the null hypothesis when in fact $\theta_0 \in \Theta_1$. The first of these errors is called a Type-I error (falsely rejecting H_0) while the second is called a Type-II error (incorrectly accepting H_0).

The most basic loss function we might think of is called *0-1 loss*, and only cares about *whether* we are right or not, i.e. $L(d, \theta_0)$ when either $d = a$ and $\theta_0 \in \Theta_0$ or $d = r$ and $\theta_0 \in \Theta_1$ (i.e. we are right), and $L(d, \theta_0)$ otherwise (we are wrong). Recall that since $\mathbf{X}$ is random, our decision $d(\mathbf{X})$ will be random, and thus we can again think about the *risk*, or expected loss, due to a particular strategy d. With the $0-1$ loss function:

$$\mathbb{E}[L(d(\mathbf{X}), \theta_0)] = \begin{cases} P(d(\mathbf{X}) = r) & \text{if } \theta_0 \in \Theta_0 \quad \text{(Type-I error)} \\ P(d(\mathbf{X}) = a) & \text{if } \theta_0 \in \Theta_1 \quad \text{(Type-II error)} \end{cases}$$

It is clear from the above that whether or not the null is actually true determines *which* probability matters in determining the risk of the test.

Since the value of θ pins down some aspect of the distribution of $\mathbf{X}$, the probability of rejecting the null will depend upon what the true value of θ_0 in fact is. Like the risk function that we saw in estimation, let us use the notation $P_\theta(d(\mathbf{X}) = r)$ to denote the probability of rejecting when the true value is θ. Viewing this as a function of θ, we define the *power function* $\beta(\theta)$ of test d.

Beyond the $0-1$ loss function, we might put a different penalty on Type-I vs. Type-II errors:

$$L(a, \theta_0) = \begin{cases} 0 & \text{if } \theta_0 \in \Theta_0 \\ \ell_{II} & \text{if } \theta_0 \notin \Theta_0 \end{cases} \qquad \text{while} \qquad L(r, \theta_0) = \begin{cases} 0 & \text{if } \theta_0 \in \Theta_1 \\ \ell_I & \text{if } \theta_0 \notin \Theta_1 \end{cases}$$

The ratio ℓ_{II}/ℓ_I will govern whether our test d should be more conservative about avoided Type-I errors, or about avoiding Type-II errors.

C.4.2 Desirable properties of a test

As with estimation problems, choosing the optimal test d is a hard problem because we don't know the distribution of $\mathbf{X}$, we can only approximate it using the dataset $\mathbf{X}$ that we actually observe, along with whatever assumptions we are willing to make. As again with estimation, there are a few principles that are used to help guide the design of statistical tests.

C.4.2.1 Size

The *size* α of a test d is the maximum probability of making a Type-I error (falsely rejecting), over all $\theta \in \Theta_0$. We can write this in terms of the power-function $\beta(\theta)$ as:

$$\alpha = \sup_{\theta_0 \in \Theta_0} \beta(\theta)$$

We'd like the size of a test d to be small; we therefore often design tests to control their size (keep it below a certain value). Often we can do this in the asymptotic limit (as $n \to \infty$) even if we do now know the size of a test in finite sample.

C.4.2.2 Power

The power of a test is given by its power function $\beta(\theta)$. We generally want to increase $\beta(\theta)$ among the $\theta \in \Theta_1$, to reduce the probability of a Type-II error.

C.4.2.3 Navigating the tradeoff

In general, the two desiderata of a) a small size; and b) large power, are at tension with one another. A test that always rejects, regardless of the data $\mathbf{X}$, will never make a Type-II error (have lots of power), but may be extremely likely to make a Type-I error (have large size). On the other hand, a test that always accepts will never make a Type-I error (have low size) but may be making lots of Type-II errors (have low power). Often we approach testing by choosing a *significance-level* p ex-ante, (e.g. $p = .05$), and then design the test so that it's size is no greater than p. Given that constraint, we then try to make the power of the test as large as possible (which usually means making it's size exactly p).

C.4.3 Constructing a hypothesis test

The most common variety of hypothesis test takes the following form: from the data $\mathbf{X}$ we compute some *test statistic*, call it T_n. Then we compare T_n to some *critical value* c, and choose to reject the null-hypothesis if and only if $|T_n|$ exceeds the critical value (a so-called *two-sided test*), or alternatively if T_n exceeds the critical value (a so-called *one-sided test*).

Tests of this form are usually motivated by knowing the asymptotic distribution of T_n, i.e. $T_n \xrightarrow{d} T$ where T has some known distribution. Then we can control the size of our test by choosing c to be such that $P(T \leq c) \geq 1 - \alpha$. We then maximize power subject to this contraint on size by choosing c to be exactly the $1 - \alpha$ quantile of T (and no lower), so that $P(T \leq c) = 1 - \alpha$.

Example: Let us close by illustrating some of the concepts of this section with an example. Suppose our statistical model is that $X_i \sim N(\theta_0, 1)$, i.e. a normal random variable with unit variance but unknown mean θ_0. We wish to test whether $H_0 : \theta_0 = 0$, that is: $\Theta_0 = \{0\}$ and $\Theta_1 = \mathbb{R}/\{0\}$. Let our test statistic be $\sqrt{n}$ times the sample mean $T_n = \sqrt{n} \cdot \bar{X}_n$. Given our model, the sample mean has the exact distribution $\bar{X}_n \sim N(\theta_0, 1/n)$ for any n, and hence $T_n \sim N(\theta_0, 1)$. Under the null, T_n is a standard normal (since then $\theta_0 = 0$) and hence for a two-sided test we can choose our critical value c to be the $1 - \alpha/2$ quantile of the standard normal distribution (then $P(|T_n| > c) = P(T_n < -c) + P(T_n > c) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$). Note that the power function $\beta(\theta)$ of this test is the probability that a $N(\theta, 1)$ random variable has absolute value greater than c, which is equal to $\Phi(c - \theta) + \Phi(-c - \theta)$, where Φ denotes the standard normal CDF.

C.4.4 Interval estimation and confidence intervals

The goal of interval estimation is to choose an a set $\hat{\Theta}$ of values that with high probability contains the true value θ_0. We call this interval estimation because $\hat{\Theta}$ typically corresponds to an interval $[a, b]$ (if θ is one-dimensional), or some higher-dimensional analog of an interval (e.g. a region). By contrast, estimation in the sense of Section C.3 is by contrast referred to as *point*-estimation.

As with point estimation and testing, our action $\hat{\Theta}$ is a function of the data (however now this is a set-valued function)—call it $s(\mathbf{X})$. The *coverage probability* of an interval estimator s is the probability that it contains the true value of θ_0 Here the tradeoff is between increasing the coverage probability, but without making the the interval too big (in which case we haven't learned much about the value of θ_0). Thus with interval estimation, we might define our loss function to depend both on the coverage probability and the length of the interval estimate.

As with estimation, we *do* care about the specific value of θ, not just whether or not some hypothesis H_0 about it is true. However, we'll now see that there is a very close connection between interval estimation and hypothesis testing.

One scenario in which we might implement interval estimation is when our parameter of interest is only partially identified (see Section C.2). In such a setting, for example, our model might only imply that $\theta_0 \in [\theta_L, \theta_H]$, where the bounds θ_L and θ_H are themselves point identified. Then we can construct an interval estimate of θ_0 with the set $\hat{\Theta} = [\hat{\theta}_L, \hat{\theta}_H]$, given estimators of each of the two bounds.

The much more common scenario in which we engage in interval estimation is when constructing a *confidence interval* for θ_0. We do this even when θ_0 is identified and we have a consistent estimator for it. A confidence interval makes a much more credible than a point estimate. In fact, point-estimation is just a special case of interval estimation in which we constrain our $\hat{\Theta}$ to be a singleton. While singleton will sets typically have zero probability containing θ_0 (though they may be very close to it with high probability), confidence intervals allow us to deliver an interval estimate of θ_0 that takes sampling uncertainty into account.

C.4.4.1 Confidence intervals by test inversion

The most popular method for constructing confidence intervals is to perform a hypothesis test having size α for the null $H_0 : \theta_0 = \theta$, for each conceivable value of θ. Then, collect the set of all values θ that are not rejected by that test to form our interval estimate of θ_0. That is:

$$\hat{\Theta} = \{\theta \in \Theta : d(\mathbf{X}) = a\}$$

This process is often referred to as *test inversion*, and the resulting $\hat{\Theta}$ is called a $(1-\alpha)$-confidence interval $\mathcal{CI}^{1-\alpha}$. For example, if we used a test with size 5%, then the resulting confidence interval is called a 95% confidence interval.

Example: Suppose we apply this principle to the example in Section C.4.3 in which $X_i \sim N(\theta_0, 1)$. There we constructed a test for the null hypothesis that $\theta_0 = 0$, but now we need to consider more general hypotheses of the form $H_0 : \theta_0 = \theta$. If we revise our test statistic to be $T_n(\theta) = \sqrt{n} \cdot (\bar{X}_n - \theta)$, we again have that T_n has a standard normal distribution asymptotically, and thus our critical value c is unchanged from the $\theta = 0$ case. A $1-\alpha$ confidence interval would thus be:

$$\mathcal{CI}^{1-\alpha} = \{\theta \in \mathbb{R} : |T_n(\theta)| \leq c\} = \{\bar{X}_n - c/\sqrt{n}, \bar{X}_n + c/\sqrt{n}\}$$

where c is the $1-\alpha$ quantile of the standard normal distribution.