2 The potential outcomes framework

This chapter introduces the concept of potential outcomes to define causal effects. The potential outcomes framework has redefined and clarified our understanding of causality. In particular, this framework allows us to establish the conditions under which a measure of association implies a causal effect. We can determine these conditions before an experiment or study is conducted, and without a reference to a particular statistical model or method of estimation.

We introduce the concept of potential outcomes to define causal effects for a single person (more generally, unit). The causal effect is a comparison of hypothetical states, one state in which treatment, which in our field is often a policy change, is applied and one in which treatment is not applied to a unit. Since we only observe one of the states because a unit either receives or not a treatment, the other state is considered a counterfactual. Thus, the challenge of estimating causal effects is to predict the counterfactual. To make predictions, we use observations from multiple units.

The next challenge is to establish the conditions under which predictions are accurate. As we will see, a key to determine these conditions is to understand the assignment mechanism. Why did some units end up receiving treatment and others did not? We will cover three assignment mechanisms: simple randomization, conditional randomization, and non-controlled mechanisms generating observational data. We then discuss the estimation of causal effects using different approaches, including one that closely follows the definition of causal effects by explicitly using predictions.

Besides clarifying the key assumptions needed to establish causal effects, this chapter introduces the potential outcomes notation that we will use in the rest of the book. This notation is somewhat difficult to learn at first, but it is important to master it because it makes difficult concepts much easier to understand.

2.1 Potential outcomes

Causality is linked to a manipulation or intervention – an action that is applied to a subject or person. We will use the general term **unit** throughout this chapter (see Box **??**). A unit could either receive the intervention or not. We often use the word “treatment,” although not all interventions are medical treatments, and in other fields “exposure” is a common term. The unit that receives treatment is the “treated” unit. In contrast, the unit that does not receive the treatment is the “control” unit. We can think of the unit as the object that is subjected to the intervention. Focusing on only two treatments simplifies the discussion and the notation in this chapter, but an intervention could have more than two levels of treatment. For example, different doses of a medication or number of stars a nursing home receives in a 5-star rating system (i.e., five levels of treatment). In general, the results we discuss extend to **multi-level** (or **multi-valued**) **treatments**. In the context of this book, the treatment is often a policy change, like assigning nursing homes a rating or expanding insurance benefits to include preventive medical visits with no cost-sharing.

For each pair of treatment and unit, there is an associated **potential outcome**. These outcomes are "potential" because not all of them will actually happen. This is a crucial – and initially confusing – aspect in this framework. We can discuss in abstract all potential outcomes, but after an intervention or action is applied to a unit, we can only observe one of the potential outcomes since a unit either received the treatment or not, but never both. In other words, one potential outcome is realized while the other remains potential or **counterfactual**. In this framework, the same unit now is not the same a minute from now – it is a different unit.

Suppose, for example, that the unit is an individual and the treatment is providing the individual with health insurance with no deductible at the start of the year. The alternative intervention is health insurance with a high deductible (a deductible is the amount paid out-of-pocket before the insurance pays any medical expense). We could define the outcome as a measure of financial strain at the end of the year or a measure of healthcare utilization, since previous evidence suggests that deductibles reduce utilization. We denote the potential outcome for this individual Y^1 if she receives the treatment. If she does not receive the treatment, the potential outcome is Y^0. After the treatment is applied, the observed outcome is Y. If she received the intervention, then the *observed* outcome Y is the same as the potential outcome Y^1, and Y^0 is the counterfactual outcome. We can generalize the notation for all individuals by adding a subscript i to potential and observed outcomes.

2.2 Definition of causal effects

The *definition* of the causal effect is a *comparison* of potential outcomes Y_i^1 and Y_i^0. In our example, the causal effect of zero deductible insurance for an individual i is the comparison of financial strain if the individual had a no-deductible health plan to the financial strain if the individual had a high-deductible plan. In other words, the causal effect of the intervention is a comparison of *a priori* hypothetical states. We do not need to conduct an intervention to *define* causal effects.

We could compare outcomes in different ways. We want to use a metric to make comparisons because we are also interested in the **magnitude** of the causal effect. In most situations, we want to determine if X causes Y, but we are also interested in measuring how large is the causal effect of X on Y. The most common metric is a difference or **contrast**:

$$\delta_i = Y_i^1 - Y_i^0.$$

We could also use a relative measure like the ratio of potential outcomes Y_i^1 / Y_i^0 or we could express the causal effect as a percent change $(\frac{Y_i^1 - Y_i^0}{Y_i^0}) \times 100$. Or perhaps we can think of Y as an event with an associated probability, which we could compare as a **relative risk** or as a **risk difference** (marginal effect) $\Pr(Y_i^1) - \Pr(Y_i^0)$. We will come back to these metrics when we discuss the *estimation* of treatment effects.

The definition of causal effects may appear both obvious and unnecessarily intricate. It is obvious in the sense that it agrees with our intuitive understanding of cause and effect. A similar notion of causality appears in most time travel movies. If we could go back in time to prevent event A from occurring, then event B would not have occurred; therefore, we think that A caused B because we imagine two potential scenarios, one with event A occurring and one with event A not occurring. Of course, we do not know the rules of time travel, but the reasoning on time travel movies reveals how we think about causality. A common plot twist is that after changing event A, it turns out that event B happens anyway. That means that A was not the cause and the main character must go back to the past to change another event.

This view of causality is also encoded in the US legal system. In the law that covers most civil cases (Tort Law), the "but-for" principle is used to establish proximate cause, the primary cause of an injury. The but-for rule considers if an injury would not have occurred *but for* a defendant's negligent act (or omission). In other words, the principle says that a jury must establish a counterfactual. Would the injury have occurred if an action had not occurred (or an omission occurred)? If the answer is yes, then the defendant's action was not the proximate cause of the injury.

The intricate part of a definition of causal effects based on potential outcomes is that it is not clear that we gain much by discussing hypothetical states. Our definition does not take into account the observed outcome Y. We can *define* treatment effects before an intervention or in the absence of one. However, we can later conduct the intervention or observe data, which would allow us to measure *some* of the potential outcomes. After conducting an intervention, we only know one version of the events. We know that events A and B occurred, but we do not know what would have happened if event A had not occurred. If the individual is provided with health insurance without deductibles, we only know what happens when the person has health insurance with no deductibles. Before an intervention or before data are collected, the potential outcomes exist only in our minds. Although defining causality using hypothetical events is intuitive, it has not always been a leading definition of cause and effect; most attempts to define causality used *observed* events (See Section 2.14 for suggested readings).

2.3 The fundamental problem of causal inference

We now have a *definition* of causal effects based on potential outcomes, but as the previous discussion foreshadows, we also have a big problem: we can only observe one potential outcome, not both. This problem is referred to as the **fundamental problem of causal inference** or the **fundamental problem of program evaluation**. Although we now have a problem, a useful consequence of defining causal effects as a comparison of potential outcomes is that it clarifies the nature of the problem – the challenge of *estimating* causal effects with observed data. What we need to do is *predict* what would have happened under the counterfactual scenario. Put it this way, we can think of causal inference as a **missing data** problem. Our task is to predict or **impute** the missing data.

If we take a medication, then we observe the presence or absence of a symptom (the outcome) under the treatment state, but we do not know the counterfactual, what would have happened if we had not taken the medication. If we wanted to estimate the causal effect of taking the medication, we would need to predict the outcome if we had not taken the medication. If the individual did not have a high-deductible plan, we would observe financial strain at the end of the year, but we could not establish the causal effect of having a high-deductible plan since we would need to predict what would have happened if the person had a no-deductible plan, the counterfactual.

To clarify these concepts, we can write the observed outcome Y_i as a function of potential outcomes under the two treatments, denoted by D_i. If a unit i receives treatment, then $D_i = 1$, and $D_i = 0$ if the unit does not receive treatment. The function relating observed outcomes to potential outcomes is

$$Y_i = Y_i^1 D_i + (1 - D_i)Y_i^0. \tag{2.1}$$

Or, alternatively,

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0. \end{cases} \tag{2.2}$$

Using this notation, we can restate the fundamental problem of causal inference: if a unit receives treatment, then $Y_i = Y_i^1$; thus, we need to predict Y_i^0 because Y_i^0 is the unobserved counterfactual. On the other hand, if the unit does not receive treatment, $Y_i = Y_i^0$. We now need to predict the counterfactual Y_i^1.

2.4 Multiple units to make predictions

To make some progress in solving the fundamental problem of causal inference, we rely on the experience of multiple units to make predictions about unobserved counterfactuals. We could, for example, observe a group of people who had no-deductible plans during the year and a group of people who had high-deductible plans. At the end of the year, we observe a measure of financial strain to make comparisons between the groups. Or going back to the example of a medication, we could use our own personal experience to estimate the counterfactual, which is how we often figure out what works for us. Each of us at different times is a different unit. If we used the medication many times before, we can make predictions about our symptoms when we take the medication and when we do

Box 2.1 Notation: Writing potential outcomes

There is no standard notation for potential outcomes. We follow the notation in Morgan and Winship (2007, 2015) by denoting potential outcomes as Y_i^1 and Y_i^0, and treatment indicator as D. This notation is similar to the notation subsequently used by Angrist and Pischke (2008), although potential outcomes are defined as Y_{1i} and Y_{0i}. One difficulty with the latter notation is that it becomes cumbersome when dealing with longitudinal or clustered data. For example, the observed outcome for unit i in county c at time t would be Y_{ict}. The potential outcome if treated would be Y_{1ict}. Writing Y_{ict}^1 and Y_{ict} makes the difference between potential and realized outcomes easier to read. Other authors, for example, Imbens and Rubin (2015), denote potential outcomes as $Y_i(1), Y_i(0)$ and treatment as W. So Equation 2.1 becomes $Y_i = Y_i(1)W_i + (1 - W_i)Y_i(0)$. The longitudinal potential outcome if treated becomes $Y_{ict}(1)$. This notation becomes hard to read when equations are enclosed in multiple parentheses. Denoting treatment using T would be more intuitive, but T and t are often used to denote time.

not take it. Thus, conceptually, the case of using different people for comparison is similar to using multiple versions of the same unit at different times.

To compare outcomes between groups, we could contrast the expected value of the units that received the intervention to the expected value of the units that did not receive the intervention,

$$\delta_{NAIVE} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]. \tag{2.3}$$

It will become clear shortly why we call this difference "naive." But before, note that we could define this comparison in different ways. We are often interested in the expected value of random variables because the expected value gives us a summary of central tendency (see Section A.7.6 in the Appendix). Furthermore, most commonly used and oldest parametric and nonparametric regression models estimate conditional expectations like those in Equation 2.3 (see Section **??**). However, we could also be interested in comparing the median outcome, or we could be interested in comparing the entire *distribution* of Y among the two groups. Or, perhaps, like in Section 2.2, we would like to express the comparison in terms of relative risks. Different statistical models estimate parameters in different metrics or **scales**, and we can often go from one scale to another. This is the topic of Chapter 4 on marginal effects. However, the choice of summary measure is not always trivial since it could affect policymakers' decisions about the value of an intervention.

We can rewrite Equation 2.3 incorporating potential outcomes as

$$\delta_{NAIVE} = E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1] + E[Y_i^0|D_i = 1] - E[Y_i|D_i = 0]. \tag{2.4}$$

Note that we just used 2.1 to rewrite observed conditional expectations in terms of counterfactuals. The term $E[Y_i^1|D_i = 1]$ is equal to $E[Y_i|D_i = 1]$. In words, after an intervention is conducted or if data for the two groups are observed, the potential outcome for the treatment group had the treatment group been treated becomes the observed outcome for the treatment group because, well, the treatment group was treated (understanding Equation 2.1 is key for these derivations). We also added two terms in the middle that equals zero ($-E[Y_i^0|D_i = 1] + E[Y_i^0|D_i = 1] = 0$).

We can re-write 2.4 in a more convenient form, noticing that $E[Y_i^1|D_i = 1]-E[Y_i^0|D_i = 1]$ is the *definition* of the average causal effect for the *treated* group – the so-called **average treatment effect on the treated** (ATET). It is a comparison of potential outcomes conditional on receiving treatment. Therefore, we have

$$\delta_{NAIVE} = \delta_{ATET} + \underbrace{E[Y_i^0|D_i = 1] - E[Y_i|D_i = 0]}_{\text{assignment bias 1}}. \tag{2.5}$$

Equation 2.5 provides a lot of insight. The observed naive comparison will provide an estimate of the average treatment effect for the treated provided the **assignment bias** (commonly known as **selection bias**) is zero, which will be if $E[Y_i|D_i = 0] = E[Y_i^0|D_i = 1]$. In words, the assignment bias is zero if the observed average outcome for the control group is the same as the average outcome for the treated group had the treated group not been treated (a counterfactual). Thus, the assignment bias is zero if the experience of the control group (as measured by the expected value of the outcome) can be used to predict what would have happened to the treated group if they were not treated.

We can perform these derivations in a different way to obtain another expression for the assignment bias, one that tells us that the observed experience of the treatment group must be a good counterfactual for the control group. Starting from Equation 2.3, we rewrite the naive comparison as

$$\delta_{NAIVE} = E[Y_i|D_i = 1] + E[Y_i^1|D_i = 0] - E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0].$$

Once again, we re-arrange terms to obtain

$$\delta_{NAIVE} = \delta_{ATEC} + \underbrace{E[Y_i|D_i = 1] - E[Y_i^1|D_i = 0]}_{\text{assignment bias 2}}, \tag{2.6}$$

where $\delta_{ATEC} = E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0]$ is the definition of the **average treatment effect on the control** (ATEC). Thus, 2.6 is the the flip side of the previous insight: the assignment bias 2 is zero if the experience of the treatment group tells us what would have happened had the control units been treated.

We also ended up showing something else. If the two forms of assignment bias, 2.5 and 2.6, are zero, then it must be true that $\delta_{ATET} = \delta_{ATEC}$, and also that the naive comparison is the same as the **average treatment effect** (ATE). Therefore, when there is no assignment bias,

$$\delta_{NAIVE} = \delta_{ATE} = \delta_{ATET} = \delta_{ATEC}. \tag{2.7}$$

In other words, when the assignment biases are zero, our observed measure of association, δ_{NAIVE}, does imply a causal effect. Now *association is causation*.

Although the derivations can be confusing at first, they illuminate concepts that are difficult to explain without careful notation. Multiple units can help us predict counterfactuals, but we showed that a naive comparison of average observed outcomes **identifies** (see Box 2.4) causal treatment effects only when the experience of each group tells us what would have happened to the other group in the counterfactual scenarios. This is the insight we gain from reframing the fundamental problem of causal inference as a missing data problem. Equations 2.5 and 2.6 tell us under which conditions a naive comparison of observed outcomes provides accurate predictions of counterfactuals.

Intuitively, there is no assignment bias when treatment and control groups are on average identical in both observed an unobserved characteristics, which includes potential outcomes. The group of with high-deductible plans must be comparable to the group with no-deductible plans so they can provide a valid counterfactual prediction (and vice versa). It turns out that understanding the **assignment mechanism** –why some units ended up receiving treatment– is the most important consideration when trying to establish if a comparison of observed outcomes identifies causal treatment effects – if the naive comparison can provide causal effects.

One subtle issue is that to identify average causal treatment effects from a comparison of average observed outcomes, we need the symmetry implied by 2.5 and 2.6. Each group has to be a good counterfactual for the other. For example, suppose that treatment and control groups are identical (on average) on all unobserved and observed characteristics except for a genetic mutation that is more prevalent in the treatment group. This mutation only affects the response to treatment (say, it makes it more likely that the medication will alleviate the symptom). In this scenario, the assignment bias in 2.5 is zero because $E[Y_i|D_i = 0] = E[Y_i^0|D_i = 1]$. That is, the experience of the control, summarized by the expected value of the outcome in the control group, would be a good predictor of the counterfactual outcome if those in the treatment group had not taken the medication. However, the assignment bias in 2.6 is not zero, because if individuals in the control group had taken the medication, the control group would not have responded as strongly. Thus, in this example, $E[Y_i|D_i = 1] > E[Y_i^1|D_i = 0]$, and consequently, $\delta_{NAIVE} \neq \delta_{ATE}$. The naive treatment effect does not reflect the causal average effect if the medication were given to the entire population. However, we could identify δ_{ATET} in this example because the control gives us a good counterfactual for the treatment group.

2.5 The central importance of the assignment mechanism

The previous section previewed the central importance of understanding the assignment mechanism to determine if a comparison of observed outcomes can provide causal treatment effects. In this section, we review different types of assignment mechanisms: **simple randomization**, **conditional randomization**, and mechanisms that generate observational

Box 2.2 **Semantics: Identification**

The word **identification** is used often in econometrics and other fields. Lewbel (2019) traced the different uses of the word, and the many words used instead of identification. In the context of this chapter, identification refers to the idea of whether a quantity or parameter "picks up" or "pins down" the feature of interest. For example, we say that δ_{NAIVE} identifies average causal treatment effects if $\delta_{NAIVE} = \delta_{ATE}$. Lebel (2019) defines identification as "model parameters or features being uniquely determined from the observable population that generates the data." Or said another way, if we knew the population, what could we say about a quantity or parameter of interest? Keep in mind that we are not distinguishing sample from population. We can discuss the identification of causal effects at the population level.

data, which are not under the control of an investigator. With observational mechanisms (or data), units may **self-select** into treatment or the assignment may depend on many different factors that complicate the problem of establishing causal effects. Some of these factors might be unobservable.

2.5.1 Simple randomization

In simple randomization, a group of units is *randomly* divided into treatment and control groups. The treatment group receives an intervention and the control group does not. After the intervention takes place, we measure an outcome Y for each unit i. A naive comparison of average treatment effects identifies average treatment effects because both forms of assignment bias are zero.

The idea that simple randomization identifies average treatment effects is familiar to most readers, but it is important to understand why given the concepts we have covered so far. Provided sample sizes are large, simple randomization makes both forms of assignment bias zero because the assignment of units into treatment is completely unrelated to any measure or unmeasured characteristic of the units including *potential outcomes*. Treatment assignment was random after all.

A *consequence* of assigning units in a random way, say, by tossing a coin, although different assignment proportions could work, is that on average, both treatment and control groups are identical in all observed and unobserved characteristics. In fact, the entire distribution for any characteristic is identical. Thus, one group provides a counterfactual for the other group. It does not matter which group ends up receiving the treatment since both groups are **exchangeable**. Since causal inference is a prediction problem, simple randomization simplifies the prediction problem: the outcome of the control group is a good prediction of what would have happened to the treatment group had they not been treated (and vice versa).

Going back to our notation, one way to to express the ideas in the previous paragraphs is to note that under simple randomization

$$\delta_{ATE} = E[Y_i|D_1 = 1] - E[Y_i|D_1 = 0] = E[Y_i|D_1 = 1] - E[Y_i^0|D_1 = 1]. \tag{2.8}$$

In words, a comparison of average observed outcomes between treatment and control groups is the same as comparing the observed outcome of the treated group to its non-observed (counterfactual) $E[Y_i^0|D_1 = 1]$ because randomization ensured that $E[Y_i^0|D_1 = 1] = E[Y_i|D_1 = 0]$ – the observed average outcome of the control group gives us the counterfactual for the treated group had they not received treatment (remember that the same logic applies to the control group).

When discussing randomization, it is helpful to frame the consequences of randomization in terms of statistical **independence** (see Section A.6 in the Appendix for more examples). In terms of events, events A and B are independent if their **joint probability** is the same as the multiplication of their individual probability of occurring: $P(A \cap B) = P(A)P(B)$. However, it is more useful to think of independence as it applies to **conditional probability**. Independence implies that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A). \tag{2.9}$$

In words, the conditional probability of event A occurring given that event B has occurred is just the probability of event A occurring. Because events A and B are independent, knowing B does not provide any information about the probability of event A occurring. Another way of denoting that events A and B are independent is to write that $A \perp B$ (A is perpendicular or orthogonal to B).

We can now re-state the consequence of randomization. Randomization ensures that treatment assignment is *independent of potential outcomes*, which is the reason both forms of assignment or selection bias are zero. This is another way of framing the “comparability” of treatment and control groups. Recall that independence implies **mean independence**. (Two random variables Y and X are mean independent if and only if $E[Y|X] = E[Y]$; the mean of Y is the same regardless of the values of X.) Since randomization ensures that treated and control group assignment is independent of potential outcomes and any other factors, observed and unobserved characteristics in each group are mean independent of treatment assignment. For this reason, if we compare the mean of observed characteristics between treatment and control, they will be, on average, the same. Not only that, in fact, randomization ensures that unobserved characteristics should be the same between treatment and control groups. This is the information that would be displayed in “Table 1” of a paper reporting trial results, a comparison of observed characteristics between treatment and control groups in which the mean, standard deviation, range or any other measures would be similar in both groups. The smaller the sample size, the more likely that there could be small differences, but if the trial was properly conducted, all differences were due to chance. If treatment and control groups do not appear to be comparable, that would provide evidence that randomization was not conducted correctly.

2.5.2 Conditional randomization

As helpful as simple randomization is to understand under which circumstances we can obtain causal effects, there is another type of randomization that provides a more informative example, especially when we discuss observational data: **conditional** or **block randomization**. In conditional randomization, the values of one or more variables are used to assign treatment. For example, a medical treatment could be given to people *conditional* on the level of severity of the disease. Rather than randomly dividing people into treatment a control, an investigator could first determine the severity of the disease for each patient. For those with a more severe condition, patients are randomly divided intro treatment and control groups, with those with a more severe condition having, say, 80% chances of receiving treatment. For patients with a less severe condition, randomization is 50%-50%. This could be done for ethical reasons if an investigator believes that the treatment is effective. The investigator may want to give severely ill patients more chances to receive the treatment. Further, we assume that the outcome Y is positively associated with disease severity.

Note the implications of this design. Because treatment assignment depends on the level of disease severity (another way of saying *conditional* on disease severity), there will be more severely ill patients in the treatment group than in the control group. Therefore, the potential outcomes between treatment and control groups are not independent of group assignment. In terms of assignment bias, we would suspect that $E[Y_i|D_i = 0] \neq E[Y_i^0|D_i = 1]$ and/or $E[Y_i|D_i = 1] \neq E[Y_i^1|D_i = 0]$. In words, the experience of each group, as measured by the expected value of their outcomes, does not provide a good prediction of their counterfactual. As a consequence, $\delta_{NAIVE} \neq \delta_{ATE}$. Since disease severity implies worse outcomes, the naive comparison would be biased towards finding that the treatment does not work. It is likely, too, that if we compared observed characteristics between treatment and control groups, we would find that they differ in factors that are likely related to the outcome. For example, if disease severity increases with age, then there would be a larger proportion of older patients in the treatment group.

Fortunately, in the case of conditional randomization in which an investigator assigns treatment, there is a simple solution to obtain causal effects. Treatment assignment is not independent of potential outcomes, but treatment assignment is **conditional independent** of potential outcomes – once we condition for disease severity.

Before we review conditional independence, which is one of the most important concepts to master when it comes to the analysis of causal inference using observational data, we can tackle this example intuitively. We now know that a naive comparison of outcomes between treatment and control groups will not identify average treatment effects. But we also know that *within* (i.e., conditional or once we know disease severity) each severity level, patients were randomly assigned to treatment. It follows that we could stratify the naive comparison of outcomes to obtain average causal effects. In terms of assignment bias, and denoting disease severity by Z, we have

$$E[Y_i|D_i = 0, Z = 1] = E[Y_i^0|D_i = 1, Z = 1], \tag{2.10}$$

and

$$E[Y_i|D_i = 0, Z = 0] = E[Y_i^0|D_i = 1, Z = 0], \tag{2.11}$$

where $Z = 1$ if a patient is severely ill and $Z = 0$ if the patient is not severely ill. Thus, both forms of assignment bias are zero once we condition for disease severity.

We can define conditional independence more formally using events, as we did for independence (see Section A.6.3 in the Appendix). Events A and B are conditionally independent given Z if $P(A \cap B|Z) = P(A|Z)P(B|Z)$. As with independence, it is more helpful to use the implication of the definition. If events A and B are conditionally independent given Z, then $P(A|B, Z) = P(A|Z)$. So, given Z, knowing B does not provide information on the conditional probability $P(A|Z)$. We can write the same as $A \perp (B|Z)$.

An example will help us make all these concepts more concrete. We use data from a simulated experiment from a sample of the US natality files. Mothers were randomized based on smoking status before pregnancy (`smokedb`). If a mother smoked before pregnancy, she had a 90% chance of being assigned to treatment (variable `D`), such as receiving visits from nurses for education and help before and after birth, like the Nurse Family Partnership in the US. The outcome is baby's birthweight (`bwgrm_post`) measured in grams.

The table below shows baseline characteristics by treatment group. Remember that independence implies mean independence. Thus, if the data below had originated from simple randomization, we would have reasons to believe randomization failed. The treatment group is more likely to be White non-Hispanic, more likely to have completed high school, less likely to be married, more likely to receive supplemental assistance for food (the Special Supplemental Nutrition Program for Women, Infants, and Children, WIC), and, of course, much more likely to have smoked before pregnancy. On the other hand, there are small or no differences in the number of prenatal visits, age, and risk factors for pregnancy. Since we know that randomization was conditional on smoking before pregnancy, the variables that are different (not **balanced**) are the variables that tend to be associated with smoking.

```
use "https://www.perraillon.com/PLH/data/cr_example_sm.dta", clear
dtable mage i.meduc i.mracehisp bmi i.married i.smokedb i.noprenatal ///
              i.wic i.rf_any, by(D, tests notestnotes nototal) novarlabel
----------------------------------------------------------
                           1 if assigned to treatment
                             0              1        Test
----------------------------------------------------------
N                      6,108 (41.8%)  8,492 (58.2%)
mage                  28.877 (5.768) 28.525 (5.723) <0.001
meduc
  Some edu               711 (11.7%)  1,236 (14.6%) <0.001
  High school          1,543 (25.4%)  2,495 (29.5%)
  Some college         1,749 (28.8%)  2,501 (29.5%)
  College              1,289 (21.2%)  1,384 (16.4%)
  Master/doctorate       791 (13.0%)    848 (10.0%)
mracehisp
  White non-hisp       3,308 (54.2%)  5,119 (60.3%) <0.001
  Black non-hisp       1,050 (17.2%)  1,256 (14.8%)
  Asian non-hisp          314 (5.1%)     345 (4.1%)
  Other/mult non-hisp     164 (2.7%)     312 (3.7%)
  Hispanic             1,272 (20.8%)  1,460 (17.2%)
bmi                   27.174 (6.822) 27.182 (6.897)  0.941
married
  0                    2,353 (38.5%)  4,056 (47.8%) <0.001
  1                    3,755 (61.5%)  4,436 (52.2%)
smokedb
  0                    5,809 (95.1%)  5,807 (68.4%) <0.001
  1                       299 (4.9%)  2,685 (31.6%)
noprenatal
  0                    5,987 (98.0%)  8,254 (97.2%)  0.002
  1                       121 (2.0%)     238 (2.8%)
wic
  0                    4,004 (65.6%)  5,095 (60.0%) <0.001
  1                    2,104 (34.4%)  3,397 (40.0%)
rf_any
  0                    4,974 (81.4%)  6,898 (81.2%)  0.754
  1                    1,134 (18.6%)  1,594 (18.8%)
----------------------------------------------------------
```

Since we know that smoking and many of the factors in the table above are related to birthweight, we would suspect that the average difference in birthweight, the outcome, between treatment and control group would not identify average treatment affects. The comparison of average birthweight is $\delta_{NAIVE} = 60.7$ grams (p-value < 0.001). (The code for this chapter includes all calculations).

We will now compare baseline characteristics conditional on smoking before pregnancy. To save space, we only present the table for the group that smoked:

```
dtable mage i.meduc i.mracehisp bmi i.married i.smokedb i.noprenatal ///
            i.wic i.rf_any if smokedb==1, ///
            by(D, tests notestnotes nototal) novarlabel
---------------------------------------------------------
                           1 if assigned to treatment
                             0              1        Test
---------------------------------------------------------
N                         299 (10.0%)  2,685 (90.0%)
mage                  26.873 (5.497) 27.389 (5.457) 0.121
meduc
  Some edu                 72 (24.1%)    587 (21.9%) 0.777
  High school             123 (41.1%)  1,075 (40.1%)
  Some college             91 (30.4%)    866 (32.3%)
  College                   11 (3.7%)     125 (4.7%)
  Master/doctorate           2 (0.7%)      26 (1.0%)
mracehisp
  White non-hisp          225 (75.3%)  1,996 (74.3%) 0.732
  Black non-hisp           37 (12.4%)    321 (12.0%)
  Asian non-hisp             0 (0.0%)      16 (0.6%)
  Other/mult non-hisp       15 (5.0%)     150 (5.6%)
  Hispanic                  22 (7.4%)     202 (7.5%)
bmi                   27.256 (6.829) 27.351 (7.304) 0.829
married
  0                       219 (73.2%)  1,921 (71.5%) 0.536
  1                        80 (26.8%)    764 (28.5%)
noprenatal
  0                       287 (96.0%)  2,556 (95.2%) 0.541
  1                        12 (4.0%)  129 (4.8%)
wic
  0                      131 (43.8%)  1,210 (45.1%) 0.680
  1                      168 (56.2%)  1,475 (54.9%)
rf_any
  0                      231 (77.3%)  2,146 (79.9%) 0.277
  1                       68 (22.7%)    539 (20.1%)
--------------------------------------------------------
```

The table above provides strong evidence that we have mean independence, and also, in the case of education and race, that the distribution of characteristics is the same in both groups. Since we know how randomization was conducted (simulated in this case), we know that conditioning on smoking gives us conditional independence of treatment and potential outcomes. One way to think about this is to imagine that you know the marital status of a mother (married or not). Since we know that there are more married women in the control group (first table), we know that it is more likely that this mother is in the control group. However, once we condition on smoking, knowing marital status does not tell us anything about the probability that a mother belongs to either treatment or control group because there are the same proportion of married women within each group given smoking status.

Now, for each smoking status group, a comparison of average outcomes identifies causal effects. The treatment effect for the group that smoked is $\delta_{ATE,smoked} = 131.3$ (p-value = 0.001) and the treatment effect for the group that did not smoke is $\delta_{ATE,notsmoked} = 102.0$ (p-value < 0.001). We could then confidently conclude (although we need other assumptions that we discuss in Section 2.8) that the trial worked regardless of smoking status. Of course, statistical significance is not all that matters. We would need to establish if a difference of about 100 grams is a meaningful change.

In Chapter **??**, we emphasized that parametric regression models always imply functional form assumptions. Since we know that we can identify average treatment effects by conditioning on smoking, we could have estimated the linear regression model

$$bwgrm\_post_i = \beta_0 + \beta_1 D_i + \beta_2 smokedb_i + \epsilon_i. \tag{2.12}$$

However, Model 2.12 *assumes* that the treatment effect does not depend on smoking status, but we just saw, using stratification, that the experiment showed a larger effect for mothers who smoked. In Model 2.12, $\hat{\beta}_1$ is a weighted average of the treatment effect in each group. Since we do not include other covariates, the treatment effects are weighted by sample sizes in each smoking status group.

The model that would capture **heterogeneous treatment effects** by smoking status would be:

$$bwgrm\_post_i = \alpha_0 + \alpha_1 D_i + \alpha_2 smokedb_i + \alpha_3 (D \times smokedb)_i + \eta_i. \tag{2.13}$$

The advantage of Model 2.13 is that we could also test for interactions. As it turns out, we do not reject the null $H_0 : \alpha_3 = 0$, so we could present a single treatment effect rather than stratified treatment effects. The combined treatment effect is 104.5 (p-value

< 0.001). Since most mothers are non-smokers, it makes sense that the combined treatment effect is closer to $\delta_{ATE,notsmoked} = 102.0$ grams. Because we can use regression models that condition for covariates, we can think of regression models as one of the oldest methods to estimate causal effects.

Incidentally, the R^2 of the regression model is 0.014. Thus, the model explains only 1.4% of the birthweight variance. In Chapter **??**, Section **??**, we explained the meaning of R^2 and emphasized that basing modeling decisions on R^2 can be misleading. This is another example. A low R^2 tells us that many other factors can explain and predict the variance of birthweight, but Model 2.13 correctly captures the effect of the treatment regardless of the value of R^2.

There are other ways to analyze this experiment. To improve the precision of the estimates, we could have added other covariates to the model, even though the estimates of treatment effect would likely not change by much because of conditional independence (review Section **??**). Or we could have used Inverse Propensity Score Weighting (IPW) (Chapter 6) because this method has some advantages in terms of relaxing functional form assumptions (i.e., **doubly-robust**). In Section 2.9, we will use a system of equations to estimate treatment effects because this method makes it explicit that regression is actually predicting counterfactual outcomes. If the outcome of interest would have been the probability of low birthweight, then we would have used logit or probit regression models for binary outcomes. What is important to keep in mind, however, is that we discussed the assumptions needed to identify causal effects *without needing to discuss how to estimate causal effects*.

2.6 Ignorability or the the conditional independence assumption

We can now more formally state one of the most important assumptions to establish causal effects, the **conditional independence assumption** (CIA) or **ignorability**: conditional on observable covariates $\boldsymbol{X}$, the assignment of units into experimental groups is independent of potential outcomes:

$$(Y_i^0, Y_i^1) \perp D_i | \boldsymbol{X}_i. \tag{2.14}$$

In the conditional randomization examples, we satisfied this assumption by conditioning on disease severity or smoking status. In simple randomization, we do not need to condition for any covariate $\boldsymbol{X}$.

One key issue that we need to determine is the set of covariates $\boldsymbol{X}$ that we need to condition on to obtain causal effects. The discussion of conditional randomization makes it clear that we need to condition for variables that are associated with treatment assignment and are related to the outcome. In other words, all *confounders*. Recall the definition of a confounder. A confounder of the relationship between Y and D is a variable X that is both associated (in any functional form) with Y and D. This condition is necessary but not

Box 2.3 Semantics: Conditional independence assumption

This assumption comes in may different names depending on the field. The most common in statistics is **ignorability** of treatment assignment (once we condition for X). In other fields, other names are **no unmeasured confounders** or (conditional) **exchangeability**. In econometrics, **selection on observables** is common. All these names tell us about a feature of the assumption and different ways of understanding the same idea. Be careful distinguishing conditional independence versus the conditional independence *assumption* (CIA).

sufficient. We also need a conceptual framework to determine if a variable is a confounder or a **mediator** plus other logical restrictions like temporality (Section 2.12.4). Thus, not all variables that are associated with Y and D are confounders. For this reason, "selection on observables" (see Box 2.6) can be somewhat deceiving. Not all the observed variables that determine selection (assignment) into treatment are relevant. We only care about the variables that determine selection and are also associated with the outcome of interest. Keep in mind an idea that is often forgotten: even within the context of the same data and research question, confounders for one outcome may not be confounders for another outcome.

Another idea that is important to keep in mind is the implication of CIA for *unobserved* confounders. If our conceptual framework tells us that to obtain causal effects we must condition for a factor that is not possible to observe or is not present in a dataset, then we cannot establish causal effects unless we use methods that do not rely on observing all confounders, such as the methods that we cover in Part IV this book.

2.7 Observational data

The preceding discussion highlights the importance of understanding (and, if possible, manipulating!) treatment assignment, and the importance of conditional independence and conditioning when assessing the ability to identify and estimate causal effects. Conditional randomization provides a useful mental model to discuss observational data because we can think of observational data as data resulting from uncontrolled "experiments" with a very complicated treatment assignment; designs in which many observable and unobservable factors could determine which units are treated.

Consider again one of the examples we used at the start of the chapter. Suppose that we are again interested in understanding the role of providing insurance with no-deductibles versus insurance with high deductibles. However, we do not conduct an experiment but instead use observational data such as a nationally representative survey that has information on insurance status, deductible level, and outcomes (measures of financial strain and healthcare utilization) for those younger than 65 (most adults 65 and older in the US obtain Medicare plans with no deductibles). Would a naive comparison of financial strain or healthcare utilization identify causal effects?

The key to making progress in answering the previous question is to reframe the question in terms of treatment assignment. Why do some people have low versus high-deductible plans? Are the factors associated with deductibles also related to financial strain or healthcare utilization? In the US, some people younger than 65 *choose* high deductible plans because they are healthy and do not expect high utilization or prefer not to go the doctor (the so-called, facetiously, "young invincibles"). Plans with high deductibles are attractive to this group because high-deductible plans have lower premiums. Other people may not have a choice because their employers only offer high-deductible plans, which is becoming increasingly common. Some plans with high deductibles tend to have more choices in terms of providers (i.e., wider networks), so it could be that people with chronic conditions or people who have a preference for more care choose high-deductible plans. Medicaid, the insurance for low-income individuals, has no deductibles. Thus, many factors are related to both treatment and outcomes. Furthermore, following the logic of conditional randomization, we cannot observe, and thus cannot condition for, all the factors that determine treatment assignment, such as preferences and planned utilization of healthcare services. We could include observable factors that could act as **proxy variables**, such as income and comorbid conditions, but still lack information on preferences. Furthermore, proxies are not perfect substitutes. As a consequence, we would need another strategy to estimate the causal effect of deductibles because we cannot condition for all the relevant covariates.

Consider another example that is common when studying the effects of health policy. In 2009, the Centers for Medicare and Medicaid (CMS) released 5-star ratings for nursing homes. All nursing homes that receive Medicare payments (over 94% of nursing homes in the country) were given a rating, so it would not be possible to establish the causal effect of receiving a rating since all nursing homes received one. If we wanted to study the impact of receiving a low versus high rating, a naive comparison of outcomes – for example, new admissions 6 months after the ratings were released – would be problematic since treatment assignment is not independent of potential outcomes: previous knowledge suggests that nursing homes size is associated with quality and ratings, and so are many factors that determine the choice of a nursing home by patients and their families, which are largely unobservable and are related to new admissions. As in this example, health policy changes usually apply to all units.

2.7.1 Strong ignorability

We will discuss in more detail issues of **overlap** and **balance** in Chapter 6 on propensity scores. Informally, overlap, sometimes referred to as **common support** (or **positivity**), is the degree to which the values of a variable share the same range of values between treatment and control groups. Under simple and conditional randomization, we saw that the distribution of covariates, not just the mean, is the same between treatment and control groups – in the case of conditional randomization, once we condition for some variables. Thus, there is no overlap problem. Furthermore, there is no balance problem either since independence and conditional independence imply mean (conditional) independence. Even more, they imply that the entire distribution of covariates should be the same.

With observational data, overlap can be a problem. For example, the treatment group consists of older patients while the control group consists of young patients. Thus, implicitly, a comparison of outcomes between treatment and control groups would involve extrapolating information from the young to the old and vice versa.

More formally, overlap requires that for all covariates $\boldsymbol{X}_i \in \omega$, where ω is the support (domain) of the covariates,

$$0 < Pr(D_i = 1|\boldsymbol{X}_i) < 1. \tag{2.15}$$

In words, for all the values of every covariate, the probability that a unit i is treated cannot be 0 or 1. In the example of older patients being only in the control group, the probability that a person who is, say, 70 years old is in the control group would be 1. Thus, $Pr(D_i = 1|age = 70) = 1$, violating overlap.

Strong ignorability says that to obtain causal effects with observational data, we must have conditional independence *and* overlap. Again, we only worry about overlap with observational data. In expression 2.15, $Pr(D_i = 1|\boldsymbol{X}_i)$ is the *definition* of the **propensity score**, the propensity to belong to the treatment group conditional on covariates. In Chapter 6, we will use the propensity score to both diagnose overlap problems and come up with solutions to the problem, assuming we also have CIA/ignorability.

2.8 Stable Unit of Treatment Values Assumption (SUTVA) and exclusion restrictions

Ignorability and strong ignorability are necessary conditions to establish causal effects but are not sufficient. We made some implicit assumptions when discussing some examples that we clarify in this section. In general, there are assumptions that can be verified, or at least can be partially checked, with data and assumptions that cannot be verified with data, often called **exclusion restrictions**. One of them is the Stable Unit of Treatment Values assumption (SUTVA).

SUTVA refers to two ideas: a) **no treatment interference** (also known as **no spillovers**) and b) **no hidden variation in treatment**. No treatment interference states that the potential outcomes for any unit do not vary with the treatment to other units. This assumption can easily be violated. For example, if treatment consists of providing nurse visits to pregnant young women, other pregnant young women who do not receive the visit (treatment) could communicate with the ones who do. In the US, when Medicare announces a new change in policy such as a new benefit or requirement for providers, Medicare patients are not the only ones to be affected by the policy because insurance companies and providers tend to follow Medicare policy. Thus, finding good control units could be problematic.

No hidden variation in treatment refers to the requirement that units should receive the same treatment. In the case of high-deductible plans, for example, many health plans vary not only by the level of deductibles but also by the providers they cover and many other subtle differences in plans that make them different from each other – in other words, the “treatment” is not just different level of deductibles. **Double-blind** trials, in which both patients and investigators do not know who receives the treatment, are considered one of the strongest designs because it ensures that treatment is not affected by subtle expectations or changes in behavior by patients or researchers. There is plenty of evidence that placebo, and its opposite, nocebo, and expectation effects are powerful. If we think a treatment will help us, the treatment tends to be more effective – at least in the short run.

Issues with violations of SUTVA are about the design of the study and relate to potential problems in the conclusions that could be drawn from the estimation of effects. Potential violations of SUTVA should always be in our minds when evaluating the design and interpretation of studies.

2.9 Estimation of treatment effects by explicitly predicting counterfactuals

We have covered so far the conditions needed to establish causal effects under different assignment mechanisms. In Section 2.5.2, we used stratification and linear regression to estimate causal effects under conditional randomization. In this section, we show another way to estimate treatment effects that *explicitly* follows the idea that causal inference requires predicting counterfactual outcomes. Linear regression *implicitly* uses predictions, as we will further discuss in Chapter 6 on propensity scores. Explicit use of predictions to estimate causal effects has many didactic advantages, and we will more clearly estimate ATET and ATEC.

2.9.1 Average Treatment Effect (ATE)

We will estimate ATE using a series of steps:

1. Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 1]$ with a linear/OLS model using only (i.e., conditional on) treated observations
2. Using estimates from 1), predict $\hat{Y}_{Ti}$ in the *entire sample*. For the control units, $\hat{Y}_{Ti}$ is a prediction/imputation of their missing conterfactual
3. Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 0]$ (using only control observations)
4. Using estimates from 3), predict $\hat{Y}_{Ci}$ in the *entire sample*. For the treated units, $\hat{Y}_{Ci}$ is a prediction/imputation
5. The difference (contrast) between $E[\hat{Y}_{Ti}] - E[\hat{Y}_{Ci}]$ is the ATE

Note the logic of the preceding steps. We use the information of the treated group to make predictions about the control group, and then we use the information of the control group to make predictions about the treatment group. Once we have predictions, the estimate of treatment effects is the difference of two means (step 5). For this reason, we could call this method of estimating causal effects a **non-parametric** or **semi-parametric** approach since, contrary to Model 2.13, the estimate of treatment effect is not based on model parameters.

The code below estimates the treatment effect using the previous example of conditional randomization.

```
quietly {
   * Steps 1 and 2
   regress bwgrm_post smokedb if D ==1
   predict double yhat_t
   * Steps 3 and 4
   regress bwgrm_post smokedb if D ==0
   predict double yhat_c
   *Step 5
   summarize yhat_t
   local pom_t = r(mean)
   summarize yhat_c
   local pom_c = r(mean)
}
   display "ATE: " ‘pom_t’ - ‘pom_c’
ATE: 107.97642
```

The ATE is 108. Note that by default Stata makes predictions using the entire sample, not just the sample used to estimate the model (we omitted `e(sample)` when making predictions). Also, recall the algebraic properties of the linear model. On average, we do not make prediction mistakes, so in step 2, for example, the average predictions in the treatment group would be the same as the average of the observed outcome for the treatment group. The predictions for the other group are all counterfactual predictions.

There is another issue to keep in mind. If we wanted to obtain a standard error for the treatment effect and do hypothesis testing, we should take into account that we estimated several models. Stata implemented this way of estimating treatment effects with the command `teffects ra`, which uses the Generalized Method of Moments (GMM) method to estimate a system of equations (Chapter **??**, Section **??**):

```
teffects ra (bwgrm_post smokedb) (D), ate
Iteration 0:   EE criterion =  2.118e-25
Iteration 1:   EE criterion =  5.692e-26
Treatment-effects estimation                    Number of obs     =     14,600
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
------------------------------------------------------------------------------
             |               Robust
  bwgrm_post | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATE          |
           D |
   (1 vs 0)  |   107.9764   11.58283     9.32   0.000     85.27448    130.6784
-------------+----------------------------------------------------------------
POmean       |
           D |
         0   |   3240.606   9.549211   339.36   0.000     3221.889    3259.322
------------------------------------------------------------------------------
```

Hopefully, you are wondering why this estimate is different from the one we obtained with linear regression in Section 2.5.2. The reason is that we now estimate fully-interacted models because we stratified by treatment group (review Section **??** of Chapter **??**). To replicate the same treatment effects, we need to estimate a fully interacted model and then obtain an effect that averages over the treatment heterogeneity:

```
margins, dydx(D)
Average marginal effects                                 Number of obs = 14,600
Model VCE: Robust
Expression: Linear prediction, predict()
dy/dx wrt:  1.D
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         1.D |   107.9764   11.58401     9.32   0.000      85.2703    130.6825
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

The connection will become clearer after reviewing marginal effects in Chapter 4. For now, the main point to remember from this example is that linear regression *implicitly* uses predictions of counterfactuals. When using observational data, overlap becomes important because we could use units with different characteristics to make predictions. The other point to remember is that parametric models always imply functional form assumptions.

2.9.2 Average Treatment Effect on the Treated (ATET)

Using explicit predictions of counterfactuals will help us estimate ATET in an intuitive way. The steps are:

1. Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 1]$ with a linear/OLS model using only treated observations.
2. Using estimates from 1), predict $\hat{Y}_{Ti}$ *only using the treated sample*. Thus, no prediction/imputation to control units.
3. Estimate $E[Y_i|\boldsymbol{X}_i, D_i = 0]$ with a linear/OLS model using only control observations.
4. Using estimates from 3), predict $\hat{Y}_{CTi}$ using only the *treated* sample. Thus, this is the counterfactual for the *treated* group only
5. The difference (contrast) between $E[\hat{Y}_{Ti}]$ and $E[\hat{Y}_{CTi}]$ is the ATET.

Keep in mind that we only use the control observations to provide information about the counterfactual for the treatment group, but we are only interested in treatment effects for the treated group. Also, note that steps 1 and 2 are not strictly necessary because of the algebraic properties. We could have just calculated the average of observed outcomes for the treatment group.

```
quietly {
   * Steps 1 and 2
   regress bwgrm_post smokedb if D== 1
   predict double yhat_t1 if D==1 & e(sample)
   * Not needed since same as
   *summarize bwgrm_post if D==1
   * Steps 3 and 4
   regress bwgrm_post smokedb if D== 0
   predict double yhat_t11 if D == 1
   *Step 5
   summarize yhat_t1
   local pom_t1 = r(mean)
   summarize yhat_t11
   local pom_t11 = r(mean)
}
   display "ATET: " ‘pom_t1’ - ‘pom_t11’
ATT:  111.24741
```

Using `teffects ra`:

```
teffects ra (bwgrm_post smokedb) (D), atet
<--- output omitted --->
------------------------------------------------------------------------------
             |               Robust
  bwgrm_post | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
           D |
   (1 vs 0)  |   111.2474   14.09677     7.89   0.000     83.61824    138.8766
-------------+----------------------------------------------------------------
POmean       |
           D |
          0  |   3219.453   12.56905   256.14   0.000     3194.819    3244.088
------------------------------------------------------------------------------
```

The connection with linear regression is the same as with ATE, but now we want treatment effects for the treatment group while using the control group to make counterfactual predictions. We again use a fully-interacted model to match the `teffects ra` specification. The option `subpop()` in marginal effects restricts predictions to a subset of the data:

```
quietly regress bwgrm_post i.D##i.smokedb, robust
margins, dydx(D) subpop(D)
<--- output omitted--->
Expression: Linear prediction, predict()
dy/dx wrt:  1.D
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           D |
          0  |          0  (empty)
          1  |   111.2474   14.09793     7.89   0.000     83.61368    138.8811
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Using the same logic, we could estimate the treatment effect for the control group (ATEC). Once gain, the important part is to always consider that when we estimate regression models to obtain causal effects, we are implicitly making predictions from one group to the other.

2.10 The importance of conceptual frameworks and the data generating process

In the previous sections, we established the conditions needed to identify treatment effects. We discussed the assumptions in terms of the assignment mechanism. To understand the assignment mechanism, we need good knowledge of the **data generating process** and how variables are related to each other, which we call the conceptual or theoretical framework. The data-generating process is the stochastic (i.e, random or probabilistic) process that results in the data we observe. We can think of the conceptual or theoretical framework as the knowledge that allows us to define relationships given a research question. For example, we know that mother's age is associated with birthweight (conceptual framework), but in the example of conditional randomization, we know that the research design made mother's age irrelevant once we condition on smoking (data generating process). In other words, we know that mother's age is conditionally independent of treatment assignment given smoking status.

Theoretical frameworks can be described in different ways. Causality has always been central in econometrics because economists use detailed theories to describe behavior. Before (approximately) 1950, economic theories used mostly language to explain relationships, but now mathematical models are the norm. For example, consumers are assumed to maximize utility subject to budget constraints. The result of the utility maximization process is a demand curve, which determines the quantity that would be demanded at given prices – interestingly, they can be thought of as counterfactuals. The other side of the market is the producers. Producers are assumed to maximize profits subject to input budget constraints. The result is a supply curve that determines the quantity that would be demanded at given prices. The intersection of supply and demand curves sets the market equilibrium price and quantity sold. If we collected data on prices and quantities for a product and estimated the relationship between prices and quantities, the detailed theoretical framework tells us that the observed data does not allow us to identify a demand or supply curve. Instrumental variables (Chapter 9) were developed about a century ago to solve this identification problem. It is a strong understating of the data-generating process and a conceptual framework that allows us to establish causality.

2.10.1 Directed Acyclic Graphs

Theoretical frameworks and the data-generating processes do not need to be described using mathematical models or words. A popular way to depict relationships in the causal inference literature in epidemiology and statistics is to use Directed Acyclic Graphs (DAGs). DAGs are graphical displays of the data-generating process and conceptual framework for a particular causal problem (and data) that include a depiction of the *causal* relationships between variables. Variables or set of variables are represented by nodes and relationship by directed lines or edges. DAGs are also used to characterize the conditions under which causal inference is possible, and, in some cases, a way to solve the identification problem.

We do not use DAGs in this book because they are not commonly used in health services research and health economics, but also because using DAGs would involve introducing many more terms to describe the same concepts covered in this chapter. We find DAGs helpful in some situations, although they can also be difficult to work with when dealing with longitudinal data. See Section 2.14 for resources on DAGs.

2.11 Brief overview of methods for causal effects with observational data

Part IV of this book covers methods to estimate causal effects when ignorability/CIA does not hold. That is, methods that do not rely on conditioning for confounders, which implies that we are not able to observe or measure key factors that determine treatment assignment and are associated with outcomes. These methods are often referred to as **natural experiments** of **quasi-experimental** designs.

One way to unify these methods is that they all exploit our understating of the assignment mechanism – a kind of "golden rule" of causal inference: if we know key aspects of treatment assignment, we could find ways to estimate treatment effects. In **difference-in-difference** designs (Chapter 8), we know that treatment (often a policy change) assignment occurred at some point in time t_0. We use data before and after t_0 for a treatment and control group to remove factors that are time-invariant within each group (treatment and control) and factors that do not change over time between groups. We do not require ignorability/CIA. In fact, treatment and control groups could be different, but the key is that this difference remains constant over time – the **constant bias** or **parallel trends** assumption. If there are factors that evolve differently over time (i.e., time-varying) between treatment and control groups and are related to the treatment, we must observe and condition for them.

With an **instrumental variables** approach (Chapter 9), causal effects can be identified by using a variable (or set of variables) that is a strong predictor of treatment assignment but is not (conditionally) related to the outcome. The variable that is a strong predictor of treatment is the **instrument**. In a sense, the instrument acts like a pseudo-randomizer of treatment, *inducing* a group of units to receive treatment. As a consequence, we can only identify a treatment effect for the units that were affected by the instrument. That is, the estimated treatment effect is a local average treatment effect (discussed in the next section). An ideal instrument would be a factor completely unrelated to potential outcomes that happens to induce treatment – an "exogenous shock."

In **regression discontinuity** designs (Chapter **??**), we take advantage of knowing (and observing) that a particular value of a continuous variable was used to assign treatment. The continuous variable used to assign treatment is called the **assignment** or **running variable**, and the particular value used to assign treatment is the cutoff point c. For example, all units with a value equal or greater than c receive the treatment while the rest are controls. Often, the assignment variable is associated with the outcome of interest, so we must condition for this variable. The key in regression discontinuity is that units with values greater and lower than c do *not* satisfy the assumption of ignorability/CIA. However, units very close to c, under some conditions, do satisfy ignorability/CIA. Thus, the identification of causal effects is at the limit, when the assignment variable approaches c.

All the previous methods require additional assumptions. Some of these assumptions can be verified, or can be partially verified, with observed data and some assumptions cannot – they are **exclusion restrictions**. Note that we do not include propensity scores in Part IV of the book. Propensity score methods require ignorability/CIA. However, propensity scores do help solve causality problems when there is lack of overlap, which is an issue related to observational data in which the assignment mechanism is more complicated. Also, keep in mind that some of these methods can be combined and there are many variations for each, which we discuss in the corresponding chapters.

Finally, we mentioned that a unit at different points in time is considered a different unit. However, observing a unit at different points in time can help with causality because at least we can discard factors that remain constant over time. Difference-in-difference designs in part take advantage of **longitudinal** or **panel data** as we will see in Chapter 8. Thus, in some sense, longitudinal data partially help us deal with causality problems.

2.12 Additional topics

The following sections introduce topics that are important to keep in mind but do not flow directly from other sections in this chapter.

2.12.1 Other causal estimands and alternative definitions

We have defined and discussed the estimation of ATE, ATET, and ATEC, but there are other types of causal effects or estimands. An **estimand** is a quantity to be estimated or identified, and it is useful to define a new word because we often need to invoke conditions under which an estimate (quantity) identifies a particular estimand of interest. For example, since we are often interested in treatment heterogeneity, we could define Conditional Average Treatment Effects (CATE):

$$\delta_{CATE} = E[Y_i^1|\boldsymbol{X}_i] - E[Y_i^0|\boldsymbol{X}_i]. \tag{2.16}$$

That is, CATE is a comparison of potential outcomes conditional on the particular value of one or more variables. As an example, we often want to estimate the treatment effect conditional on biological sex. Recall that when we write something like $E[Y_i|\boldsymbol{X}_i]$, we always mean $E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}_0]$. That is, conditional on specific values $\boldsymbol{x}_0$ of a variable or set of variables.

In subsequent chapters, we will introduce the Local Average Treatment Effect (LATE). As with CATE, LATE are treatment effects that apply to a particular population, although with some methods, it can be difficult to precisely define this population.

Finally, we define ATE, ATET, and ATEC again using different notations because these causal effects are presented in different ways by different authors.

The average treatment effect includes all units:

$$ATE = E[Y_i^1] - E[Y_i^0] = \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0 = \frac{1}{N}\sum_{i=1}^{N}(Y_i^1 - Y_i^0) = E[Y_i^1 - Y_i^0]. \quad (2.17)$$

The average treatment effect on the treated includes only treated units:

$$\begin{aligned} ATET &= E[Y_i^1|D_i = 1] - E[Y_i^1|D_i = 1] = \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i Y_i^1 - \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i Y_i^0 \\ &= \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i(Y_i^1 - Y_i^0) = E[Y_i^1 - Y_i^0|D_i = 1]. \end{aligned} \quad (2.18)$$

Finally, the definition of average treatment effect on the controls includes only controls:

$$\begin{aligned} ATEC &= E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0] \\ &= \frac{1}{\sum_{i=1}^{N}(1 - D_i)}\sum_{i=1}^{N}(1 - D_i)Y_i^1 - \frac{1}{\sum_{i=1}^{N}(1 - D_i)}\sum_{i=1}^{N}(1 - D_i)Y_i^0 \\ &= E[Y_i^1 - Y_i^0|D_i = 0]. \end{aligned} \quad (2.19)$$

The above expressions might look somewhat esoteric, but they are straightforward once you realize that $\sum_{i=1}^{N} D_i$ is the number of treated units and $\sum_{i=1}^{N}(1 - D_i)$ is the number of controls. For the treated group, $D_i Y_i^1 = Y_i^1$ and $D_i Y_i^1 = 0$ for the controls. $(1 - D_i)$ is 0 for the treated and 1 for the controls.

Keep in mind that we still need counterfactuals for ATET and ATEC when *estimating* treatment effects, as the examples in Section 2.9 showed.

2.12.2 Internal versus external validity

Two concepts that are often used and are important to keep in mind when interpreting study results are internal and external validity. **Internal validity** refers to the extent to which a study can provide causal effects. Thus, we could say that in this chapter we have established the assumptions needed to determine the internal validity of a research design or study.

External validity refers to the extent to which study results can be generalized to other populations or situations. For example, it is often said that randomized clinical trials have strong internal validity but often have poor external validity. They have strong internal validity because, as we have discussed, simple and conditional randomization results in zero assignment biases. Clinical trials tend to have poor external validity because the characteristics of patients enrolled in trials are often different than those of the general population. In part, this is because of enrollment requirements. In trials of antidepressant medications, for example, patients are required to be diagnosed with major clinical depression but may not have other psychiatric disorders. They could also be required to not have other medical diagnoses. In the past, women were excluded from clinical trials because it was thought that hormonal variations would complicate the interpretation of findings. (In the US, the National Institutes of Health requires the inclusion of women in trials, or a strong justification for not including women.) Few trials include children. The problem is, of course, that once a medication is approved, it will be used by the general population. Thus, there is always uncertainty as to whether the medication will work in the same way as in the clinical trial.

Similar external validity concerns apply to LATE. Sometimes, it can be easy to argue that LATE probably does not apply to an entire population, but in some cases it is difficult to determine the population to which LATE applies. We will discuss these issues when covering propensity scores, instrumental variables, and regression discontinuity, but it is important to always keep in mind potential external validity threats when interpreting study findings.

2.12.3 Average treatment effects versus unit-level treatment effects

To estimate treatment effects, we use multiple units to predict counterfactuals. As a consequence, we estimate *average* causal effects, but we do not estimate unit-level treatment effects because a single unit is either treated or controlled but never both. Our best guess is to assume that everybody in the population will experience the average treatment effect. Unfortunately, we know that this is not the case in most situations. Treatment heterogeneity is the norm rather than the exception. We know that medications affect people in different ways. We know that a viral infection can be deadly for some and a minor annoyance for others. The same health policy may be beneficial for some but detrimental for others – the so-called **unintended consequences** of policy changes. Even by applying a treatment to the same unit repeatedly, it is difficult to keep all factors constant, although it can help establish cause and effect.

We might be able to estimate conditional average treatment effects, for example, by race or smoking status, but we know that people of the same race and smoking status may experience a treatment in different ways. The concept of **personalized** or **precision medicine** is in essence an attempt to estimate heterogeneous treatment effects conditional on observed patient characteristics. Although we do not cover new methods attempting to estimate individual treatment effects – often under strong assumptions, it is an area of ongoing development.

2.12.4 Mediator, moderator, and “bad control” variables

Two concepts that are easy to confuse because the common dictionary definitions of the words do not help us much in distinguishing them are treatment moderation and mediation. We already covered moderation when discussing treatment heterogeneity and conditional average treatment effects. In general, a **moderator** is a variable that can change the relationship between two other variables. In the context of causal inference, a moderator is a variable that can change the treatment effect. In other words, the treatment effect depends on the value of another variable. When using regression models to estimate causal effects, we can add an interaction term between treatment and another variable to estimate effect moderation, as we did in Model 2.13. Another strategy would be to stratify the analysis, if possible. Using an approach like `teffects` implies obtaining a treatment effect that averages over all the heterogeneity.

On the other hand, a variable is a **mediator** if it explains or is a consequence of treatment. It can be part of the mechanism that explains causal effects, and in general, occurs after or at the same time as the treatment. It is often said that a mediator is in the "causal pathway" between treatment and outcome. For example, we now know that smoking causes lung cancer. A mediator of the relationship between smoking and lung cancer is the amount of carcinogens that enter the body. It is the main *mechanism* by which smoking causes lung cancer.

Note that both mediators and moderators appear similar to confounders. Both mediators and moderators are variables that are likely associated (in any functional form) with the treatment and the outcome. If we were to condition for a moderator or mediator in a regression model, it is likely that the treatment effect estimate would change. Therefore, we need a conceptual framework or theory to distinguish confounders from mediators and moderators.

Several estimation issues arise with mediators. In general, we should not condition for mediators when estimating causal effects. For example, if we were to estimate the impact of smoking on the probability of lung cancer, including a measure of carcinogens would change the estimate of treatment effects, but we would not be able to separate the effect of smoking from the effect of carcinogens. In other others, the contribution of smoking to cancer is not solely due to the level of carcinogens. In general, mediators are "bad control" variables, and so are variables that are a consequence of treatment. In difference-in-difference designs, for example, we cannot include in our regression models factors that occurred after a policy change and are a consequence of the change in policy.

Understanding mediation effects, however, is of great scientific importance. We are interested in understanding cause and effect, but we are also interested in understanding the mechanisms by which X causes Y. Mediation analysis is an area that has seen important advances recently. A goal of mediation analysis is to decompose a treatment effect into parts that are explained by different factors (**direct** and **indirect effects**). We recommend Vanderweele (2015) for a comprehensive introduction.

2.12.5 Multiple causes, moderation, and mediation

In this chapter, we discussed cause and effect as our understanding of a variable causing another. Does X causes Y? We have a tendency to seek "one" cause for a problem. Medical doctors often attempt to find the one disease that explains all symptoms. However, cause and effect can take more difficult forms, and this posses a challenge when designing and interpreting studies.

As an example, think about COVID-19 infections. In what sense can we say that COVID causes Long COVID when most people that are infected with the SARS-CoV-2 virus do not develop Long COVID?

If we use counterfactuals and some of the concepts we covered in this chapter, we can clarify this issue. We could say that COVID causes Long COVID because if a person had not had COVID, then they would not have Long COVID. However, COVID is necessary but not sufficient because there must be an interaction of COVID with one or more factors that result in Long COVID.

The challenge is complex because we also need to develop a strong biological theoretical framework to distinguish factors that are moderators and factors that are mediators of the relationship between COVID and Long COVID.

It is important to always keep in mind interactions, mediation, and multiple factors that could explain an outcome, and the complexity of both defining and estimating cause and effect.

2.12.6 Framing causality using the linear/OLS model

Students with an economics background are familiar with discussions of causality because causality has been central to econometrics since the creation of the field. Causality was traditionally introduced in the context of linear regression. For example, suppose that we want to estimate treatment effects with the following population regression model:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i. \tag{2.20}$$

We can take the expectation, conditional on treatment, for the treatment group to obtain:

$$E[Y_i|D_i = 1] = \beta_0 + \beta_1 + E[\epsilon_i|D_i = 1]. \tag{2.21}$$

For the control group, we obtain:

$$E[Y_i|D_i = 0] = \beta_0 + E[\epsilon_i|D_i = 0]. \tag{2.22}$$

The average treatment effect is the difference between 2.21 and 2.22:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \beta_1 + E[\epsilon_i|D_i = 1] - E[\epsilon_i|D_i = 0]. \tag{2.23}$$

Thus, β_1 (the naive estimate, δ_{NAIVE}) provides an estimate of average causal effects provided $E[\epsilon_i|D_i = 1] = E[\epsilon_i|D_i = 0]$. In words, the expected value of the error term must be mean independent of treatment assignment. Since we know that independence implies mean independence, we can obtain an estimate of causal effects if treatment is independent of the error term. But note that independence is not needed. We just need the weaker (or milder) mean independence assumption.

Thus, β_1 is not going to identify a causal effect if the model does not include all relevant confounders since any variable not included in the model is part of the error term. If we used Model 2.21 to estimate the conditional randomization example of Section 2.5.2, smoking is part of the error term, and we know that smoking is not independent of treatment assignment. Therefore, β_1 would not identify the causal effect. A common way to describe the fact that the error term is correlated to the treatment is to say that treatment is **endogenous**, a term that originates in mathematical modelling. Thus, saying that the treatment is endogenous is for practical purposes the same as saying that ignorability/CIA does not hold.

Although using linear regression as the basis for discussing causality can be helpful, it can also be confusing. For example, one of the algebraic properties of linear regression says that the correlation between any explanatory variable and the residual is *always* zero, but now we are saying that we must *assume* that the treatment assignment be independent of the error term. However, there is no contradiction because Model 2.20 is a postulated population model, not the estimated model.

Thus, the discussion of causality is whether an estimated regression model from a sample identifies the postulated population model. The other confusing aspect is that in non-linear models like logit, probit, or negative binomial models, there is no separable and additive error term, so it becomes cumbersome to invoke assumptions that hold for a different type of regression model and data-generating mechanism.

For these reasons, we prefer to discuss causality without referring to a particular estimation method or regression model, a key advantage of the potential outcomes framework.

2.13 Missing data

A field in statistics deals with missing data and the problem of **imputing** or predicting the missing data (e.g., **multiple imputation** techniques). Imputing missing data is especially important in the analysis of survey data. Each unit is meant to represent a group of units in the population of interest, so removing units from the analysis because of some missing values affects inference and representativeness.

Two concepts about the *mechanism* for missingness are related to our discussion of causal effects. **Missing Completely at Random** (MCAR) states that a missing value is entirely due to randomness. The missingness of the data is unrelated to any other variable. On the other hand, **Missing at Random** (MAR) states that the missigness is entirely at random *after* conditioning on a set of covariates.

Therefore, the conditional independence assumption can be expressed in terms of the potential outcomes being MAR once we condition for all relevant covariates, which must be observed. Under simple randomization, the potential outcomes are MCAR.

2.14 Further readings and additional material

There are many excellent introductions to the potential outcomes framework. This framework is often called the **Rubin Causal Model** or the **Neyman-Rubin** potential outcomes model. Neyman (1923, 1990) explicitly introduced the idea of potential outcomes in the context of experimental designs, and Rubin (1974) extended the framework to observational studies. We recommend Imbens and Rubin (2015), Chapter 2, for a historical overview (and a complete set of references), and Chapter 3 to conceptualize the assignment mechanism in terms of a probabilistic process. The assignment mechanism is characterized as the probability of any assignment, conditional on all variables and potential outcomes. This approach permits to develop a taxonomy of assignment mechanisms. In this chapter, we provided a simpler approach to conceptualize controlled and uncontrolled assignments.

This chapter has in part followed the structure of Inbems and Rubin (2015) Chapter 1. Angrist and Pischke (2008) is a good introduction to potential outcomes from an economics perspective, and is a bridge between the potential outcomes framework and the understanding of causality using the linear model. Section 2.12.6 is based on Chapter 2 of Angrist and Pischke (2008). Chapter 20 of Wooldridge (2010) also offers a comprehensive introduction to potential outcomes and the idea that causal effects are nonparametrically identified (Section 2.9). We emphasized the importance of understanding the concept of conditional independence. Dawid (1979, 1998) are two classic discussions of conditional independence. Rosembaum and Rubin (1983) introduced propensity scores and strong ignorability, and it is worth reading this important contribution. When reading Rosembaum and Rubin (1983), keep in mind that “conditioning” for a variable does not necessarily imply adding that variable in a regression model, a point we have emphasized in this chapter.

Morgan and Winship (2007, 2015) is a great overview of the potential outcomes framework from a social research perspective. Cook et al. (2002) discusses quasi-experimental designs in great detail without the use of much mathematics. From an epidemiological perspective, Hernan and Robins (2023) is a standard reference. One drawback of this work is that the authors define concepts using different terminology than in most fields. We have attempted to use some of their language since readers might be familiar with this work. Hernan and Robins (2022) use DAGs to depict conceptual frameworks and data-generating processes, as does Pearl (2009), an approach to causality originating in computer science. Pearl (2009) and Pearl et al. (2016) are also good introductions to DAGs. In Pearl’s approach, DAGs provide not only a depiction of the data-generating mechanism, but also a way to solve the identification of causal effects. We also recommend Pearl (2020) for an accessible discussion of causality from different perspectives, including historical and philosophical discussions of cause and effect (see also Reiss 2016).

In some corners of econometrics, the potential outcomes framework is known as the **Roy model** because of Roy (1951). This is rather puzzling because Roy (1951) does not *explicitly* discuss potential outcomes or counterfactuals. Roy (1951) is an excellent and entertaining paper about a specific data-generating mechanism that makes it clear that the observed data cannot identify the underlying process, that is, the causal effect. As Roy (1951) writes, the paper is about “one possible and fairly simple method of selection.” As was the norm at the time, Roy discussed the selection model (into two professions, fishing and hunting) with words, although it is clear that he was translating equations into words.

We do not cover in this chapter **selection models** (Heckman 1980) because they are not commonly used in health services research. Selection models leverage a precise theoretical understanding of the assignment mechanism to both model selection into treatment and outcomes under strong assumptions. For this reason, the approach can be conceptualized as a **structural approach**. Since the assignment mechanism is modeled, this approach has some connections to propensity scores (see Heckman and Todd, 2009). Stata's command `heckman` implements this approach. We recommend Stata's manual entry for an introduction and concrete examples.

There has been intense discussions in economics about the merits and drawbacks of a "natural experiments" approach versus a structural approach to answer research questions. We recommend Imbens (2010) and Angrist (2010) for a discussion. Chapter 9 on instrumental variables introduces some of the terminology related to structural approaches. One part of the discussion is more about the research questions posed than the methods used once a research question is posed. The other part is about the strength of the theoretical approach used when posing and answering research questions. The arguments provide important insights about the methods we cover in this book and their application in practice.

2.15 Summary

This chapter introduced the concept of potential outcomes to define causal effects for a single unit. Potential outcomes are hypothetical outcomes under different treatment states. Causal effects are defined as comparisons of potential outcomes if treated or untreated, although the definition can be extended to more than two treatment states. We can define causal effects without the need to conduct an experiment or collect observational data. After conducting an intervention or collecting data, we only observe some outcomes. Thus, the main challenge of estimating causal effects is that we need to make predictions about the unobserved potential outcomes. We make predictions by using information from multiple units. However, for the predictions to be accurate, treatment and control groups must be comparable. We defined the assumptions needed to identify causal effects under simple randomization, conditional randomization, and uncontrolled data generating mechanisms like observational data. Conditional randomization serves as fundamental mental representation to conceptualized observational data. We reviewed assumptions that can be verified with data and assumptions that cannot be verified with data (exclusion restrictions), which highlights the importance of conceptual/theoretical frameworks and an understanding of the data generating process to establish causality. This chapter also introduced the notation that we use for the rest of the book. Key concepts to remember are the different meanings of conditioning for covariates, the fundamental importance of the assignment mechanism, and the different definition of causal effects, such as local average treatment effects.

Problems

2.1 Use the same data as Section 3.5.2 on conditional randomization. Stata code:

```
use "https://www.perraillon.com/PLH/data/cr_example_sm.dta", clear
```

a. Show that the distribution of BMI and age have good balance and overlap once you condition on smoking status before pregnancy.

b. Estimate the average treatment effect using a linear regression model (use Model 3.12).

c. Verify that there is no treatment heterogeneity by smoking status. Write down the test and null hypothesis.

d. Using the same model as in b) as the basis, add other covariates known to be strongly associated with birthweight: mother's age, mother's education, race, BMI, and pregnancy risk factors (variable *rf_any*). Why does the treatment effect not change by much if we know that these variables are strong predictors of birthweight?

e. Pretend you present the model estimates in a seminar. A statistician in the room – who was falling asleep before you showed the slide with the model – tells you that your model has terrible fit. The R^2 is less than 5%. What would you (politely) say? (To ensure she goes back to sleep?)

2.2 Use the counterfactual framework and the assignment mechanism language to discuss the following study: The Centers for Medicare and Medicaid (CMS) released a new rating of all nursing homes in the US. Nursing homes are assigned stars, from 1 to 5, much like Yelp or Amazon ratings. The stars are meant to reflect nursing homes' quality of care. CMS used data from extensive health inspections, staffing levels, and clinical outcomes to create the ratings. Assume that your research question is to determine how the number of stars affects admissions six months after the release of the ratings. Keep in mind that the intent of the ratings is to motivate consumers to choose nursing homes of better quality, which, in turn, may motivate nursing homes to invest in quality. Therefore, the number of admissions is a metric of the policy success. For simplicity, assume that you want to study the effect of getting less than 3 stars versus getting 3 or more stars (i.e. below average versus average or above average quality).

a. What is the intervention or manipulation (treatment)?

b. What are the potential outcomes?

c. What is the assignment mechanism? (Describe the type of assignment.)

d. Is the assumption of ignorable treatment assignment (or no unmeasured confounders, selection on observables, conditional independence) likely to be satisfied? (You need knowledge of the subject to do a good job so make assumptions if needed and focus on at least two factors.)

e. Describe an ideal experiment to answer the research question. (Ignore ethical or practical concerns. Pretend you are a CMS dictator who can design any experiment. Be creative.)

f. Same as e) but now be more realistic. What you propose in e) may not be ethical or practical. What else could you do?

2.3 Take the following equation about conditional randomization based on severity, where

D is the treatment, Z is severity (the outcome is not a part of this equation): $E[D|Z = 1] \neq E[D|Z = 0]$

a. Why is this true?
b. Severity, Z, is also associated with age and likely associated with the outcome, Y. Do we need to control for age when estimating treatment effects with statistical adjustment? Why or why not?

2.4 (Please think carefully about this question. It will be the key for understanding a method that is not easy to understand.) Suppose that you run an experiment in which you randomly assign unemployed workers to have the opportunity to take a free programming class. You are only giving them the opportunity, but they may opt not to take it. Assume the outcome is the probability of getting a job 3 months after the class is over. Consider estimating ATE and ATET. For each of these estimators (ATE and ATET):

a. Is age a confounder?
b. Is motivation a confounder?
c. Is having young children a confounder?
d. Bonus question: Can you think of a way you could estimate ATET?

2.5 Using the example in the previous question, describe some scenarios that would imply that SUTVA is violated. Remember that SUTVA has two components.

2.6 Explain conditional independence using two examples. Besides verbal examples, use math symbols to make sure you understand the mathematical definition. In other words, use letters for events to present the definition of conditional independence in your examples with math.

2.7 The conditional independence assumption (ignorability, exchangeability, and so on) says that:

$Y_{1i}, Y_{0i} \perp D_i|X_i$.

We could write the same as:

$(Y_{1i}, Y_{0i}) \perp D_i|(X_{1i}, ..., X_{ji})$

Statistical theory does not tell us which are the relevant variables $X_1, ..., X_j$ that we must condition for. We an understanding of the data generating mechanism (i.e., conceptual framework). Why is this assumption about potential outcomes and not about observed outcomes? In other words, why do we not write: $Y_i \perp D_i|(X_{1i}, ..., X_{ji}$?

2.8 True/False. Explain in about 2 sentences why the following statements are true or false. If the statement is not 100% true, then it's false.

a. The identification of causal effects requires conducting an intervention.
b. In conditional randomization, treatment assignment is independent of potential outcomes.
c. A confounder of the relationship between Y and D is a variable X that is associated with Y, but not with D.
d. External validity refers to the extent to which a study can provide causal effects.
e. With observational data, overlap requires that the means of the covariates share the same range of values between treatment and control groups.

f. To estimate the Average Treatment Effect on the Treated (ATET), we only need the data from the treated group.
g. All variables that determine selection into treatment must be conditioned on to identify a causal effect.
h. Without a full understanding of the assignment mechanism, you can still identify a causal effect by including **all** covariates that are associated with the treatment and outcome in your dataset.

2.9 For each of the following, explain whether and how they should be controlled for when estimating a treatment effect, and what the impact would be if they were incorrectly accounted for in your model: confounder, mediator, and moderator.

2.10 If the ”no spillovers” component of the SUTVA assumption were violated, how (directionally) would your treatment effect estimate be affected?

2.11 Strong ignorability requires both conditional independence and overlap. Explain the issue with estimating a counterfactual if the overlap assumption is not satisfied.