This material primarily comes directly from me, but you should also read H: 2.30. I didn't directly refer to it while I was writing these notes, but, if you want additional related material, you can consult Scott Cunningham's *The Mixtape* (particularly chapters 1 and 4).

These notes provide an introduction to causal inference, particularly with observational data. In addition they cover: (i) how should you interpret regressions under treatment effect heterogeneity? and (ii) what are some alternative approaches to estimation besides linear regression that might be useful in this context?

Causal Effects

H 2.30 (though much of the material below is not included in the textbook)

Now, let's move to thinking about causal effects. I'll talk briefly about how to think about this conceptually and then how this is related to regression derivatives and linear regression.

> **Notation**
> In cases (like in the current section) where we are interested in understanding the effect of particular variable, I may denote it by D (which is common in many academic papers), while referring to all remaining regressors as X (I'll probably also use the term "covariates" for these other regressors).

Binary Treatment

Work on understanding the effect of a particular variable of interest on some outcome is typically called the "treatment effects literature". This terminology originates from the biostatistics literature where a treatment could literally refer to a medical treatment. We'll use the term "treatment' ' more broadly to refer to a policy or some intervention that we are interested in studying.

Let's start with the case where the treatment is binary; that is $D_i = 1$ if a unit participates in the treatment and $D_i = 0$ if a unit does not participate in the treatment.

We'll also define **potential outcomes** $Y_i(1)$ and $Y_i(0)$ – these are the outcomes that a unit would experience if it participated in the treatment or if it did not participate in the treatment, respectively. Introducing potential outcomes notation provides a natural way to think about causal effects. In particular, we can define the **treatment effect** (or **causal effect**) of the treatment for unit i as $TE_i = Y_i(1) - Y_i(0)$. This is the difference between the potential outcome that would occur if unit i participates in the treatment and the potential outcome that would occur if unit i does not participate in the treatment, i.e., how the treatment causes the outcome to change for unit i.

For any, particular unit, the researcher only observes one of these potential outcomes; that is, for treated units, we observe their treated potential outcomes, and for untreated units, we observe their untreated potential outcomes. We can therefore write the observed outcome as

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

and, it is convenient to note that this can also be written as

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

which follows just by re-arranging terms from the previous equation.

Target Parameters

In the context of a binary treatment, much research targets one of the following two parameters:

$$ATE := \mathbb{E}[Y(1) - Y(0)]$$
$$ATT := \mathbb{E}[Y(1) - Y(0)|D = 1]$$

ATE stands for "average treatment effect" and ATT stands for "average treatment effect on the treated". ATE is the average difference between treated and untreated potential outcomes for the entire population. ATT is the average difference between treated and untreated potential outcomes among those that participate in the treatment.

It may seem like ATE is inherently more interesting than ATT, but I don't think this is necessarily the case. To give an example, suppose you are interested in studying the causal effect of job training on people's earnings. Presumably, the effect of job training is exactly 0 for a large portion of the population. In this case, ATT is probably the more relevant parameter to aim to identify — it is the average effect of job training among those that actually participate.

For much of the course, we will target identifying the ATT — at the beginning of the course, this is mainly to make the arguments more concise, and we could instead target ATE. That said, there are some cases where we will explicitly target ATE, and there will be some other case (particularly when we discuss panel data) where it would require different sorts of arguments to identify ATE relative to ATT.

Experiments

If we had access to an experiment (that is, that we could randomly assign units to either participate in the treatment or not), it would follow that

$$(Y(1), Y(0)) \perp\!\!\!\perp D \tag{2}$$

In words, if we can randomly assign treatment, then (by construction) potential outcomes are independent of participating in the treatment. More informally, there is "nothing special" about units that participate in the treatment relative to those that do not participate in the treatment in terms of their potential outcomes.

Let's think about identifying ATT under random assignment as in Equation 2. Notice that

$$\begin{aligned} ATT &= \mathbb{E}[Y(1) - Y(0)|D = 1] \\ &= \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1] \\ &= \underbrace{\mathbb{E}[Y|D = 1]}_{\text{Easy}} - \underbrace{\mathbb{E}[Y(0)|D = 1]}_{\text{Hard}} \end{aligned}$$

The previous display indicates that ATT is equal to the average outcome actually experienced by the treated group relative to the average outcome among those in the treated group if they had not participated in the treatment. The first term is "easy" because those outcomes are observed outcomes. The second term is "hard" because we do not observe untreated potential outcomes for the treated group.

However, Equation 2 implies that $\mathbb{E}[Y(0)|D = 1] = \mathbb{E}[Y(0)|D = 0]$. That is, because untreated potential outcomes are independent of treatment, the average untreated potential outcome among the treated group is the same as the average untreated potential outcome among the untreated group. This, therefore, implies that (given random assignment):

$$ATT = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0]$$

That is, we can recover the ATT by comparing the average outcomes among the treated group relative to the average outcomes among the untreated group.

> **Practice:** Given the above expression for ATT, what is the natural way to estimate ATT?

Connection to Regression

Now, let's think about how to estimate causal effects using a regression (and given random assigment) — this is going to be very simple, but I think it is worth explaining so that we can use the same sorts of procedures in more complicated cases below.

Let's write an extremely simple model for untreated potential outcomes:

$$Y_i(0) = \beta_0 + e_i \tag{3}$$

By construction, we have that $\mathbb{E}[e] = 0$, but random assigment also implies that $\mathbb{E}[e|D = d] = 0$ for $d \in \{0, 1\}$. To see this, notice that $\mathbb{E}[Y(0)|D = d] = \beta_0 + \mathbb{E}[e|D = d]$. Recall that random assignment implies that $\mathbb{E}[Y(0)|D = 1] = \mathbb{E}[Y(0)|D = 1]$, therefore it must be the case that $\mathbb{E}[e|D = 1] = \mathbb{E}[e|D = 0] = 0$.

Let's also make an additional assumption called **treatment effect homogeneity**. In math, we can write this as $Y_i(1) - Y_i(0) = \alpha$. This means that the effect of participating in the treatment is the same for all units (and is equal to α). This is probably a strong assumption; in my view, one would expect that the effect of participating in most any treatment could conceivably vary across

units (especially in economics, social sciences, and most business applications). But let's just make this assumption for now — we'll talk about it much more in the future.

Next, notice that

$$\begin{aligned} Y_i &= Y_i(0) + D_i(Y_i(1) - Y_i(0)) \\ &= Y_i(0) + \alpha D_i \\ &= \beta_0 + \alpha D_i + e_i \end{aligned} \tag{4}$$

where the first equality comes from Equation 1, the second equality holds by treatment effect homogeneity, and the last equality holds from Equation 3 and by rearranging terms. Moreover, because $\mathbb{E}[e|D] = 0$, this suggests estimating α (the causal effect of the treatment) by running a regression of Y on D.

To conclude this discussion, it is interesting to notice that, the regression in Equation 4 is fully saturated. This means that the number of values that the conditional expectation can take (here 2, one for each value of D) is equal to the number of parameters in the model, and we can therefore exactly map the regression coefficients to the conditional expectations. In particular, notice that

$$\begin{aligned} \mathbb{E}[Y|D=1] &= \beta_0 + \alpha \\ \mathbb{E}[Y|D=0] &= \beta_0 \end{aligned}$$

and subtracting the second equation from the first equation and re-arranging implies that

$$\alpha = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$$

which further implies that $\alpha = ATT$. This is interesting because we derived the regression in Equation 4 under the extra condition of treatment effect homogeneity. However, that $\alpha = ATT$ implies that this regression is *robust* to treatment effect heterogeneity.

Unconfoundedness

In most application in economics, researchers do not have access to an experiment (or, alternatively, do not have the ability to randomly assign units to participate in the treatment or not). In cases with "observational" data (meaning: non-experimental data), one of the most common assumptions for thinking about causal effects is the following unconfoundedness assumption (you may also sometimes hear this called selection-on-observables, and the textbook refers to this as a conditional independence assumption):

$$(Y(1), Y(0)) \perp\!\!\!\perp D \big| X$$

Unconfoundedness says that potential outcomes are independent of the treatment *after conditioning on some covariates* X. Informally, unconfoundedness means that, among units with the same characteristics X, the distribution of treated and untreated potential outcomes is the same among

the treated and untreated group (though the distribution of X could differ across groups). If you want to assume unconfoundedness, this often needs to be rationalized (perhaps informally) theoretically.

> **Side Comment:** Sometimes the assumption that $Y(0) \perp\!\!\!\perp D|X$ can be meaningfully weaker that what I have called unconfoundedness above. In particular, this assumption just implies that treated and untreated units with the same characteristics X have the same distribution of untreated potential outcomes (but would allow for treated units to, for example, have systematically better treated potential outcomes that untreated units). The assumption in this comment is strong enough to identify ATT, but it is not strong enough to identify ATE.

Under unconfoundedness, notice that

$$\begin{aligned}
ATT &= \mathbb{E}[Y(1) - Y(0)|D=1] \\
&= \mathbb{E}\big[\underbrace{\mathbb{E}[Y(1) - Y(0)|X, D=1]}_{ATT(X)}|D=1\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y(1)|X, D=1] - \mathbb{E}[Y(0)|X, D=1]|D=1\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y(1)|X, D=1] - \mathbb{E}[Y(0)|X, D=0]|D=1\big] \\
&= \mathbb{E}\big[\mathbb{E}[Y|X, D=1] - \mathbb{E}[Y|X, D=0]|D=1\big]
\end{aligned}$$

where the first equality is just the definition of ATT, the second equality holds by the law of iterated expectations, the third equality holds by pushing the expectation through the difference, the fourth equality holds by unconfoundedness, and the last equality holds by re-writing potential outcomes in terms of their observed counterparts.

This implies that ATT is **nonparametrically identified** under the assumption of unconfoundedness — that is, it can be related to population quantities that we have analogues of in the data that we observe.

To understand this expression, let us work from the inside out. The term $\mathbb{E}[Y|X, D=1] - \mathbb{E}[Y|X, D=0]$ is the difference between the average outcome for the treated group relative to the average outcome for the untreated group given a particular value of X. Under unconfoundedness, this comparison delivers a causal effect of the treatment for that value of X. The outside expectation averages this causal effect over the distribution of X for the treated group. This is a natural way to think about the average treatment effect for the treated group: it is the average of the causal effects for each value of X weighted by how common that value of X is among the treated group.

To give an example, suppose that the treatment is whether or not a person goes to college; further, suppose that we are willing to assume unconfoundedness conditional on parents' income (note: this assumption is not likely to be plausible, but let's just go with it here). In this case, $\mathbb{E}[Y|X, D=1]$ is the average earnings of people who went to college conditional on parents' income. $\mathbb{E}[Y|X, D=0]$ is the average earnings of those that did not go to college conditional on their parents' income.

By the unconfoundedness assumption, this is equal to $\mathbb{E}[Y(0)|X, D = 1]$—the average earnings that people who (i) went to college and (ii) have the same value of parents' income *would have experienced* if they had not gone to college. Thus, this comparison yields the causal effect of college for a particular value of parents' income. The outside expectation averages the causal effect over the distribution of parents' income *among those that went to college*. This latter step allows for the distribution of parents' income to differ (perhaps significantly) among those that went to college and those that did not go to college.

The expression for ATT above is, in my view, the easiest one to interpret, but you can simplify this expression as follows, which will be especially useful when we discuss how to estimate ATT

$$ATT = \mathbb{E}[Y|D = 1] - \mathbb{E}\big[\mathbb{E}[Y|X, D = 0]\big|D = 1\big]$$

which follows immediately from the previous expression by applying the law of iterated expectations to the first term. This expression highlights that the main challenge in estimation will be for the term $\mathbb{E}[Y|X, D = 0]$, and, because of our focus on the ATT, we can actually side-step estimating $\mathbb{E}[Y|X, D = 1]$ and simply estimate $\mathbb{E}[Y|D = 1]$—the overall average outcome for the treated group.

Although the discussion above shows that ATT is nonparametrically identified, it may be practically difficult to (nonparametrically) estimate the ATT using the above expression. This would particularly be the case if the dimension of X is relatively large as estimating $\mathbb{E}[Y|X, D = 0]$ would start to suffer from the curse of dimensionality that we talked about in the introductory slides. Thus, in many applications, it might be desirable to have simpler (i.e., more feasible) estimation strategies. And, for this reason, we are going to try to connect unconfoundedness to running regressions. This will involve some extra assumptions, but it will result in (very) simple estimation approaches.

To connect this to running a regression, let's make some additional assumptions. **Linear model for untreated potential outcomes:**

$$Y_i(0) = X_i'\beta + e_i$$

This is a linearity assumption for untreated potential outcomes. Notice that unconfoundedness implies that $\mathbb{E}[Y(0)|X, D = 1] = \mathbb{E}[Y(0)|X, D = 0]$ which (given linearity) implies that $\mathbb{E}[e|X, D = d] = 0$ for $d \in \{0, 1\}$. Next, let's make the treatment effect homogeneity assumption that $Y_i(1) - Y_i(0) = \alpha$. Then,

$$\begin{aligned} Y_i &= Y_i(0) + D_i(Y_i(1) - Y_i(0)) \\ &= Y_i(0) + \alpha D_i \\ &= \alpha D_i + X_i'\beta + e_i \end{aligned}$$

where the first equality holds by Equation 1, the second equality holds by the treatment effect homogeneity condition, and the third equality holds by the model for untreated potential outcomes and by rearranging. Moreover, because $\mathbb{E}[e|X, D] = 0$, this is a correctly specified linear CEF model.

This equation suggests estimating the causal effect of D on Y by running a regression of Y on D and X and interpreting the estimated coefficient on D as an estimate of the causal effect.

Unlike in the earlier case of random assignment, this regression is not robust to violations of treatment effect homogeneity. Later in the semester, we will talk about exactly what this regression recovers in the presence of treatment effect heterogeneity, and we will also talk about some alternative methods that are more robust to violations of treatment effect homogeneity. It is also not robust to violations of the linear model for untreated potential outcomes. I am not totally sure about this, but my sense is that, in cases where unconfoundedness holds, that the "empirical relevance" of violations of treatment effect homogeneity and linearity of untreated potential outcomes are relatively small. And, at any rate, under unconfoundedness, running a regression of Y on D and X is by far the most common approach used in empirical work.

Continuous Treatment

So far, we have talked about the case with a binary treatment. Next, let's move to the case where the treatment can take on a continuum of values. I'll talk here about the case where the treatment can take values in $\mathcal{D} = \{0\} \cup [d_L, d_U]$. In other words, it is possible that some units do not participate in the treatment at all, but, otherwise, the treatment is continuous in the range from d_L to d_U. I won't cover intermediate cases such as a multi-valued discrete treatment, but the arguments would basically be a combination of the ones in this section with the ones in the previous section with binary treatment. To fix ideas, you can think of continuous treatment examples such as the amount of "dose" of some medical treatment (e.g., number of Advils to treat a headache or the "amount" of a Covid-19 vaccine); as an economics example, one example is intergenerational income mobility where the outcome is child's income and the continuous treatment is parents' income, and another example is quantity demanded where the outcome is quantity demanded and the continuous treatment is price.

We use D_i to denote the actual amount of the treatment that unit i experiences. We'll define potential outcomes using a slightly extended notation from the previous extension. In particular, let $Y_i(d)$ denote the outcome that would occur for unit i if they were to experience dose d. The observed outcome is given by

$$\begin{aligned} Y_i &= Y_i(D_i) \\ &= Y_i(0) + (Y_i(D_i) - Y_i(0)) \end{aligned} \tag{5}$$

In other words, we observe outcomes corresponding to the actual amount of the treatment for a particular unit. The second equality holds by adding and subtracting $Y_i(0)$ and will be helpful in some derivations below. As a side-comment, in cases where it is not possible to be untreated or where defining untreated potential outcomes is somehow "awkward"; the arguments below will follow with trivial modifications by replacing "untreated" with the smallest possible amount of the

treatment.

Let's briefly talk about the sorts of parameters that you could be interested in for this case. A natural starting point is to consider **level effects** such as

$$ATE(d) := \mathbb{E}[Y(d) - Y(0)]$$

These are quite similar to ATE that we talked about in the case with a binary treatment. $ATE(d)$ is the overall average difference between potential outcomes under dose d relative to untreated potential outcomes.

When the treatment is continuous, it also makes sense to think about "slope effects" that are derivatives of the above parameters. For example, one could be interested **average causal response**

$$ACR(d) := \frac{\partial\, ATE(d)}{\partial\, d}$$

This is how much outcomes causally increase on average under a marginal increase in the dose/treatment.

> **Side Comment:** $ACR(d)$ is a functional parameter — you could plug in different values of d and $ACR(d)$ could take a different value. Many times researchers would like to report a single number to summarize the causal effect of a treatment. In this case, a natural summary measure is
>
> $$ACR^o := \mathbb{E}[ACR(D)|D > 0]$$
>
> which is just $ACR(d)$ averaged over the distribution of the dose. Below, when we talk about regressions, these generally output a single number, and it is natural to compare that number to ACR^O (ideally, we would like the regression to deliver ACR^O).

Let's start with the case where the amount (sometimes this is called the "dose") of the treatment is randomly assigned. This implies that, for all $d \in \mathcal{D}$,

$$Y(d) \perp\!\!\!\perp D$$

In other words, potential outcomes are independent of the amount of the treatment.

Let's show that some of the parameters of interest above are identified. First, let's consider $ATE(d)$. In this case,

$$\begin{aligned} ATE(d) &= \mathbb{E}[Y(d)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y(d)|D = d] - \mathbb{E}[Y(0)|D = 0] \\ &= \mathbb{E}[Y|D = d] - \mathbb{E}[Y|D = 0] \end{aligned}$$

where the first equality is just the definition of $ATE(d)$, the second equality holds by random assignment, and the third equality re-writes potential outcomes in terms of their observed counterparts. This shows that, under random assignment, $ATE(d)$ is identified. And, in particular, it is given by the mean outcome among those that experienced dose d relative to the mean outcome among those that were untreated. This is not surprising: random assignment means that to think about average treatment effects, we can take units that experienced some particular amount of the treatment (because of random assignment their outcomes are not systematically different from outcomes among those that experienced some other amount of the treatment) and we can compare these outcomes to the mean of outcomes experienced by the untreated group (under random assignment, these outcomes are not systematically different from the outcomes others would have experienced if they had been untreated).

We can also recover $ACR(d)$ by taking the derivative of the previous expression; that is,

$$ACR(d) = \frac{\partial \mathbb{E}[Y|D=l]}{\partial l}\Big|_{l=d}$$

which holds because $\mathbb{E}[Y|D=0]$ does not depend on d.

The above discussion implies that $ATE(d)$ and $ACR(d)$ are both nonparametrically identified. Now, let's think about nonparametrically estimating these. As long as you have access to an untreated group, then the term $\mathbb{E}[Y|D=0]$ is easy to estimate — just subset the data down to untreated observations and calculate their average outcome. However, when the treatment is continuously distributed, $\mathbb{E}[Y|D=d]$ is trickier to estimate; in particular, if the treatment is truly continuous then there are likely to be 0 observations that have dose exactly equal to 0 which suggests that the same "subsetting" strategy is not likely to work. Instead, most nonparametric estimation strategies take observations that are "close" to d and average them together (we will leave the definition of "close" vague for now as there are several ways to think about this and this discussion can become quite technical). Broadly, this strategy should work pretty well. There is no curse of dimensionality here since D is a scalar. Estimating $ACR(d)$ is somewhat more challenging (intuitively, it should make sense that estimating derivatives of functions well is more challenging than estimating the function itself though they are clearly related). I'm not going to talk about how you would do it now, but, in many application, it is probably feasible to do this too.

In my view, if you are in this case, you ought to seriously consider the nonparametric estimation approaches discussed above, but I think that it is much more common to use regressions in this case too. My sense is that this is for two reasons: (i) although the nonparametric approaches mentioned above are likely to be "feasible", they are definitely more complicated than running a regression, (ii) you need to choose some way to define "close" and, it turns out, that results can be quite sensitive to this choice, but regressions (for better or worse) side-step this choice.

Now, let's discuss how you can connect the previous discussion to running a regression. As in the case with a binary treatment, let's start by making a treatment effect homogeneity assumption:

for all $d \in \mathcal{D}$, $Y_i(d) - Y_i(0) = \alpha d$. Notice that this implies that

$$\begin{aligned} Y'(d) &:= \lim_{h \to 0} \frac{Y(d+h) - Y(d)}{h} \\ &= \lim_{h \to 0} \frac{\alpha(d+h) - \alpha d}{h} \\ &= \alpha. \end{aligned}$$

where the second line uses the treatment effect homogeneity assumption, and the last line follows just from canceling terms. This means that α should be interpreted as how much outcomes causally increase under a one unit increase in the dose, and (under the assumptions we have made) this is constant across units and across different amounts of the dose.

As in the previous section, treatment effect homogeneity is likely to be very strong. As in the case with a binary treatment, it restricts treatment effects to be constant across units. In this case it is additionally potentially restrictive in that it requires that the causal effect of more dose is the same regardless of the "starting dose" (for example, it would be a very strong assumption to assume that every time you increase the number of Advil that you take it reduces your headache by the same amount). As before, let us delay trying to relax this assumption and/or thinking about what potential issues it could cause and just go with it for now.

Finally, let's use the same model for untreated potential outcomes as in Equation 3, where from random assignment, it holds that $\mathbb{E}[e|D = d] = 0$.

Now, notice that

$$\begin{aligned} Y_i &= Y_i(0) + (Y_i(D_i) - Y_i(0)) \\ &= Y_i(0) + \alpha D_i \\ &= \beta_0 + \alpha D_i + e_i \end{aligned}$$

where the first equality uses Equation 5, the second equality uses treatment effect homogeneity, and the third equality uses Equation 3 and re-arranges terms. Thus, under random assignment of the amount of the treatment and the version of treatment effect homogeneity discussed above, the CEF of Y given D is linear and the coefficient on D has a causal interpretation. This discussion suggests to run a regression of Y on D and interpret α as the causal effect of a marginal increase in the dose.

Like the case of unconfoundedness above, treatment effect homogeneity matters in a potentially meaningful way here. We'll come back to this issue in a few weeks and discuss how α can be interpreted without treatment effect homogeneity. As in the previous case, my sense is that running the above regression would still be the leading approach to estimating causal effects in this case though, and it is not entirely clear to me how much using alternative approaches that are robust to treatment effect heterogeneity actually matter.

To conclude this section, let's briefly consider the case of a continuous treatment under unconfoundedness. That is, let's assume that, for all $d \in \mathcal{D}$,

$$Y(d) \perp\!\!\!\perp D|X$$

Practice: Show that $ATE(d)$ and $ACR(d)$ are nonparametrically identified under the above unconfoundedness assumption and provide an expression for them.

If you complete the above practice problem, you will see that $ATE(d)$ and $ACR(d)$ depend on terms like $\mathbb{E}[Y|X, D = d]$ (given our above discussion about unconfoundedness with a binary treatment, this should not come as a surprise to you). Although these sorts of terms are identified, they can be very challenging to nonparametrically estimate particularly when X is moderate- or high-dimensional. For this reason, it is often empirically useful to provide conditions under which one can estimate causal effects of a continuous treatment using a regression. As earlier, the benefit here is a (much) simpler estimation strategy, and the cost is some extra assumptions.

Let's make some assumptions that lead to using a regression to estimate the causal effect of a small increase in the dose. As in the case of a binary treatment under unconfoundedness, let's assume that untreated potential outcomes are generated by the following linear model:

$$Y_i(0) = X_i'\beta + e_i$$

where the linearity is the key assumption here. Given linearity, we have that $\mathbb{E}[e|X] = 0$. Unconfoundedness additionally implies that $\mathbb{E}[e|X, D = d] = 0$ for all $d \in \mathcal{D}$. Next, let's make the treatment effect homogeneity assumption that, for all $d \in \mathcal{D}$, $Y_i(d) - Y_i(0) = \alpha d$. Then, following the same sorts of arguments that we have been using in earlier sections

$$\begin{aligned} Y_i &= Y_i(0) + (Y_i(D_i) - Y_i(0)) \\ &= Y_i(0) + \alpha D_i \\ &= \alpha D_i + X_i'\beta + e_i \end{aligned}$$

which holds using similar arguments as we have used before and suggests estimating the causal effect of a marginal increase in the dose by running a regression of Y on D and X.

As you would expect (given that this is the most complicated setup we have considered so far), this regression is not fully robust to (i) violations of treatment effect homogeneity or (ii) misspecification of the model for untreated potential outcomes. That said, we'll re-visit what exactly α is under treatment effect heterogeneity and potential misspecification in several weeks.

Covariate Balance

Let us suppose, momentarily (and probably unrealistically for most applications) that the distribution of X is the same for the treated group as it is for the untreated group, which we can write

as

$$D \perp\!\!\!\perp X$$

Sticking with the ATT (though noting that similar arguments would apply for ATE), we have, under this condition, that

$$\begin{aligned} ATT &= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}[Y|X,D=0]\Big|D=1\Big] \\ &= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}[Y|X,D=0]\Big|D=0\Big] \\ &= \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \end{aligned}$$

where the first equality holds from the discussion in the previous section, the second equality from the condition mentioned above, and the last equality by the law of iterated expectations. This is an interesting result as it says that, even though we need to compare units with the same characteristics in order to recover causal effects, when there is **covariate balance**, the unadjusted comparison of means effectively does compare units with the same characteristics; i.e., you could reverse engineer the discussion above to show that $\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] = \mathbb{E}\Big[\mathbb{E}[Y|X,D=1] - \mathbb{E}[Y|X,D=0]\Big|D=1\Big]$ under the conditions we have used here.

It is common to check covariate balance in empirical applications. For example, you can compare the estimated means or variances of covariates for the treated group relative to the untreated group. It would be uncommon for them to be perfectly balanced, but, in general, causal inference is easier when the distribution of covariates is not-too-different across groups than if they are much different. Some approaches to estimation involve trying to force balance in the distribution of covariates between the treated and untreated group, but we will save that discussion for later on in these notes.

The discussion above is conceptually related to our discussions of omitted variable bias earlier in the semester: it is okay not to control for variables that are not correlated with the regressor of interest.

Proxies

The other point that I'd like to emphasize is that it implies that we don't actually need to control for every covariate that we'd like to condition on when we make comparisons between the treated group and untreated group—just the ones that are distributed differently across the treated and untreated groups. This can be quite useful in practice because (1) the number of covariates that you might want to condition on could be quite large, and (2) in many applications, you may not observe all of the covariates that you would like to condition on. In fact, suppose that

$$\Big(Y(1), Y(0)\Big) \perp\!\!\!\perp D|(X,W)$$

where X is a vector of observed covariates and W is a vector of unobserved covariates. Now, consider another vector R_i of proxies of X_i and W_i, such that

$$D \perp\!\!\!\perp (X, W)|R \quad \text{and} \quad \Big(Y(1), Y(0)\Big) \perp\!\!\!\perp R|(X, W)$$

where the first part ways that the distribution of (X, W) is the same for the treated group as for the untreated group conditional on R. The second part is a redundancy condition that says that, effectively, the potential outcomes are unrelated to R as long as we condition on (X, W).

Example (Minimum Wage): Suppose you are interested in studying the effect of state-level minimum wage changes on employment and that you have access to county-level data. In this case, you might think that unconfoundedness holds after conditioning on variables such as a county's population density, income distribution, industry mix, political preferences, religiosity, or other things. Some of these variables you might be able to observe directly (e.g., population density), others you might observe related quantities (e.g., a county's median income and/or poverty rate, which are often observed, are related to its income distribution), and others you might not observe at all (e.g., industry mix could be hard to observe and summarize and religiosity might not be observed at all). In this case, we might take R_i to include variables such as the region of the country it is located in or whether or not it is a rural county. [To be clear, you could take X to include region and rural, or you could take R to include population density, median income, and poverty rate, but, to make things concrete, let's just say that these are separate from each other.]

In this application, the first assumption mentioned above states that, after conditioning on region and rural, the distribution of all the other variables that show up in the unconfoundedness assumption is the same for each group. The second part says that, if we could control for population density, income distribution, industry mix, political preferences, religiosity, etc., then the potential outcomes would be unrelated to region and rural.

Under the conditions discussed above, we have that

$$\begin{aligned}
ATT &= \mathbb{E}[Y|D=1] - \mathbb{E}[Y(0)|D=1] \\
&= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}[Y(0)|R, D=1]\Big|D=1\Big] \\
&= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y(0)|X, W, R, D=1]\big|R, D=1\big]\Big|D=1\Big] \\
&= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y(0)|X, W, D=0]\big|R, D=1\big]\Big|D=1\Big] \\
&= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[Y(0)|X, W, R, D=0]\big|R, D=0\big]\Big|D=1\Big] \\
&= \mathbb{E}[Y|D=1] - \mathbb{E}\Big[\mathbb{E}[Y(0)|R, D=0]\Big|D=1\Big]
\end{aligned}$$

where the first equality is the definition of ATT, the second and third equalities hold by applying the law of iterated expectations, the fourth equality holds by the redundancy condition and unconfoundedness, the fifth equality holds by the redudancy condition again and because the distribution of (X, W) is the same for the treated group and untreated group conditional on R, and the last equality holds by the law of iterated expectations. Everything is identified in the last line.

The result above shows that, in applications with proxies like R above, we can recover the ATT by comparing treated and untreated units that are alike in terms of R, under the assumption of unconfoundedness given (X, W). This is kind of a hybrid result that sits in between targeting ATT directly by conditioning on covariates X and the case with observed covariate balance that we talked about above. Just like the covariate balance case discussed above, the condition that $D \perp\!\!\!\perp (X, W)|R$ is partially testable. We can test whether the distribution of X is the same for the treated group and untreated group conditonal on R. In the minimum wage example mentioned above, we could check if, e.g., the mean of median income is the same for treated and untreated counties conditional on region and rural. We cannot test the part of the condition about W, but, if our approach balances observed X, then you might at least hope that it also balances unobserved W.

Interpreting Regressions under Treatment Effect Heterogeneity

Next, let's move to estimation. Very early on in the semester, we thought some about why one would want to use a regression in order to try to answer research questions — where research questions typically involve trying to answer "causal" questions when the researcher had access to observational data, under the assumption of unconfoundedness mentioned above. In addition to unconfoundedness, we also used the following two additional assumptions.

- **Treatment Effect Homogeneity:** $Y_i(1) - Y_i(0) = \alpha$ for all units
- **Linear model for untreated potential outcomes:** $Y_i(0) = X_i'\beta + e_i$.

In this context, we showed that you could run the following regression:

$$Y_i = \alpha D_i + X_i'\beta_0 + e_i \tag{6}$$

and interpret α as an estimate of the causal effect of D on Y.

For this section, I would like to maintain the unconfoundedness assumption while thinking about relaxing the treatment effect homogeneity and (sometimes) the linear model assumptions. We will think about how to interpret α in this context.

My impression is that if a researcher was running the regression in Equation 6 in the presence treatment effect heterogneity, it's likely that the researcher would hope that α would be equal to (or at least related to) the ATE. For this reason, in the arguments below, I'll mainly focus on the relationship between α and ATE, but you could develop very similar arguments for the relationship between α and ATT.

Preliminaries

Before discussing in detail how to interpret α in the regression discussed above, we need to introduce some additional notation as well as cover some useful preliminary results.

We will use the following notation. First, some of the results below involve **conditional average treatment effects** which we define as $CATE(X) := \mathbb{E}[Y(1) - Y(0)|X]$. We will allow for the possibility that these vary across different values of X. Next, let $p = \mathrm{P}(D = 1)$. And let $p(x) = \mathrm{P}(D = 1|X = x)$ denote the **propensity score** which is the probability of being treated conditional on having covariates $X = x$.

Next, since we are talking about trying to understand regression coefficients, a number of the arguments below involve linear projections. I'll mostly stick to "population" arguments rather than "sample" arguments below, so we'll write population versions of linear projection. In particular, we'll denote the linear projection of D on X by

$$\mathrm{L}(D|X) := X'\gamma = X'\mathbb{E}[XX']^{-1}\mathbb{E}[XD]$$

and we'll also use linear projections of the outcome on covariates separately for the treated group and the untreated group; we'll denote these by

$$\begin{aligned}\mathrm{L}_1(Y|X) &:= X'\beta_1 = X'\mathbb{E}[XX'|D = 1]^{-1}\mathbb{E}[XY|D = 1]\\ \mathrm{L}_0(Y|X) &:= X'\beta_0 = X'\mathbb{E}[XX'|D = 0]^{-1}\mathbb{E}[XY|D = 0]\end{aligned}$$

One property of linear projections that we will use below is the following:

$$\begin{aligned}\mathbb{E}[\mathrm{L}(D|X)\mathrm{L}_d(Y|X)|D = d] &= \mathbb{E}\Big[(\gamma'X)X'\mathbb{E}[XX'|D = d]^{-1}\mathbb{E}[XY|D = d]\big|D = d\Big]\\ &= \mathbb{E}[\gamma'XY|D = d]\\ &= \mathbb{E}[\mathrm{L}(D|X)Y|D = d] \qquad (7)\end{aligned}$$

where the first equality holds because $Y = X'\beta_d + e_d$, the second equality holds because $\mathbb{E}[Xe_d|D = d] = 0$, because e_d is a projection error (conditional on $D = d$), and the last equality holds by using the definition of $\mathrm{L}(D|X)$ again.

Next, let me mention a few "tricks" that we will use below. First, notice that the law of iterated expectations implies that

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}\Big[\mathbb{E}[Y|D]\Big]\\ &= \sum_{d\in\{0,1\}} \mathbb{E}[Y|D = d]\mathrm{P}(D = d)\\ &= \mathbb{E}[Y|D = 1]p + \mathbb{E}[Y|D = 0](1 - p) \qquad (8)\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds because

the outside expectation is over the distribution of D and because D is binary, and the last equality holds pretty much immediately. A similar argument implies that

$$\mathbb{E}[DY] = \mathbb{E}[DY|D=1]p + \underbrace{\mathbb{E}[DY|D=0]}_{=0}(1-p) = \mathbb{E}[Y|D=1]p$$
$$\mathbb{E}[(1-D)Y] = \underbrace{\mathbb{E}[(1-D)Y|D=1]}_{=0}p + \mathbb{E}[(1-D)Y|D=0](1-p) = \mathbb{E}[Y|D=0](1-p)$$

Below, we'll actually more often use the rearranged versions of these that

$$\mathbb{E}[Y|D=1] = \mathbb{E}\left[\frac{D}{p}Y\right] \qquad \text{and} \qquad \mathbb{E}[Y|D=0] = \mathbb{E}\left[\frac{1-D}{1-p}Y\right] \tag{9}$$

Intuitively, you can think the following: $\mathbb{E}[DY]$ would mix together the mean of Y for the treated group with a bunch of 0's for the untreated group (because $D=0$ for the untreated group). In order for this to be equal to $\mathbb{E}[Y|D=1]$, you need to "inflate" it to account for the 0's. Dividing by p (which is between 0 and 1) is what does this.

Similar sorts of arguments apply for conditional expectations. In particular, by the law of iterated expectations

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, D=1]p(X) + \mathbb{E}[Y|X, D=0](1-p(X)) \tag{10}$$

and, using the same sort of arguments as above

$$\mathbb{E}[DY|X] = \mathbb{E}[Y|X, D=1]p(X) \tag{11}$$

A number of terms that we will consider below look like $\mathbb{E}[g(X)]$ for some function g. This sort of term involves averaging over the population distribution of X. It will sometimes be useful for us to switch to averaging over the distribution of X for the treated group or untreated group. Of course, in general, $\mathbb{E}[g(X)] \neq \mathbb{E}[g(X)|D=1]$. However, notice that (for simplicity, I am going to implicitly assume here that X is continuously distributed and has a pdf f, so this is more of a sketch of an argument, but the result below holds more generally)

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int g(x)\, f(x)\, dx \\
&= \int g(x)\, \frac{f(x)}{f(x|D=1)} f(x|D=1)\, dx \\
&= \int g(x)\, \frac{f(x)\, p}{p(x) f(x)} f(x|D=1)\, dx \\
&= \mathbb{E}\left[\frac{p}{p(X)} g(X) \middle| D=1\right]
\end{aligned}$$

where the second equality holds by multiplying and dividing by $f(x|D=1)$, the third equality holds by applying the definition of conditional probability twice, and the last equality holds by canceling

the $f(x)$ terms and then by the definition of expectation. What this result is saying is that $\mathbb{E}[g(X)]$ can be computed by calculating the mean of $g(X)$ among the treated group *after re-weighting it*. Notice that the term $1/p(X)$ (which will be large when $p(X)$ is small/close to 0 and will be small when $p(X)$ is large/close to 1) will put more "weight" on treated units that have characteristics that are relatively more common among the untreated group and will put less weight on treated units that have characters that are common in the treated group than among the untreated group. Intuitively, you can think of this expression as "balancing" the distribution of covariates for the treated group relative to be the same as the overall distribution of covariates.

Using the same sort of arguments, you can similarly show that

$$\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}\left[\frac{(1-p)}{(1-p(X))}g(X)\middle|D=0\right] \\
\mathbb{E}[g(X)|D=1] &= \mathbb{E}\left[\frac{p(X)}{p}g(X)\right] \\
\mathbb{E}[g(X)|D=1] &= \mathbb{E}\left[\frac{p(X)(1-p)}{(1-p(X))p}g(X)\middle|D=0\right] \\
\mathbb{E}[g(X)|D=0] &= \mathbb{E}\left[\frac{(1-p(X))}{(1-p)}g(X)\right] \\
\mathbb{E}[g(X)|D=0] &= \mathbb{E}\left[\frac{(1-p(X))p}{p(X)(1-p)}g(X)\middle|D=1\right]
\end{aligned}$$

I'll leave showing these results as practice exercises. I'm not sure if I'd recommend memorizing these (you may be able to notice a pattern above), but you can "derive" them using the same arguments as above.

Side-Comment: There's one more technical detail that I ought to mention here. The arguments above additionally require an **overlap condition**. This sort of condition amounts to, for any possible value of the covariates, you need to be able to find both treated and untreated units with those characteristics. In math, we can write the overlap condition as $0 < p(X) < 1$ (the important part is that we rule out $p(X) = 0$ and $p(X) = 1$). In the context of the expressions above, you can see that the overlap condition avoids possible divide by 0 issues in those expressions.

Interpreting α under treatment effect heterogeneity

Using population versions Frisch-Waugh types of arguments (recall that we discussed this earlier in the semester on the last page here), we have that

$$\alpha = \frac{\mathbb{E}[(D - \mathrm{L}(D|X))Y]}{\mathbb{E}[(D - \mathrm{L}(D|X))^2]} \tag{12}$$

In order to understand α in the presence of treatment effect heterogeneity, I am going to provide three results that build on each other.

Decomposition of α 1: α can be decomposed as follows:

$$\alpha = \mathbb{E}\Big[w(D,X)(\mathrm{L}_1(Y|X) - \mathrm{L}_0(Y|X))\Big]$$

where $w(D,X)$ are weights that are given by

$$w(D,X) = \frac{D(1 - \mathrm{L}(D|X))}{\mathbb{E}\Big[(D - \mathrm{L}(D|X))^2\Big]}$$

which have the properties that (i) $\mathbb{E}[w(D,X)] = 1$ and (ii) it is possible that $w(D,X)$ can be negative for some values of D and X.

Proof

Starting with the numerator of Equation 12, notice that

$$\begin{aligned}
\mathbb{E}\Big[(D - \mathrm{L}(D|X))Y\Big] &= \mathbb{E}\Big[(1 - \mathrm{L}(D|X))Y\Big|D = 1\Big]p - \mathbb{E}\Big[\mathrm{L}(D|X)Y\Big|D = 0\Big](1-p) \\
&= \mathbb{E}\Big[(1 - \mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big|D = 1\Big]p - \mathbb{E}\Big[\mathrm{L}(D|X)\mathrm{L}_0(Y|X)\Big|D = 0\Big](1-p) \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big] - \mathbb{E}\Big[(1-D)\mathrm{L}(D|X)\mathrm{L}_0(Y|X)\Big] \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\Big(\mathrm{L}_1(Y|X) - \mathrm{L}_0(Y|X)\Big)\Big] + \mathbb{E}[(D - \mathrm{L}(D|X))\mathrm{L}_0(Y|X)] \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\Big(\mathrm{L}_1(Y|X) - \mathrm{L}_0(Y|X)\Big)\Big]
\end{aligned} \tag{13}$$

where the first equality holds by the law of iterated expectations, the second equality holds by Eq.(7), the third equality holds by the law of iterated expectations, the fourth equality comes from adding and subtracting $\mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\mathrm{L}_0(Y|X)\Big]$ and cancels terms, and the last equality holds because

$$\mathbb{E}[(D - \mathrm{L}(D|X))\mathrm{L}_0(Y|X)] = \mathbb{E}[DX']\beta_0 - \mathbb{E}\Big[\mathbb{E}[DX']\mathbb{E}[XX']^{-1}X(X'\beta_0)\Big] = 0 \tag{14}$$

Plugging Eq.(13) back into Eq.(12) gives the above expression for α. To show that the weights

have mean one, notice that the denominator of the weights is given by

$$\begin{aligned}\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big] &= \mathbb{E}\Big[(D-\mathrm{L}(D|X))D\Big] \\ &= \mathbb{E}\Big[D-D\mathrm{L}(D|X)\Big] \\ &= \mathbb{E}\Big[D(1-\mathrm{L}(D|X))\Big]\end{aligned}$$

where the first equality holds because $(D-\mathrm{L}(D|X))$ is the projection error from the (population) linear projection of D on X which is uncorrelated with $X'\gamma = \mathrm{L}(D|X)$, the second equality holds because $D^2 = D$ (because D is binary), the third equality holds by factoring out D. Notice that the term on the third line is the mean of the numerator of $w(D,X)$. Thus, this implies that $\mathbb{E}[w(D,X)] = 1$. Finally, $w(D,X)$ can be negative for treated units such that $\mathrm{L}(D|X) > 1$.

This first decomposition of α is interesting for a couple of reasons. First, it will be the basis of our interpretation of α under treatment effect heterogeneity. Second, it's easy to compute. In particular, $\hat{\alpha}$ (i.e., the estimated value of α from the regression) will be equal to the sample analogue of the expression on the right hand side of the decomposition, and everything here is easy to estimate — in particular, the linear projection terms can be estimated by just recovering the predicted values from the regression of D on X and the regression of Y on X among the subsets of treated/untreated observations.

Decomposition of α 2: α can be decomposed as

$$\alpha = \mathbb{E}\left[w(D,X)\Big(\mathbb{E}[Y|X,D=1]-\mathbb{E}[Y|X,D=0]\Big)\right] \tag{15}$$

$$+\,\mathbb{E}\left[w(D,X)\Big(\mathbb{E}[Y|X,D=0]-\mathrm{L}_0(Y|X)\Big)\right] \tag{16}$$

Proof

Starting from the expression in the previous decomposition, we have that

$$\begin{aligned}\alpha &= \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X)}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathrm{L}_1(Y|X)-\mathrm{L}_0(Y|X)\Big)\right] \\ &= \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathbb{E}[Y|X,D=1]-\mathbb{E}[Y|X,D=0]\Big)\right] \\ &- \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big\{\Big(\mathbb{E}[Y|X,D=1]-\mathrm{L}_1(Y|X)\Big)-\Big(\mathbb{E}[Y|X,D=0]-\mathrm{L}_0(Y|X)\Big)\Big\}\right] \quad (17)\end{aligned}$$

where the first equality comes from the previous decomposition, and the second equality holds by

adding and subtracting $\mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\left[(D-\mathrm{L}(D|X))^2\right]}\Big(\mathbb{E}[Y|X,D=1]-\mathbb{E}[Y|X,D=0]\Big)\right]$. Next, notice that

$$\begin{aligned}\mathbb{E}\Big[D(1-\mathrm{L}(D|X))\mathbb{E}[Y|X,D=1]\Big] &= \mathbb{E}\Big[(1-\mathrm{L}(D|X))\mathbb{E}[Y|X,D=1]\Big|D=1\Big]\,p \\ &= \mathbb{E}\Big[(1-\mathrm{L}(D|X)Y\Big|D=1\Big]\,p\end{aligned}$$

where both equalities hold by applying the law of iterated expectations. Additionally, notice that

$$\begin{aligned}\mathbb{E}\Big[D(1-\mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big] &= \mathbb{E}\Big[(1-\mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big|D=1\Big]\,p \\ &= \mathbb{E}\Big[(1-\mathrm{L}(D|X))Y\Big|D=1\Big]\,p\end{aligned}$$

where the first equality holds by the law of iterated expectations, and the second equality holds by Eq.(7). That the previous terms are equal to each other implies that

$$\mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathbb{E}[Y|X,D=1]-\mathrm{L}_1(Y|X)\Big)\right]=0$$

and simplifies Eq.(17) to correspond to what is in the decomposition.

Relative to the previous decomposition, this decomposition of α may be more challenging to compute (in particular, the terms like $\mathbb{E}[Y|X,D=d]$ are likely to be challenging to nonparametrically estimate), but this decomposition will be the main basis for our interpretation of α in the next result.

Result on Interpreting α: Suppose that unconfoundedness and overlap both hold. In addition, suppose that either (i) $p(X)=\mathrm{L}(D|X)$ or (ii) $\mathbb{E}[Y|X,D=0]=\mathrm{L}_0(Y|X)$, then

$$\alpha=\mathbb{E}\left[w(D,X)CATE(X)\right]$$

where $w(D,X)$ are defined above and have mean 1. In addition, if condition (i) holds (that $p(X)=\mathrm{L}(D|X)$), then the weights are non-negative.

I am going to leave the proof of this result as an exercise. Given the previous decomposition, it is not too difficult to show; as a hint: under condition (ii), the result holds essentially immediately, but you need to do a bit of work to show that it holds under condition (i).

This result says that, if in addition to unconfoundedness, either of two additional conditions hold, then α will be equal to a weighted average of conditional average treatment effects. Let's discuss the conditions first and then how to interpret this weighted average. Condition (i), that $p(X)=\mathrm{L}(D|X)$, would hold under **linearity of the propensity score**; in other words, the linear probability model is true for D on X. My sense is that you would not generally expect for this to

be true (though perhaps it reasonable to think that it is not "too far" from being true). That said, there is an important leading case where the propensity score is linear is when all of the covariates are discrete and the model is "saturated" in the covariates (i.e., all interactions between covariates are included). Condition (ii) is **linearity of the model for untreated potential outcomes**; this corresponds to "linearity of untreated potential outcomes" that we discussed at the beginning of this section. This condition may or may not hold in practice, though unlike the propensity score, often it would be the case that the most natural model for these conditional expectations is linear. And, perhaps it is reasonable to think that in many applications, the conditional expectations are not "too far" from being linear. This condition would also be satisfied in the case where the covariates are discrete and the model includes the full set of interactions.

Next, it is not immediately obvious whether this is a positive result or not (in fact, different papers on these sort of results often seem to have different opinions about whether this sort of result supports using a regression in this context or not). First, that the result holds under additional linearity conditions is probably not surprising (we are estimating a linear model after all), and I think it is fair to see those conditions as the "price" of using a simple estimation strategy.

The weights are more interesting though (and arguably more troubling). First, notice that, ideally, we'd have that $w(D,X) = 1$ (this would imply that $\alpha = ATE$); in that case, since we are averaging over the distribution of X, $CATE(X)$'s would get more weight for common values of X. This still happens for α for the same reason; however, the weight on a particular $CATE(X)$ also comes from $\mathbb{E}[w(D,X)|X] = \frac{p(X)(1-\mathrm{L}(D|X))}{\mathbb{E}\left[(D-\mathrm{L}(D|X))^2\right]}$. $p(X)(1-\mathrm{L}(D|X))$ is closely related to (but not exactly equal to) $\text{var}(D|X) = p(X)(1-p(X))$. So, roughly, conditional on the "frequency" of a particular value of X, the regression puts more weight on $CATE(X)$'s where there is more variation in treatment status. If treatment effects were homogeneous, this weighting scheme makes sense, but in cases where there is treatment effect heterogeneity, these are at least peculiar weights, in my view. For example, it would be unlikely that a researcher would choose as their target parameter this particular weighted average of conditional average treatment effects. Moreover, this means that α could be far away from the ATE when $CATE(X)$ varies across different values of X.

Along these lines, let's introduce one more assumption: **treatment effect homogeneity across covariates** so that $CATE(X)$ is constant across X (this is slightly weaker than full treatment effect homoegeneity that we had discussed previously, though still likely to be a very strong assumption). If this condition holds in addition to the ones in the previous result, then

$$\alpha = ATE$$

This follows because, in this case, $CATE(X) = ATE$; therefore,

$$\alpha = \mathbb{E}\left[w(D,X)CATE(X)\right] = ATE \times \mathbb{E}[w(D,X)] = ATE$$

where the third equality holds because the weights have the property that $\mathbb{E}[w(D,X)] = 1$.

Alternative Approaches

Now, let's give some alternative approaches that can recover causal effect parameters directly without requiring treatment effect homogeneity assumptions.

At the risk of creating a little bit of confusion, I am going to switch back to targeting ATT at this point. The arguments for ATT are slightly simpler and tend to involve less cumbersome notation (as you can recover $\mathbb{E}[Y|D=1]$ directly). So if you target ATE instead, the expressions below will not be identical, but, conceptually, the arguments will be essentially the same.

Regression adjustment

If we are willing to believe (i) unconfoundedness and (ii) the linear model for untreated potential outcomes, then Equation **??** implies that

$$\begin{aligned} ATT &= \mathbb{E}[Y|D=1] - \mathbb{E}[\mathbb{E}[Y|X,D=0]|D=1] \\ &= \mathbb{E}[Y|D=1] - \mathbb{E}[\mathrm{L}_0(Y|X)|D=1] \\ &= \mathbb{E}[Y|D=1] - \mathbb{E}[X'\beta_0|D=1] \\ &= \mathbb{E}[Y|D=1] - \mathbb{E}[X'|D=1]\beta_0 \end{aligned}$$

where the second equality uses linearity of the model for untreated potential outcomes. This type of expression is called **regression adjustment** as it suggests estimating ATT by running a regression of Y on X using the subset of untreated observations — from this step we have estimated $\widehat{\beta}_0$ — and then computing an estimate of ATT by

$$\widehat{ATT} = \bar{Y}_{D=1} - \bar{X}'_{D=1}\widehat{\beta}_0$$

The case for using regression adjustment relative to the regression in Eq.(6) seems pretty strong to me (that said, in practice, it is substantially less popular). The main benefit is that if the model for untreated potential outcomes is, in fact, linear, then regression adjustment will directly provide an estimate of ATT rather than recovering a peculiar weighted average of some conditional average treatment effects.

Inverse Propensity Score Weighting (IPW)

The regression adjustment strategy above worked for the case when the model for untreated potential outcomes was linear (or least correctly specified). Here, we'll develop an alternative approach based on modeling/estimating the propensity score; this approach won't require linearity of the model for untreated potential outcomes. As a quick intuition, notice that, if the distribution of X were the same across the treated and untreated groups, then (even under unconfoundedness) we could just compute $ATT = \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$. To see this, notice that the challenging term

for identifying the ATT here is

$$\begin{aligned}\mathbb{E}[Y(0)|D=1] &= \mathbb{E}\big[\mathbb{E}[Y|X,D=0]|D=1\big] \\ &= \mathbb{E}\big[\mathbb{E}[Y|X,D=0]|D=0\big] = \mathbb{E}[Y|D=0]\end{aligned}$$

where the first equality holds by unconfoundedness, the second equality (which is the interesting one here) holds only when the distribution of covariates is the same for the treated and untreated group (the outside expectation is over the distribution of covariates for the treated group but can be replaced with the distribution of covariates from the untreated group if these two distributions are the same), and the last equality holds by the law of iterated expectations.

The idea of propensity score weighting will essentially be to weight observations in the untreated group in a way that, in the re-weighted data, they will have the same distribution of covariates as the treated group.

To show this more formally (and, just to be clear, this derivation holds without assuming that the distribution of the covariates is the same for each group), notice that we can write

$$\begin{aligned}\mathbb{E}[Y(0)|D=1] &= \mathbb{E}\Big[\mathbb{E}[Y|X,D=0]|D=1\Big] \\ &= \mathbb{E}\left[\frac{p(X)(1-p)}{p(1-p(X))}\mathbb{E}[Y|X,D=0]\big|D=0\right] \\ &= \mathbb{E}\left[\frac{p(X)(1-p)}{p(1-p(X))}Y\big|D=0\right]\end{aligned}$$

where the first equality holds by unconfoundedness, the second equality holds by the re-weighting results earlier in these notes (and by just viewing $\mathbb{E}[Y|X,D=0]$ as a function of X), and the third equality holds by the law of iterated expectations. This implies that the ATT is identified and that we can recover it by taking the mean outcomes for the treated group relative to a weighted average of outcomes for the untreated group where the weights depend on the propensity score. If you think about these weights, they will be large for untreated units who have characteristics that are relatively common among treated units (so that $p(X)$ is large) and they will be small for untreated units who have characteristics that are relatively uncommon among treated units (so that $p(X)$ is small).

It is common to re-write the expression for the ATT as follows:

$$\begin{aligned}ATT &= \mathbb{E}\left[\frac{D}{p}Y\right] - \mathbb{E}\left[\frac{p(X)(1-p)}{p(1-p(X))}Y|D=0\right] \\ &= \mathbb{E}\left[\frac{D}{p}Y\right] - \mathbb{E}\left[\frac{(1-D)p(X)}{p(1-p(X))}Y\right] \\ &= \mathbb{E}\left[\left(\frac{D}{p} - \frac{(1-D)p(X)}{p(1-p(X))}\right)Y\right] \qquad (18)\end{aligned}$$

where the first line holds from our previous discussion, the second equality holds by the re-weighting results earlier in the notes, and the second line in the expression for ATT holds by the law of iterated

expectations, and the last line holds by combining terms.

Given the expression for ATT in Eq.(18), it suggests estimating ATT by

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\hat{p}} - \frac{(1-D_i)\hat{p}(X_i)}{\hat{p}(1-\hat{p}(X_i))}\right)Y_i$$

where $\hat{p}$ is just the fraction of treated observations in the data, and $\hat{p}(X_i)$ comes from estimating a propensity score model (e.g., leading choices would be logit or probit of the treatment on covariates) and computing predicted values for each X_i in the data.

Notice that the above estimation strategy involves specifying/estimating a model for the propensity score, but side-steps needing to impose a linear model for untreated potential outcomes. This approach is likely to be more attractive than regression adjustment when you feel more confident about correctly specifying a model for the propensity score than for the outcome regression model.

Augmented Inverse Propensity Score Weighting (AIPW)

At this point, you might notice that, relative to α from the regression in Eq.(6), regression adjustment seemed attractive in the case when we knew the right model for untreated potential outcomes and that propensity score re-weighting seemed attractive in the case where we knew the right model for the propensity score. But, our previous results for α, suggested that you could get a weighted average of conditional average treatment effects if *either* the outcome model for untreated potential outcomes was linear or the propensity score was linear. In this section, we will develop an approach can directly target ATT if *either* a model for untreated potential outcomes or for the propensity score is correctly specified.

In particular, one can additionally show that

$$ATT = \mathbb{E}\left[\left(\frac{D}{p} - \frac{(1-D)p(X)}{p(1-p(X))}\right)(Y - \mathbb{E}[Y|X, D=0])\right]$$

This expression is more complicated than the previous ones for the ATT, but it has the very useful property of being **doubly robust**. Recall that the main estimation challenge here is for the the propensity score, $p(X)$, and the outcome regression, $\mathbb{E}[Y|X, D=0]$. The regression adjustment approach that we discussed above will deliver consistent estimates of the ATT if we correctly specify a model for $\mathbb{E}[Y|X, D=0]$ while the propensity score weighting approach will deliver consistent estimates of the ATT if we correctly specify the model for $p(X)$. A doubly robust estimator is one that will deliver consistent estimates of the target parameter (here the ATT) if *either* (but not necessarily both) the propensity score model or the outcome regression model is correctly specified. This gives a researcher two chances to correctly specify a model.

In order to study the properties of this expression for the ATT, it is helpful to re-write it as

$$\begin{aligned} ATT &= \mathbb{E}\left[\frac{D}{p}(Y - \mathbb{E}[Y|X, D=0])\right] - \mathbb{E}\left[\frac{(1-D)p(X)}{p(1-p(X))}(Y - \mathbb{E}[Y|X, D=0])\right] \\ &= \underbrace{\mathbb{E}[Y|D=1] - \mathbb{E}\big[\mathbb{E}[Y|X, D=0]|D=1\big]}_{ATT} - \underbrace{\mathbb{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(Y - \mathbb{E}[Y|X, D=0])\Big|D=0\right]}_{=0 \text{ by LIE}} \end{aligned}$$

Now, let's show that this expression is actually doubly robust. Suppose that we specify parametric models for the propensity score and the outcome regression. Even in cases where these are misspecified for the "true" propensity score and/or outcome regression, if you estimate them, the estimated parameters still converge to "pseudo true values" (i.e., these are just defined as whatever these parameters converge to but allowing for the models to be misspecified). I'll use the notation $p(X;\theta^*)$ to denote the propensity score under some model (e.g., probit) and where θ^* denotes the pseudo true value of the parameter. Likewise, let $m(X,\beta^*)$ denote a parametric model for the outcome regression and where β^* denotes the pseudo true value of the parameter. Given this notation, our estimate of ATT would be given by

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{p} - \frac{(1-D_i)p(X_i,\widehat{\theta})}{p(1-p(X_i;\widehat{\theta}))}\right)(Y_i - m(X_i;\widehat{\beta})) \xrightarrow{p} ATT^*$$

where

$$ATT^* = \mathbb{E}\left[\frac{D}{p}(Y - m(X;\beta^*))\right] - \mathbb{E}\left[\frac{(1-D)p(X;\theta^*)}{p(1-p(X;\theta^*))}(Y - m(X;\beta^*))\right]$$

where ATT^* denotes the corresponding pseudo ATT under the parametric working models for the propensity score and outcome regression. The question is whether or not $ATT^* = ATT$. Next, we will show that $ATT^* = ATT$ if either $p(X;\theta^*) = p(X)$ (i.e., the propensity score working model is correctly specified) or $m(X,\beta^*) = \mathbb{E}[Y|X, D=0]$ (i.e., the outcome regression working model is correctly specified).

Case 1: Outcome Regression Model Correctly Specified In this case, $m(X;\beta^*) = \mathbb{E}[Y|X, D=0]$, but that it could be the case that $p(X;\theta^*) \neq p(X)$. Therefore, the first term in the expression for ATT^* is equal to ATT. For the second term, notice that it is equal to

$$\mathbb{E}\left[\frac{(1-p)p(X;\theta^*)}{p(1-p(X;\theta^*))}(Y - m(X;\beta^*))\Big|D=0\right] = \mathbb{E}\left[\frac{(1-p)p(X;\theta^*)}{p(1-p(X;\theta^*))}\underbrace{\mathbb{E}[(Y - m(X;\beta^*))|X, D=0]}_{=0 \text{ in this case}}\Big|D=0\right]$$

and where the second equality uses the law of iterated expectations. This implies that $ATT^* = ATT$ in this case.

Case 2: Propensity Score Model Correctly Specified In this case, we have that $p(X;\theta^*) = p(X)$, but that it could be the case that $m(X;\beta^*) \neq \mathbb{E}[Y|X, D=0]$. In this case, the first term in

the expression for ATT^* is given by

$$\mathbb{E}[Y|D=1] - \mathbb{E}[m(X,\beta^*)|D=1] \tag{19}$$

which may not be equal to the ATT because $m(X,\beta^*)$ may not be equal to $\mathbb{E}[Y|X,D=0]$. For the second term in the expression for ATT^*, it is given by

$$\begin{aligned}\mathbb{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(Y - m(X;\beta^*))\Big|D=0\right] &= \mathbb{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(\mathbb{E}[Y|X,D=0] - m(X;\beta^*))\Big|D=0\right] \\ &= \mathbb{E}\big[\mathbb{E}[Y|X,D=0]|D=1\big] - \mathbb{E}\big[m(X;\beta^*)|D=1\big] \end{aligned} \tag{20}$$

where the first equality holds by the law of iterated expectations and the second equality switches from integrating over the distribution of X conditional on $D=0$ to integrative over the distribution of X conditional on $D=1$ (as we have done before and which involves re-weighting).

Subtracting Equation 20 from Equation 19 implies that $ATT^* = ATT$ when the model for the propensity score is correctly specified.

AIPW and Machine Learning

Doubly robust estimands often have additional nice properties in estimation. In fact, a main focus of the econometrics literature over the past few years has been to study how **machine learning** approaches, which have been developed primarily for predicting things, can be adapted to be useful for estimating partial effects which are often the objects of interest in research.

This turns out to be quite a tricky problem because most machine learning approaches essentially allow for some bias while reducing the variance of estimates, which can often result in better predictions (particularly in cases where the number of regressors is very large). However, this bias often does not disappear fast enough that we can ignore it and use conventional asymptotic theory / inference arguments.

One promising line of research about partial effects after using machine learning uses (i) doubly robust estimands like the ones we have considered before along with (ii) cross fitting (e.g., sample splitting). We will not do a full treatment of this sort of approach, but let me sketch how you could use machine learning to estimate ATT in the context that we have been considering:

Step 1: Split data into K folds (i.e., groups). K would typically be a relatively small number such as 2 or 5.

Step 2: For the kth fold, estimate $p(X)$ and $\mathbb{E}[Y|X,D=0]$ using all observations that are not in the kth fold. You could use Lasso, ridge regression, random forest, neural nets, etc. for estimating these functions.

Step 3: Use data from the kth fold to compute

$$\widehat{ATT}(k) = \frac{1}{n_k} \sum_{i \in k\text{th fold}} \left(\frac{D_i}{p} - \frac{(1-D_i)\hat{p}(X_i)}{p(1-\hat{p}(X_i))} \right) (Y_i - \hat{m}(X_i))$$

where n_k is the number of observations in the kth fold, $\hat{p}$ and $\hat{m}$ were estimated in Step 2, and $\hat{p}(X_i)$ and $\hat{m}(X_i)$ are just the predicted values of each of these for unit i.

Step 4: Repeat steps 2 and 3 for all K folds. This gives you $\widehat{ATT}(k)$ for each fold.

Step 5: Compute $\widehat{ATT} = \frac{1}{K} \sum_{k=1}^{K} \widehat{ATT}(k)$.

I am not an expert on this front, but machine learning approaches seem promising to me in that, intuitively, they sit somewhere in between parametric models and trying to fully nonparametrically estimate terms like $\mathbb{E}[Y|X, D=0]$.

A useful and (relatively) introductory treatment of using machine learning to estimate partial effects is:

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. The Econometrics Journal, 21(1).

A full treatment of machine learning along these lines is beyond the scope of this class though.

Continuous Treatment

All of our arguments above have been for the case with a binary treatment. I am going to skip the case with a continuous treatment. In general, the continuous treatment case is more complicated than the binary treatment case (although, to my knowledge, it has not been nearly as extensively studied). Intuitively, this suggests that the limitations of regressions would be more severe in this case than in the binary treatment case. If you are interested, a recent relevant paper is:

- Ishimaru, Shoya. "Empirical Decomposition of the IV–OLS Gap with Heterogeneous and Nonlinear Effects." The Review of Economics and Statistics (2022): 1-45.