Alternative Approaches to Causal Inference

These notes do not come from the textbook and are primarily geared towards answering (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?

Interpreting Regressions under Treatment Effect Heterogeneity

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.

For simplicity (and to think things from getting too long), I'll mainly consider the case with a binary treatment. In that context, we made the following three main assumptions:

- **Unconfoundedness:** $(Y(1), Y(0)) \perp\!\!\!\perp D|X$
- **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:

$$Y_i = \alpha D_i + X_i'\beta_0 + e_i \tag{1}$$

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.

As a reminder, recall that ATT is identified in this context:

$$\begin{aligned} ATT &= \mathbb{E}[Y(1) - Y(0)|D = 1] \\ &= \mathbb{E}[Y|D = 1] - \mathbb{E}\Big[\mathbb{E}[Y(0)|X, D = 1]|D = 1\Big] \\ &= \mathbb{E}[Y|D = 1] - \mathbb{E}\Big[\mathbb{E}[Y(0)|X, D = 0]|D = 1\Big] \end{aligned} \tag{2}$$

where the first equality holds by the definition of ATT, the second equality by the law of iterated expectations, and the last equality by unconfoundedness. The last line implies that ATT is identified because we observe untreated potential outcomes for the untreated group. That said, in terms of estimation, as we talked about at length earlier in the semester, $\mathbb{E}[Y|X, D = 0]$ may be challenging to estimate in general without making further assumptions — this is one of the main motivations for running the regression in Equation 1. Using similar arguments, you can show that ATE is also identified here and is given by

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

My impression is that if a researcher was running the regression in Equation 1 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 $p = \mathrm{L}(D = 1)$. And let $p(x) = \mathrm{L}(D = 1|X = x)$ 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:

$$\begin{aligned}\mathbb{E}[\mathrm{L}(D|X)\mathrm{L}_d(Y|X)|D = d] &= \mathbb{E}\Big[(\gamma'X)X'\mathbb{E}[XX'|D = d]^{-1}\mathbb{E}[XY|D = d]\Big|D = d\Big] \\ &= \mathbb{E}[\gamma'XY|D = d] \\ &= \mathbb{E}[\mathrm{L}(D|X)Y|D = d] \qquad (3)\end{aligned}$$

where the second quality holds by moving the nonrandom terms outside the expectation and then canceling terms.

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

expectations implies that

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}\Big[\mathbb{E}[Y|D]\Big] \\ &= \sum_{d\in\{0,1\}} \mathbb{E}[Y|D=d]\mathrm{L}(D=d) \\ &= \mathbb{E}[Y|D=1]p + \mathbb{E}[Y|D=0](1-p) \end{aligned} \tag{4}$$

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

$$\mathbb{E}[Y|D=1] = \mathbb{E}\left[\frac{D}{p}Y\right] \qquad \text{and} \qquad \mathbb{E}[Y|D=0] = \mathbb{E}\left[\frac{1-D}{1-p}Y\right] \tag{5}$$

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

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, D=1]p(X) + \mathbb{E}[Y|X, D=0](1-p(X)) \tag{6}$$

and, using the same sort of arguments as above

$$\mathbb{E}[DY|X] = \mathbb{E}[Y|X, D=1]p(X) \tag{7}$$

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

$$\alpha = \frac{\mathbb{E}[(D - \mathrm{L}(D|X))Y]}{\mathbb{E}[(D - \mathrm{L}(D|X))^2]} \tag{8}$$

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 8, notice that

$$\begin{aligned}
\mathbb{E}\Big[(D - \mathrm{L}(D|X))Y\Big] &= \mathbb{E}\Big[(1 - \mathrm{L}(D|X))Y\Big|D = 1\Big]p - \mathbb{E}\Big[\mathrm{L}(D|X)Y\Big|D = 0\Big](1-p) \\
&= \mathbb{E}\Big[(1 - \mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big|D = 1\Big]p - \mathbb{E}\Big[\mathrm{L}(D|X)\mathrm{L}_0(Y|X)\Big|D = 0\Big](1-p) \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\mathrm{L}_1(Y|X)\Big] - \mathbb{E}\Big[(1-D)\mathrm{L}(D|X)\mathrm{L}_0(Y|X)\Big] \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\Big(\mathrm{L}_1(Y|X) - \mathrm{L}_0(Y|X)\Big)\Big] + \mathbb{E}[(D - \mathrm{L}(D|X))\mathrm{L}_0(Y|X)] \\
&= \mathbb{E}\Big[D(1 - \mathrm{L}(D|X))\Big(\mathrm{L}_1(Y|X) - \mathrm{L}_0(Y|X)\Big)\Big] \qquad (9)
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds by Eq.(3), 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

$$\mathbb{E}[(D-\mathrm{L}(D|X))\mathrm{L}_0(Y|X)] = \mathbb{E}[DX']\beta_0 - \mathbb{E}\Big[\mathbb{E}[DX']\mathbb{E}[XX']^{-1}X(X'\beta_0)\Big] = 0 \tag{10}$$

Plugging Eq.(9) back into Eq.(8) 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, $\widehat{\alpha}$ (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

$$\alpha = \mathbb{E}\left[w(D,X)\Big(\mathbb{E}[Y|X,D=1] - \mathbb{E}[Y|X,D=0]\Big)\right] \tag{11}$$

$$+\,\mathbb{E}\left[w(D,X)\Big(\mathbb{E}[Y|X,D=0] - \mathrm{L}_0(Y|X)\Big)\right] \tag{12}$$

Proof

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

$$\begin{aligned}
\alpha &= \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X)}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathrm{L}_1(Y|X)-\mathrm{L}_0(Y|X)\Big)\right] \\
&= \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathbb{E}[Y|X,D=1]-\mathbb{E}[Y|X,D=0]\Big)\right] \\
&- \mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big\{\Big(\mathbb{E}[Y|X,D=1]-\mathrm{L}_1(Y|X)\Big)-\Big(\mathbb{E}[Y|X,D=0]-\mathrm{L}_0(Y|X)\Big)\Big\}\right] \quad (13)
\end{aligned}$$

where the first equality comes from the previous decomposition, and the second equality holds by adding and subtracting $\mathbb{E}\left[\frac{D(1-\mathrm{L}(D|X))}{\mathbb{E}\Big[(D-\mathrm{L}(D|X))^2\Big]}\Big(\mathbb{E}[Y|X,D=1]-\mathbb{E}[Y|X,D=0]\Big)\right]$. 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.(3). 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.(13) 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) $\mathrm{var}(D|X) = p(X)(1 - p(X))$. 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 2 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 $\hat{\beta}_0$ — and then computing an estimate of ATT by

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

The case for using regression adjustment relative to the regression in Eq.(1) 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.

Propensity Score Weighting

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:

$$\begin{aligned} ATT &= \mathbb{E}\left[\frac{D}{p}Y\right] - \mathbb{E}\left[\frac{p(X)(1-p)}{p(1-p(X))}Y|D=0\right] \\ &= \mathbb{E}\left[\frac{D}{p}Y\right] - \mathbb{E}\left[\frac{(1-D)p(X)}{p(1-p(X))}Y\right] \\ &= \mathbb{E}\left[\left(\frac{D}{p} - \frac{(1-D)p(X)}{p(1-p(X))}\right)Y\right] \end{aligned} \tag{14}$$

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.(14), 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.

Doubly Robust

At this point, you might notice that, relative to α from the regression in Eq.(1), 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

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

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

$$\mathbb{E}[Y|D=1] - \mathbb{E}[m(X,\beta^*)|D=1] \tag{15}$$

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

$$\begin{aligned}\mathbb{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(Y - m(X;\beta^*))\Big|D=0\right] &= \mathbb{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(\mathbb{E}[Y|X, D=0] - m(X;\beta^*))\Big|D=0\right] \\ &= \mathbb{E}\big[\mathbb{E}[Y|X, D=0]|D=1\big] - \mathbb{E}\big[m(X;\beta^*)|D=1\big] \end{aligned} \tag{16}$$

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 16 from Equation 15 implies that $ATT^* = ATT$ when the model for the propensity score is correctly specified.

Doubly Robust 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.