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:** $Y(0) \perp D|X$
- **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:

$$Y_i = \alpha D_i + X_i'\beta + e_i$$

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:

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

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, $\mathrm{E}[Y|X, D = 0]$ may be challenging to estimate in general without making further assumptions.

We will use the following notation. 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 $X = x$.

For the arguments below about interpreting regressions, it also will be helpful to write down some notation for the linear projection of Y on X, and the linear projection of D on X:

$$Y = X'\gamma_Y + u_Y$$
$$D = X'\gamma_D + u_D$$

Sometimes, I'll also use notation like $\mathrm{L}(Y|X)$ to denote the linear projection of Y on X (so that, in this context, $\mathrm{L}(Y|X) = X'\gamma_Y$).

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

$$\alpha = \frac{\mathrm{E}[Du_Y]}{\mathrm{E}[u_D^2]}$$

We would like to relate α to ATT along with any additional other terms that show up in the presence of treatment effect heterogeneity.

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

$$\mathrm{E}[Y] = \mathrm{E}[Y|D=1]p + \mathrm{E}[Y|D=0](1-p) \tag{2}$$

A similar argument implies that

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

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

$$\mathrm{E}[Y|D=1] = \mathrm{E}\left[\frac{D}{p}Y\right] \qquad \text{and} \qquad \mathrm{E}[Y|D=0] = \mathrm{E}\left[\frac{1-D}{1-p}Y\right] \tag{3}$$

Intuitively, you can think the following: $\mathrm{E}[DY]$ 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 $\mathrm{E}[Y|D=1]$, 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

$$\mathrm{E}[Y|X] = \mathrm{E}[Y|X, D=1]p(X) + \mathrm{E}[Y|X, D=0](1-p(X)) \tag{4}$$

and, using the same sort of arguments as above

$$\mathrm{E}[DY|X] = \mathrm{E}[Y|X, D=1]p(X) \tag{5}$$

Let's return to interpreting α. To start with, let's consider the numerator. We have that

$$\begin{aligned}
\mathrm{E}[Du_Y] &= \mathrm{E}[D(Y - \mathrm{L}(Y|X))] \\
&= \mathrm{E}[D(Y - \mathrm{E}[Y|X])] + \mathrm{E}[D(\mathrm{E}[Y|X] - \mathrm{L}(Y|X))] \\
&=: A + B
\end{aligned}$$

where the first equality holds by the definition of u_Y, and the second equality just adds and subtracts $\mathrm{E}\big[D\mathrm{E}[Y|X]\big]$.

Now, let's further consider each of these terms.

$$\begin{aligned}
A &= \mathrm{E}\Big[\mathrm{E}[DY|X] - p(X)\mathrm{E}[Y|X]\Big] \\
&= \mathrm{E}\left[\mathrm{E}[Y|X, D=1]p(X) - p(X)\Big(\mathrm{E}[Y|X, D=1]p(X) + \mathrm{E}[Y|X, D=0](1-p(X))\Big)\right] \\
&= \mathrm{E}\left[p(X)(1-p(X))\Big(\mathrm{E}[Y|X, D=1] - \mathrm{E}[Y|X, D=0]\Big)\right] \\
&= \mathrm{E}\left[p(1-p(X))\Big(\mathrm{E}[Y|X, D=1] - \mathrm{E}[Y|X, D=0]\Big)\Big|D=1\right] \\
&= \mathrm{E}\left[p(1-p(X))ATT(X)\Big|D=1\right]
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds by Equations 5 (for the first term) and 4 (for the second term), the third equality holds by rearranging terms. To see the fourth equality, notice that we are averaging a function (for simplicity called g below) over X, which we can write as

$$\begin{aligned}
\int g(x)\, f(x)\, dx &= \int g(x)\, \frac{f(x)}{f(x|D=1)} f(x|D=1)\, dx \\
&= \int g(x)\, \frac{f(x)\mathrm{P}(D=1)}{\mathrm{P}(D=1|X=x)f(x)} f(x|D=1)\, dx
\end{aligned}$$

where the second equality holds by applying the definition of conditional probability twice. The fifth equality holds by unconfoundedness (and where we define $ATT(X) = \mathrm{E}[Y(1) - Y(0)|X, D=1]$).

For term B, first, there is a law of iterated projections (broadly similar to the law of iterated expectations) that implies that:

$$\mathrm{L}(Y|X) = \mathrm{L}(Y|X, D=1)\mathrm{L}(D|X) + \mathrm{L}(Y|X, D=0)(1 - L(D|X)) \tag{6}$$

which we use below.

$$
\begin{aligned}
B &= \mathrm{E}\Big[\mathrm{E}[Y|X] - \mathrm{L}(Y|X)\Big|D=1\Big]p \\
&= \mathrm{E}\Big[\Big(\mathrm{E}[Y|X,D=1]p(X) + \mathrm{E}[Y|X,D=0](1-p(X)\Big) \\
&\quad - \Big(\mathrm{L}(Y|X,D=1)\mathrm{L}(D|X) + \mathrm{L}(Y|X,D=0)(1-L(D|X))\Big)\Big|D=1\Big]\,p \\
&= \mathrm{E}\Big[\Big(\mathrm{E}[Y|X,D=1] - \mathrm{L}(Y|X,D=1)\Big)p(X) + \Big(\mathrm{E}[Y|X,D=0] - \mathrm{L}(Y|X,D=0)\Big)(1-p(X)) \\
&\quad - \Big(\mathrm{L}(Y|X,D=1) - \mathrm{L}(Y|X,D=0)\Big)(\mathrm{L}(D|X) - p(X))\Big|D=1\Big]\,p
\end{aligned}
$$

where the first equality holds by the same sort of law of iterated expectations argument as for Equations 2 and 3, the second equality holds by the law of iterated expectations (Equation 4) and the law of iterated projections (Equation 6), and the last equality holds by adding and subtracting $\mathrm{L}(Y|X,D=1)p(X)$ and $\mathrm{L}(Y|X,D=0)(1-p(X))$ and rearranging terms.

Finally, consider the denominator in the expression for α. You can show that $\mathrm{E}[D\mathrm{L}(D|X)] = \mathrm{E}[\mathrm{L}(D|X)^2]$ (by just plugging in the definition of $\mathrm{L}(D|X)$ and simplifying), and therefore

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

Let's put all of this back together. We now have that

$$
\begin{aligned}
\alpha &= \mathrm{E}[w_1(X)ATT(X)|D=1] && (7)\\
&\quad + \mathrm{E}[w_{2a}(X)(\mathrm{E}[Y|X,D=1] - \mathrm{L}(Y|X,D=1))|D=1] && (8)\\
&\quad + \mathrm{E}[w_{2b}(X)(\mathrm{E}[Y|X,D=0] - \mathrm{L}(Y|X,D=0))|D=1] && (9)\\
&\quad + \mathrm{E}[w_3(X)(\mathrm{L}(Y|X,D=1) - \mathrm{L}(Y|X,D=0))|D=1] && (10)
\end{aligned}
$$

where

$$
\begin{aligned}
w_1(X) &= \frac{1-p(X)}{\mathrm{E}[1-\mathrm{L}(D|X)|D=1]} \\
w_{2a}(X) &= \frac{p(X)}{\mathrm{E}[1-\mathrm{L}(D|X)|D=1]} \\
w_{2b}(X) &= \frac{1-p(X)}{\mathrm{E}[1-\mathrm{L}(D|X)|D=1]} \\
w_3(X) &= \frac{p(X)-\mathrm{L}(D|X)}{\mathrm{E}[1-\mathrm{L}(D|X)|D=1]}
\end{aligned}
$$

This is an interesting, but perhaps confusing result. Equation 7 is roughly a weighted average of conditional ATT's. Equations 8 and 9 can be thought of as bias terms due when those conditional expectations are potentially nonlinear. Equation 10 amounts to a bias term coming from possible nonlinearity of the propensity score.

The main takeaway at this point is that α is not generally equal to ATT under treatment effect heterogeneity.

Let's make some additional assumptions to try to make some progress on understanding this a bit more. **Linearity of the propensity score** $p(X) = \mathrm{L}(D|X)$. 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 a a 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). Anyway, under linearity of the propensity score, the weights simplify:

$$\begin{aligned} w_1(X) &= \frac{1-p(X)}{\mathrm{E}[1-p(X)|D=1]} \\ w_{2a}(X) &= \frac{p(X)}{\mathrm{E}[1-p(X)|D=1]} \\ w_{2b}(X) &= \frac{1-p(X)}{\mathrm{E}[1-p(X)|D=1]} \\ w_3(X) &= 0 \end{aligned}$$

so that there is 0 weight on the term in Equation 10. Moreover, $\mathrm{E}[w_1(X)|D=1] = \mathrm{E}[w_{2b}(X)|D=1] = 1$ (so that these weights have mean 1, which is typically a good property for weights to have). $\mathrm{E}[w_{2a}(X)|D=1] = \mathrm{E}[p(X)|D=1]/(1-\mathrm{E}[p(X)|D=1])$ so this term would tend to get a large amount of weight when $\mathrm{E}[p(X)|D=1]$ is close to 1 (in general, these weights don't have mean 1 except in the case where $\mathrm{E}[p(X)|D=1] = 1/2$).

Now, let's make the additional assumption of **linearity of conditional expectations of outcomes**: for $d \in \{0,1\}$, $\mathrm{E}[Y|X, D=d] = \mathrm{L}(Y|X, D=d)$. Notice that when $d=0$, this corresponds to "linearity of untreated potential outcomes" that we discussed at the beginning of this section, but also needs to hold for $d=1$. 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. In this case, the bias terms in Equation 8 and 9 are equal to 0.

Under these two assumptions, $\alpha = \mathrm{E}[w_1(X)ATT(X)|D=1]$. Interestingly, even in this case (which already involves "extra" assumptions), α still suffers from what is sometimes called a **weight reversal** property. In particular, notice that the largest weights on $ATT(X)$ are for values of X where $p(X)$ is small — this means that these are values of the covariates that are relatively

uncommon for the treated group relative to the untreated group. Likewise, the smallest weights are put on values of X where $p(X)$ is small; that is, values of the covariates that are relatively common for the treated group. These are peculiar/undesirable weights in my view. And, in particular, this means that α could be far away from the ATT when $ATT(X)$ varies across different values of X.

Along these lines, let's introduce one more assumption: **treatment effect homogeneity across covariates** so that $ATT(X)$ is constant across X. Under all three of the additional conditions above,

$$\alpha = ATT$$

Alternative Approaches

Now, let's give some alternative approaches that can recover the ATT under weaker assumptions

Regression adjustment

If we are willing to believe (i) unconfoundedness and (ii) the linear model for untreated potential outcomes (both of which we'd already need to believe for the regression to work), then Equation 1 implies that

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

For thinking about the asymptotic theory of this type of estimator, it is sometimes useful to re-write this as

$$ATT = \mathrm{E}\left[\frac{D}{p}Y\right] - \mathrm{E}\left[\frac{D}{p}X'\right]\beta$$

which can be estimated by

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}X_i'\hat{\beta}$$

and where $\hat{\beta}$ comes from the regression of Y on X using untreated observations only.

Propensity Score Weighting

In some cases, you might not want to use the linear model for untreated potential outcomes. Here, we'll show that an alternative can be to model/estimate the propensity score. 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 $ATT = \mathrm{E}[Y|D=1] - \mathrm{E}[Y|D=0]$ (make sure

that you understand why this is the case). 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, notice that we can write

$$\begin{aligned} ATT &= \mathrm{E}\left[\frac{D}{p}Y\right] - \mathrm{E}\left[\frac{p(X)(1-p)}{p(1-p(X))}Y|D=0\right] \\ &= \mathrm{E}\left[\frac{D}{p}Y\right] - \mathrm{E}\left[\frac{(1-D)p(X)}{p(1-p(X))}Y\right] \\ &= \mathrm{E}\left[\left(\frac{D}{p} - \frac{(1-D)p(X)}{p(1-p(X))}\right)Y\right] \end{aligned} \tag{11}$$

where the second part of the first line holds because you can think of $\mathrm{E}[Y|X, D=0]$ inside the expression for ATT in Equation 1 as a function of X and because

$$\begin{aligned} \int g(x)f(x|D=1)\,dx &= \int g(x)\frac{f(x|D=1)}{f(x|D=0)}f(x|D=0)\,dx \\ &= \int g(x)\frac{p(x)}{p}\frac{1-p}{1-p(x)}\frac{f(x)}{f(x)}f(x|D=0)\,dx \end{aligned}$$

and the second line in the expression for ATT holds by the law of iterated expectations, and the last equality holds just by combining terms. You can think of this expression as showing that ATT is equal to the average Y among the treated group adjusted by a weighted average of Y among the untreated group (due to the $(1-D)$ term) where the weights are driven by the propensity score and more weight is given to untreated units with a high propensity score (indicating that they have characteristics that are relatively more common among the treated group).

Side-Comment: Notice that the denominator in the expression for ATT involves $1-p(X)$. In order to avoid a divide by 0, we need to the assumption that $p(x) < 1$ for all possible values of x. This is an **overlap condition**, and intuitively it means that there are no values of the covariates where are all units with those covariates are treated; alternatively: for any treated unit, we can always find untreated "matches" with the same covariates X.

Given the expression for ATT in Equation, 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

Finally, one can additionally show that

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

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, $\mathrm{E}[Y|X, D=0]$. The regression adjustment approach that we discussed above will deliver consistent estimates of the ATT if we correctly specify a model for $\mathrm{E}[Y|X, D=0]$ 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

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

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

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

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 $m(X,\beta^*) = \mathrm{E}[Y|X, D=0]$ (i.e., the outcome regression working model is correctly specified).

Case 1: Outcome Regression Model Correctly Specified In this case, $m(X;\beta^*) = \mathrm{E}[Y|X, D=0]$, 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

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

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 $m(X;\beta^*) \neq \mathrm{E}[Y|X, D=0]$. In this case, the first term in the expression for ATT^* is given by

$$\mathrm{E}[Y|D=1] - \mathrm{E}[m(X,\beta^*)|D=1] \tag{12}$$

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

$$\begin{aligned}\mathrm{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(Y - m(X;\beta^*))\Big|D=0\right] &= \mathrm{E}\left[\frac{(1-p)p(X)}{p(1-p(X))}(\mathrm{E}[Y|X, D=0] - m(X;\beta^*))\Big|D=0\right] \\ &= \mathrm{E}\big[\mathrm{E}[Y|X, D=0]|D=1\big] - \mathrm{E}\big[m(X;\beta^*)|D=1\big] \end{aligned} \tag{13}$$

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 13 from Equation 12 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 $\mathrm{E}[Y|X, D = 0]$ 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 $\mathrm{E}[Y|X, D = 0]$.

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

Because we are running out of time this semester, I am going to (at least for the moment) skip the case with a continuous treatment. In general, the continuous treatment case is more complicated than the binary treatment case. Intuitively, this suggests that the limitations of regressions would be more severe in this case than in the binary treatment case.