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).

Linear Regression Notes 1: Why Linear Regression?

H 2.18, H 2.11

We will spend a lot of time this semester studying properties of β defined as

$$\beta = \arg\min_b \mathrm{E}[(Y - X'b)^2]$$

In the preliminary notes, we showed that (under some regularity conditions) this had the solution

$$\beta = \mathrm{E}[XX']^{-1}\mathrm{E}[XY]$$

Moreover, in the preliminary notes, we provided some traditional reasons to be interested in β:

1. $X'\beta$ is the best linear predictor of Y given X (see H 2.18)
2. If we additionally (somehow) know that $\mathrm{E}[Y|X] = X'\beta$, then $X'\beta$ is the best predictor of Y given in X (see H 2.11)

Both of these are properties related to making predictions. This is interesting/useful in lots of contexts. For example, suppose that you work as an appraiser and want to predict how much a house will sell for, this suggests that you could run a regression of the price that houses sell for on their characteristics (e.g., number of square feet, number of bathrooms, etc). This would give you an estimate of $\hat{\beta}$. Suppose that you wanted to predict the selling price of a house with characteristics x; the above results suggest that $x'\hat{\beta}$ should be a good prediction (relative to trying to use the same information in some other way to make a prediction).

These sorts of prediction problems are extremely common (you can especially imagine that this important in numerous business/tech applications), and there have been major advancements in using data to make predictions over the past 20-30 years.

That said, most research in economics (and social sciences, business fields, etc.) is not so much interested in predictions, *per se*. For example, my secondary research interest is in labor economics. In labor economics, there is tons of work where the outcome is a person's earnings. However, I have never seen any researcher who was primarily interested in taking a person's characteristics (say, their years of education, demographic characteristics, etc.) and making predictions about what their earnings will be.

Instead, much research in labor economics concerns the effects of different policies (e.g., minimum wage policies) or other interventions (e.g., a person participating in a union, going to college, or losing their job) on earnings. Learning about "effects" is related to making predictions (and we'll see that

most of the main tools for prediction are also useful for evaluating effects of policies/interventions), but there are also some subtle and important differences.

Regression Derivatives

H 2.14

Following the textbook, we'll use the shorthand notation $m(x) := \mathrm{E}[Y|X = x]$. We will often be interested in the **regression derivative**. An example of a regression derivative is

$$\frac{\partial \, \mathrm{E}[Y|X = x]}{\partial \, x_1}$$

which holds when x_1 is continuously distributed. This derivative should be interpreted as how much Y changes, on average, when x_1 increases by one unit holding the other regressors constant.

You can also define a regression derivative when X_1 is discrete. For example, suppose that X_1 is binary (so it only takes the value 0 or 1), then the regression derivative is given by

$$\mathrm{E}[Y|X_1 = 1, X_2 = x_2, \ldots, X_k = x_k] - \mathrm{E}[Y|X_1 = 0, X_2 = x_2, \ldots X_k = x_k]$$

You could similarly define a regression derivative for the case where X_1 was discrete but took more possible values.

In order to unify notation, we write

$$\nabla_1 m(x) := \begin{cases} \frac{\partial \, \mathrm{E}[Y|X=x]}{\partial x_1} & \text{if } x_1 \text{ is continuous} \\ \mathrm{E}[Y|X_1 = 1, X_2 = x_2, \ldots, X_k = x_k] - \mathrm{E}[Y|X_1 = 0, X_2 = x_2, \ldots X_k = x_k] & \text{if } x_1 \text{ is binary} \end{cases}$$

There is nothing unique about defining partial effects for just X_1, and we can likewise define partial effects for $X_2, \ldots, X_k$, for example, $\nabla_2 m(x)$ is the partial effect of X_2.

The regression derivatives above are also sometimes called the "partial effect" of x_1 or the "marginal effect" of x_1.

Some Comments

- First, partial effects hold other regressors constant. But they do not hold other variables that are not in the model constant.
- Second, you should notice that $\nabla_1 m(x)$ is a function of x. If you plug in different values of x, then the value of this function could change. For example, if you take X_1 to be a binary variable indicating whether or not an individual attended college, Y to be their earnings, and X_2 to be a person's age, you could imagine that the partial effect of college differs depending on a person's age.
- Third, partial effects are really about averages rather than individual-level effects. Continuing the example of the return to going to college – you can easily imagine that, holding age

constant, the effect of going to college on a person's earnings may vary (perhaps tremendously across different people). The regression derivative averages over all of these individual-level effects while holding age constant.

Under the linear CEF model where $m(x) = x'\beta$, $\nabla_1 m(x) = \beta_1$ (up to cases where there are included interaction terms, quadratic terms, etc.).

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. 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:

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

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

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

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 (at least in terms of their potential outcomes).

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

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

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 $\mathrm{E}[Y(0)|D=1] = \mathrm{E}[Y(0)|D=0]$. 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):

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

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?

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

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 $\mathrm{E}[e|D] = 0$, 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, given the regression in Equation 4,

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

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

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

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):

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

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 $Y(0) \perp D|X$ 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

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

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.

And, in particular, the above result implies that ATT is equal to the mean actual outcomes of the treated group adjusted by the the mean outcomes for the treated group conditional on X, but then averaged over the distribution of X for the treated group.

For 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, the first term in the ATT of going to college is equal to the mean earnings of those that went to college. The inside part of the second term, $\mathrm{E}[Y|X,D=0]$, is the average earnings of those that did not go to college conditional on their parents' income (by assumption this is equal to the average earnings of people that (i) went to college and (ii) have the same value of parents' income *would have experienced* if they had not gone to college), and the outside expectation averages the conditional expectation 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.

Although the previous result implies 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 $\mathrm{E}[Y|X,D=0]$ 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. First, let's assume a 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 $\mathrm{E}[Y(0)|X,D=1] = \mathrm{E}[Y(0)|X,D=0]$ which (given linearity) implies that $\mathrm{E}[e|X,D=d]=0$ 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. 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. One sort of parameters are **level effects** such as

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

These are quite similar to ATT and ATE that we talked about in the case with a binary treatment. $ATT(d)$ is the average difference between potential outcomes under dose d relative to untreated potential outcomes among those that actually experienced dose d. $ATE(d)$ 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: Another interesting target parameter would be a derivative of $ATT(d)$, though this is somewhat conceptually harder to think about. In particular, let's expand the notation above to define

$$ATT(d|d') = \mathrm{E}[Y(d) - Y(0)|D = d']$$

so that this is the average difference between potential outcomes under dose d relative to untreated potential outcomes among those that experienced dose d' — which breaks the connection between the dose for the potential outcomes and the dose being conditioned on. Then, one can define the **average causal response on the treated**

$$ACRT(d|d') := \frac{\partial\, ATT(l|d')}{\partial\, l}\Big|_{l=d}$$

This is the causal effect of a marginal increase in the treatment (relative to dose d) among those that actually experienced dose d'.
At the cost of somewhat stronger assumptions (in some cases), we'll mostly target $ACR(d)$, mostly for simplicity.

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

$$ACR^O := \mathrm{E}[ACR(D)|D > 0]$$

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 $d \in \mathcal{D}$,

$$Y(d) \perp D$$

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,

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

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,

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

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

Practice: Show that $ATT(d)$ is identified under random assignment and provide an expression for it.

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 $\mathrm{E}[Y|D=0]$ is easy to estimate — just subset the data down to untreated observations and calculate their average outcome. However, when the treatment is continuously distributed, $\mathrm{E}[Y|D=d]$ 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 $\mathrm{E}[e|D = d] = 0$.

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. This discussion suggests (in the case where the amount of the treatment is randomly assigned and under treatment effect homogeneity) 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 $d \in \mathcal{D}$,

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

> **Practice:** Show that $ATE(d)$, $ATT(d)$, 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 $ATE(d)$, $ATT(d)$, and $ACR(d)$ all depend on terms like $\mathrm{E}[Y|X, D = d]$ (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 $\mathrm{E}[e|X] = 0$. Unconfoundedness additionally implies that $\mathrm{E}[e|X, D = d] = 0$ 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.