Panel Data

These notes cover (i) traditional panel data models, (ii) why these sorts of models can be useful in the context of causal research, and (iii) alternative approaches that are more robust to treatment effect heterogeneity. Some of the material in these notes comes from Chapters 17 and 18 in the textbook, but other parts are not available in the textbook.

Motivation

For thinking about causal effect, this semester we have typically relied on unconfoundedness assumptions. For this section, I want to continue to talk about unconfoundedness, but a slightly altered version of it. In particular, suppose that you are willing to believe the following assumption:

Unconfoundedness: $Y(0) \perp D | (X, W)$

This is exactly the same sort of setup that we have considered before except that I am splitting the variables that we need to condition on into X and W. And, in particular, let us now consider the case where X are observed in our data while W are not observed.

In my view, when you think about an unconfoundedness type of assumption, you ought to do it before you see what's in available in your data. Then, given your *ex ante* reasoning/model, you can check which "covariates" are actually available in your data and which are not. A classic example along these lines though is in labor economics where a researcher is studying the effect of some treatment and thinks that unconfoundedness holds after conditioning on a person's "ability" or "motivation" (both of which are hard to measure though it seems reasonable to expect that they affect lots of different individual-level outcomes). If you are studying industrial organization, a firm may have latent (unobserved) productivity that might be important to condition on. If you are studying ag econ, a particular location's soil fertility may be unobserved but important to condition on. You can probably come up with other sorts of examples along these lines. Just so we have something concrete to think about in this section, I'll use the running example of a researcher who wants to study the effect of job displacement (this essentially just means getting laid off from a job) on a person's earnings.

For simplicity, let's consider the case where there are not any observed covariates that show up in the unconfoundedness assumption; that is, $Y(0) \perp D | W$ — we'll just do this for now because it is relatively straightforward to account for observed covariates, and the main complication is due to W. Let's also suppose that we are willing to invoke the extra assumption of linearity of the model for untreated potential outcomes; that is,

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

Linearity + unconfoundedness implies that $\mathrm{E}[e|W, D] = 0$. Moreover (following arguments we have used several times before),

$$\begin{aligned} ATT &= \mathrm{E}[Y|D=1] - \mathrm{E}[Y(0)|D=1] \\ &= \mathrm{E}[Y|D=1] - \mathrm{E}[W'\beta + e|D=1] \\ &= \mathrm{E}[Y|D=1] - \mathrm{E}[W'|D=1]\beta \end{aligned} \tag{1}$$

and where β would come from the regression of Y on W using the untreated group only. This is exactly the same as we have done before except that this strategy is now *infeasible* — that is, we cannot hope to estimate β or $\mathrm{E}[W|D=1]$ using the available data because W is not observed.

Side-Comment: The above issues are related to the issue of omitted variable bias that we talked about earlier in the semester. Consider the feasible comparison of means in outcomes between the treated group and untreated group:

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

where the first equality plugs in for $\mathrm{E}[Y|D=1]$ using Equation (1) and by the law of iterated expectations for the second term, the second equality holds by linearity of untreated potential outcomes, and the last equality holds by rearranging terms.

This suggests that the comparison of means (which ignores W) is equal to the ATT plus a leftover term that is not generally equal to 0. Like omitted variable bias, the second term *can* be equal to 0 if either (i) $\mathrm{E}[W|D=1] = \mathrm{E}[W|D=0]$ (i.e., that the mean of W is the same across groups), or (ii) $\beta = 0$ (i.e., that W has no effect on untreated on potential outcomes). That said, we typically do not have a good way of checking whether the mean of W is the same across groups, and we'd probably be unlikely to have thought that W would show up in the unconfoundedness assumption if we thought β were equal to 0 (and we don't have the data where we could test this either).

Using Panel Data (2 period case)

Now, let's consider the case where we observe two periods of **panel data**. In particular, suppose that we observe $\{Y_{i1}, Y_{i2}, D_{i1}, D_{i2}\}_{i=1}^n$ which are iid across units, and where the second index denotes the time period. For example, D_{i2} indicates whether or not unit i was treated in period 2. As additional notation, sometimes I'll write Y_{it} as a generic way to indicate the outcome for unit i in time period t. Also, let's define the **first difference** operation as $\Delta Y_{it} = Y_{it} - Y_{it-1}$. Finally, let's also define $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$. Similar notation along these lines applies for D_{it}.

When there are more time periods, we also need to think more carefully about our notion of potential outcomes. In particular, let $\mathbf{d}$ denote a 2×1 vector where each element is either 0 or 1. Then, let $Y_{it}(\mathbf{d})$ denote the potential outcome in time period t for unit i under treatment "regime" $\mathbf{d}$. For example $Y_{i2}(0,0)$ is the outcome that unit i would experience if it was not treated in either time period. Also, let $\mathbf{Y}_i(\mathbf{d}) = (Y_{i1}(\mathbf{d}), Y_{i2}(\mathbf{d}))'$.

This notation/setup can be quite cumbersome (it is already fairly cumbersome, but we'll want to consider the case with more time periods later where all of this will start to explode in notational complexity). In order to circumvent this, I am going to suppose that there is **staggered treatment adoption**; that is, $D_{i1} = 1 \implies D_{i2} = 1$; that is, once a unit becomes treated, it remains treated. With two time periods, this means that there are no units that follow the treatment path $(1,0)$. When there are more periods, this condition will eliminate more treatment paths. I have a longer discussion below of staggered treatment adoption in the case with multiple periods, but the main practical benefit is that we can fully summarize a unit's entire path of participating in the treatment by it's "group" – the time period when it becomes treated. Here, we have 3 possible groups: group 1, group 2, and the never-treated group. I'll use the notation for G_i.

This is sort of ambiguously defined for units that do not participate in the treatment in any time period (the never-treated group). For the never-treated group, I'll set $G_i = T + 1$ where T is the total number of time periods; thus, with two time periods, $G_i = 3$ for the never-treated group. It's also convenient to define the variable U_i indicating whether or not a unit is in the never-treated group; $U_i = 1$ for units in the never-treated group, and $U_i = 0$ for units that participate in the treatment in any time period.

Next, we'll define potential outcomes $Y_{it}(g)$ as the outcome that unit i would experience in time period t if they were in group g. We'll also write $Y_{it}(0)$ for the outcome that unit i would experience in time period t if they did not participate in the treatment in any time period. And we'll define $\mathbf{Y}_i(0)$ as the entire vector of untreated potential outcomes and $\mathbf{Y}_i(g)$ as the vector of potential outcomes for unit i under group g. The outcomes that we observe are $\mathbf{Y}_i = \mathbf{Y}_i(G_i)$ (the potential outcome corresponding to the unit's actual group).

Finally, we'll add a **no anticipation** condition that says that participating in the treatment does not affect outcomes in pre-treatment periods. In math, we can write this as:

For $t < G_i$ (pre-treatment periods for unit i)

$$Y_{it}(G_i) = Y_{it}(0)$$

In the case with two time periods, this imposes the restriction that $Y_{i1}(2) = Y_{i1}(0)$, which says that, for units in group 2, their outcome in the first period is the same as it would have been if they had not participated in the treatment in any time period.

Further, note that ruling out anticipation does not rule out things like treatment effect dynamics (that the effect of treatment can depend on how long a unit has been treated).

Given panel data, the natural analogue of the unconfoundedness assumption is that

$$\mathbf{Y}_i(0) \perp G | \mathbf{W}_i$$

In other words, conditional on having (unobserved) covariates over time $\mathbf{W}_i$, there is nothing special about the distribution of untreated potential outcomes (in either time period) for any group; we'll come back to the issue of including observed covariates $\mathbf{X}_i$ later, but conceptually it would be straightforward to include them here too. Next, probably the most natural way to write the model for untreated potential outcomes is just to put a time subscript on everything. That is,

$$Y_{it}(0) = W_{it}'\beta_t + e_{it}$$

Together with unconfoundedness, it holds that, for all t, $\mathrm{E}[e_t|\mathbf{W}, G] = 0$.

Given this setup, we will run into the same sort of issues as we did before (you can try it!) due to W_{it} not being observed. However, let's suppose that

$$Y_{it}(0) = \theta_t + W_i'\beta + e_{it}$$

As one additional comment, we are slightly abusing notation by separating θ_t out of β_t, but this allows for trends in the untreated potential outcomes over time (for job displacement, this would mean that people's earnings could tend to be increasing over time). θ_t is called a **time fixed-effect**.

We'll talk about how this change is useful momentarily, but first let's talk about whether or not it is reasonable. We have made two imporant changes here:

- Moving from W_{it} to W_i indicates that W does not change over time. This may or may not be reasonable in applications. For example, in the job displacement application discussed earlier, is it reasonable to think that a person's "ability" or "motivation" do not change over time? I am not 100% sure, though perhaps it is reasonable to think that these are close to constant over time, at least over short time horizons.

- Moving from β_t to β indicates that the "effect" of W on untreated potential outcomes is constant over time. Again, this may or may not be reasonable. In our example on job displacement, over longer time horizons, I think that there is evidence that the return to "ability" has increased over time (suggesting that β does in fact vary over time). Perhaps over shorter time horizons, it is not-too-far from being time invariant (it is not totally clear).

In applications where you would hope to "exploit" panel data to estimate causal effect parameters, these are the types of conditions that you had ought to think about. And, an implication will be that, having access to panel data does not automatically guarantee being able to recover the ATT or other causal effect parameters.

At this point, notice that the entire term $W_i'\beta$ does not vary over time. It is common to replace this term generically with $\eta_i := W_i'\beta$. η_i is called a **unit fixed effect** (or sometimes an individual fixed effect). Now, let's explicitly write the model for time periods 2 and 1, and subtract them:

$$\begin{aligned} Y_{i2}(0) &= \theta_2 + \eta_i + e_{i2} \\ Y_{i1}(0) &= \theta_1 + \eta_i + e_{i1} \\ \implies \Delta Y_{i2}(0) &= \Delta\theta_2 + \Delta e_{i2} \end{aligned}$$

and, further,

$$\begin{aligned} \mathrm{E}[\Delta e_2|G] &= \mathrm{E}[e_2|G] - \mathrm{E}[e_1|G] \\ &= \mathrm{E}\big[\underbrace{\mathrm{E}[e_2|\mathbf{W},G]}_{=0}|G\big] - \mathrm{E}\big[\underbrace{\mathrm{E}[e_1|\mathbf{W},G]}_{=0}|G\big] \\ &= 0 \end{aligned}$$

where the second equality uses the law of iterated expectations and holds under the version of unconfoundedness that we have been using. Thus, we have that, for any g,

$$\mathrm{E}[\Delta Y_2(0)|G=g] = \Delta\theta_2 \tag{2}$$

In other words, the average change in untreated potential outcomes over time is the same across all groups (and it is equal to $\Delta\theta_2$). More commonly, this sort of condition can be written as, for any group g,

$$\mathrm{E}[\Delta Y_2(0)|G=g] = \mathrm{E}[\Delta Y_2(0)]$$

This condition is called the **parallel trends assumption**. This sort of condition is potentially useful because we do not observe $\Delta Y_{i2}(0)$ for units in group 1 or group 2, but we do observe it for untreated units (because we observe their untreated potential outcomes in both periods) – we will exploit this below.

Let's define the **group-time average treatment effect**

$$ATT(g,t) = \mathrm{E}[Y_t(g) - Y_t(0)|G=g]$$

This is the average treatment effect in period t of becoming treated in period g relative to not being treated in either time period among those that were in group g. $ATT(g,t)$ provides a natural way to generalize ATT from the case with cross-sectional data to a case with panel data and staggered treatment adoption.

Practice: How do you interpret $ATT(2,1)$? Show that $ATT(2,1) = 0$ under the conditions that we have been considering.

Let's consider trying to recover $ATT(2,2)$ (in other words, the average treatment effect for group

2 in period 2). Notice that

$$\begin{aligned} ATT(2,2) &= \mathrm{E}[Y_2(2) - Y_2(0)|G=2] \\ &= \mathrm{E}[Y_2(2) - Y_1(0)|G=2] - \mathrm{E}[Y_2(0) - Y_1(0)|G=2] \\ &= \mathrm{E}[Y_2(2) - Y_1(0)|G=2] - \mathrm{E}[Y_2(0) - Y_1(0)|U=1] \\ &= \mathrm{E}[\Delta Y_2|G=2] - \mathrm{E}[\Delta Y_2|U=1] \qquad (3) \end{aligned}$$

where the first equality is just the definition of $ATT(2,2)$, the second equality holds by adding and subtracting $\mathrm{E}[Y_1(0)|G=2]$, the third equality uses the parallel trends assumption for the second term, and the last equality holds by writing potential outcomes in terms of their observed counterparts (to be clear, that $Y_1 = Y_1(0)$ for group 2 holds by no anticipation condition; or, more informally, holds because group 2 isn't treated yet in the first time period and therefore we observe their untreated potential outcomes in the first time period).

Here are some additional things to notice:

- The expression for $ATT(g,t)$ on the right side of Equation 3 involves the average difference in outcomes over time among group 2 relative to the average difference in outcomes over time for the never-treated group. This double differencing is what leads to this strategy being called **difference-in-differences**. This is a very common (maybe the most common) identification strategy in economics.

- The same strategy would not work for other groups. For example, suppose that you were interested in $ATT(1,2)$ (the ATT among those that participated in the treatment in both periods). You might consider $\mathrm{E}[\Delta Y_2|G=1] - \mathrm{E}[\Delta Y_2|U=1]$. However, this will not generally be equal to $ATT(1,2)$ – see the practice question below.

- Moreover, group 1 is not used in the expression for $ATT(2,2)$. This suggests group 1 is not useful at all here. For this reason, it is common in DID applications to drop units that are already treated in the first period (i.e., units where $D_{i1} = 1$). If you do this, you can refer to units with $D_{i2} = 1$ as the "treated group" and define $ATT = ATT(2,2)$ which makes for "lighter" notation.

Practice: Show that, under the conditions in this section,

$$\mathrm{E}[\Delta Y_2|G=1] - \mathrm{E}[\Delta Y_2|U=1] = ATT(1,2) - ATT(1,1)$$

How can you interpret $ATT(1,2) - ATT(1,1)$?

Regression Approaches

Most often, DID identification strategies are implemented using what are called two-way fixed effects (TWFE) regressions. That's what we'll start to discuss in this section.

For simplicity, let's suppose that no units are treated in the first time period. If we additionally impose treatment effect heterogeneity: $Y_{i2}(2) - Y_{i2}(0) = \alpha$ which is constant for all units, then the observed outcome in the secon period can be written as

$$\begin{aligned} Y_{i2} &= Y_{i2}(0) + D_{i2}(Y_{i2}(2) - Y_{i2}(0)) \\ &= \theta_2 + \eta_i + e_{i2} + \alpha D_{i2} \end{aligned}$$

where the first equality we have used before, and the second equality holds by the linear model for untreated potential outcomes (and implicitly relies on W_{it} and β_t not varying over time). Furthermore, for the observed outcome in the first time period,

$$Y_{i1} = Y_{i1}(0) = \theta_1 + \eta_i + e_{i1}$$

where the first equality holds by no anticipation. Thus (because $D_{i1} = 0$ for all units), we can generally write

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + e_{it}$$

This is called a two-way fixed effects regression because it includes both a time and unit fixed effect. I have more details about estimating this sort of model below, but you can use R packages like `fixest` or `plm` to estimate this sort of model very easily.

In the case with only two time periods, this regression is exactly equivalent to running the regression

$$\Delta Y_{i2} = \Delta\theta_2 + \alpha D_{i2} + \Delta e_{i2}$$

which amounts to estimating α by running a regression of the change in the outcome on whether or not a unit participated in the treatment.

Interestingly, this particular TWFE regression is robust to treatment effect heterogeneity – see the practice question below.

> **Practice:** Try showing that $\alpha = ATT$. This implies that this regression is robust to treatment effect heterogeneity.
>
> **Hint:** The arguments to show this are very similar to the ones for using a regression in the context of random treatment assignment that we discussed early in the semester.

Practice (*): It's also interesting to think about α when there are units that are already treated in the first period; that is, there are units where $G_i = 1$. Provide an expression for α in this case in terms of underlying group-time average treatment effects.
Hint: Notice that there are only two possible values for ΔD_i; it is equal to 1 for units that become treated (i.e., when $G_i = 2$), and equal to 0 for units that are never-treated ($U_i = 1$) or always-treated ($G_i = 1$).

Multiple Periods

For this part, we'll consider the case where we observe T periods of panel data. In particular, suppose that we observe $\{Y_{i1}, Y_{i2}, \ldots, Y_{it}, D_{i1}, D_{i2}, \ldots, D_{iT}\}_{i=1}^n$ which are iid across units.

We'll continue to make the assumption of **staggered treatment adoption**: for all $t = 2, \ldots, T$, $D_{it-1} = 1 \implies D_{it} = 1$. This means that, once a unit becomes treated, then it remains treated. This is common in applications in economics where, for example, once a location implements a policy, the policy remains in place in subsequent time periods. It also happens when treatments are "scarring"; for example, in job displacement, once a person becomes displaced, it would be typical to think of them as permanently moving into the treated group. Staggered treatment adoption allows for the timing of the treatment to vary across units though. In some sense, this assumption is not necessary, but it will greatly simplify notation below. In particular, it means that we can define a unit's "group", G_i, as the time period when unit i becomes treated. Once we know a unit's group, under staggered treatment adoption, we know it's entire path of participating in the treatment. We'll also set $G_i = T + 1$ for units that do not participate in the treatment in any time period. Let's also use introduce the additional notation, $\mathcal{G}$ to denote the set of all groups (we'll continue to drop group 1 if there is an already treated group), and $\bar{\mathcal{G}}$ to denote the set of all groups excluding the untreated group. Likewise, define $p_g = \mathrm{P}(G = g)$ and $\bar{p}_g = \mathrm{P}(G = g | U = 0)$.

Thus, we can write potential outcomes indexed by group; that is, let $Y_{it}(g)$ denote unit i's outcome in period t if it became treated in period g. Observed outcomes are therefore given by $Y_{it} = Y_{it}(G_i)$. No anticipation implies that $Y_{it}(G_i) = Y_{it}(0)$ for all periods where $t < G_i$ (i.e., periods before the treatment started).

In this case, we'll continue to be interested in $ATT(g, t)$. We'll also make a multi-period version of the parallel trends assumption. In particular, we'll suppose that, for $t = 2, \ldots, T$, and for all groups g

$$\mathrm{E}[\Delta Y_t(0) | G = g] = \mathrm{E}[\Delta Y_t(0)]$$

In other words, the path of untreated potential outcomes is the same for all groups. As in the two-period case, this assumption is very closely related to (i) a version of unconfoundedness conditional on time-invariant unobservables, and (ii) a linear model for untreated potential outcomes such that $Y_{it}(0) = \theta_t + W_i'\beta + e_{it}$ with $\mathrm{E}[e_{it} | W, G] = 0$.

Moreover, note that, for any $t \geq g$ (i.e., post-treatment periods for group g), we have that

$$\begin{aligned}
ATT(g,t) &= \mathrm{E}[Y_t(g) - Y_t(0)|G = g] \\
&= \mathrm{E}[Y_t(g) - Y_{g-1}(0)|G = g] - \mathrm{E}[Y_t(0) - Y_{g-1}(0)|G = g] \\
&= \mathrm{E}[Y_t(g) - Y_{g-1}(0)|G = g] - \mathrm{E}[Y_t(0) - Y_{g-1}(0)|G = 0] \\
&= \mathrm{E}[Y_t - Y_{g-1}|G = g] - \mathrm{E}[Y_t - Y_{g-1}|G = 0]
\end{aligned}$$

where the first equality is just the definition of $ATT(g,t)$, the second equality adds and subtracts $\mathrm{E}[Y_{g-1}(0)|G = g]$ (the average untreated potential outcome in the period before group g becomes treated).

This is very similar to the arguments for the two periods case above except we use period $g - 1$ as the "base period" (the period that we difference with respect to). The reason for this is that it is the most recent period for which we observe untreated potential outcomes for group g.

This expression also suggests estimating $ATT(g,t)$ by computing the average of $Y_{it} - Y_{ig-1}$ for group g relative to the average of $Y_{it} - Y_{ig-1}$ for the untreated group.

This approach should make intuitive sense too. In order to figure out $ATT(g,t)$, we take the path of outcomes that group g actually experienced from its base period to period t and adjust it by the path of outcomes that it would have experienced if it had not participated in the treatment (under parallel trends, this path of outcomes can be recovered from the untreated group).

Aggregating group-time average treatment effects

Group-time average treatment effects are useful causal effect parameters. Once you set them as the target parameter, the DID identification strategies that we have been discussing are relatively straightforward. Group-time average treatment effects can be useful for highlighting treatment effect heterogeneity across groups, time-periods and/or length of exposure to the treatment. However, in many applications, there can be lots of them; perhaps too many to easily report.

In many cases, we would like to recover a more aggregated (i.e., lower-dimensional) parameter. First, let's consider an overall ATT that is a single number summarizing the average effect of participating in the treatment. Towards this end, among units that ever participate in the treatment, define

$$\overline{TE}_i = \frac{1}{T - G_i + 1} \sum_{t=G_i}^{T} (Y_{it} - Y_{it}(0))$$

which is the average treatment effect for unit i across all of its post-treatment time periods; notice that $T - G_i + 1$ is the total number of post-treatment periods for unit i. Also, define

$$ATT^O = \mathrm{E}[\overline{TE}|U = 0]$$

which is the average treatment effect among units that are treated in any time period. ATT^O can be expressed in terms of underlying group-time average treatment effects. In particular, notice that

$$\begin{aligned} ATT^O &= \sum_{g \in \bar{\mathcal{G}}} \mathrm{E}[\overline{TE}|G = g]\bar{p}_g \\ &= \sum_{g \in \bar{\mathcal{G}}} \frac{1}{T - g + 1} \sum_{t=g}^{T} \mathrm{E}[Y_t - Y_t(0)|G = g]\bar{p}_g \\ &= \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} \underbrace{\frac{\bar{p}_g}{T - g + 1}}_{w^O(g,t)} ATT(g, t) \end{aligned}$$

That is, you can see this as a weighted average of $ATT(g, t)$. In other words, ATT^O is a weighted average of underlying group-time average treatment effects where the weights are given by $w^O(g, t)$ (which are weights that you can easily compute). The weights depend on the relative size of the group (through $\bar{p}_g$) and on the number of time periods that a particular group is treated ($ATT(g, t)$'s get less weight for groups that were exposed to the treatment as the weight is split across more time periods).

Another common target parameter in DID applications is the **event study**. The idea is to compute the average treatment effect as a function of the length of exposure to the treatment. For some e (which you can think of as defining the length of exposure to the treatment), among units such that $G_i + e \in [2, T]$, define $TE_i(e) = (Y_{iG_i+e} - Y_{iG_i+e}(0))$ which is the causal effect of the treatment e periods after exposure to the treatment. Then, we can define

$$ATT^{ES}(e) = \mathrm{E}[TE(e)|G + e \in [2, T], U = 0]$$

which is the average effect of having been exposed to the treatment for e periods (conditional on being observed having participated in the treatment for e periods, from the condition that $G + e \in [2, T]$) and participating in the treatment (the condition $U = 0$). This can also be written as an average

$$ATT^{ES}(e) = \sum_{g \in \mathcal{G}} \underbrace{\mathbf{1}\{g + e \in [2, T], U = 0\}\mathrm{P}(G = g|G + e \in [2, T], U = 0)}_{=w^{ES}(g,e)} ATT(g, g + e)$$

which is just the average of $ATT(g, g + e)$ averaged across all ever-treated groups that are ever observed to have participated in the treatment for e periods. (This expression looks more complicated than it is – you can just find all available group-time average treatment effects corresponding to $ATT(g, g + e)$ and average them together weighted by relative group size).

It is also common in applications to report $ATT^{ES}(e)$ for negative values of e. When e is negative, this is an estimate of the average effect of the treatment in periods *before* the treatment takes place. This is useful because, if the parallel trends assumption holds in pre-treatment periods, then it should be the case that $ATT^{ES}(e) = 0$ for $e < 0$. This strategy is called **pre-testing**. To be clear, even if $ATT(e) = 0$ for $e < 0$, it could still be the case that parallel trends is violated in post-treatment periods (which would imply that our estimates of $ATT(g,t)$ would likely be poor). That said, I think it is fair to see this as a validation exercise for the identification strategy. If you see large violations of parallel trends in pre-treatment periods, it should make you feel very worried about your approach.

Regressions under treatment effect homogeneity

Now, let's think about how well running a regression can work in this case. In particular, if we additionally suppose treatment effect homogeneity, then we can get to

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + e_{it}$$

This is exactly the same TWFE regression that we discussed in the case with two time periods. Using arguments quite similar to the ones we have used before, if you estimate this regression, under treatment effect homogeneity you can interpret $\hat{\alpha}$ as an estimate of the causal effect of the treatment. In the next two sections we'll cover (i) the mechanics of how to estimate the sort of model, and (ii) how to interpret α from this model under treatment effect heterogeneity.

Within Estimator

For this section, let's consider how to estimate this sort of TWFE model using a regression. Instead of the specific case of a binary treatment and no other covariates, I'm going to follow the textbook and consider estimating the following model.

$$Y_{it} = \theta_t + \eta_i + X_{it}'\beta + e_{it}$$

For example, X_{it} could include D_{it} or it could include other variables as well.

To start with, define

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t \qquad \bar{e}_i = \frac{1}{T}\sum_{t=1}^{T} e_{it}$$

which are the average outcome (across time periods) for unit i, the average regressors for unit i across time periods, and the the average time fixed effect across time periods.

Throughout this section, we'll maintain the assumption of **Strict Exogeneity** For $t = 1, \ldots, T$, $\mathrm{E}[e_t|\mathbf{X}_i] = 0$ where $\mathbf{X}_i$ is the $T \times k$ matrix that includes X_{it} for all time periods

$$\mathbf{X}_i = \begin{bmatrix} X_{i1}' \\ X_{i2}' \\ \vdots \\ X_{iT}' \end{bmatrix}$$

Strict exogeneity would hold in the model with a binary treatment under the unconfoundedness assumption that we have been considering.

We'll also denote the **within transformation** for a particular random variable by $\dot{C}_{it} = C_{it} - \bar{C}_i$. Moreover, notice that

$$\bar{Y}_i = \bar{\theta} + \bar{X}_i'\beta + \bar{e}_i$$

which implies that

$$(Y_{it} - \bar{Y}_i) = (\theta_t - \bar{\theta}) + (X_{it} - \bar{X}_i)'\beta + (e_{it} - \bar{e}_i)$$

or, equivalently,

$$\dot{Y}_{it} = \dot{X}_{it}'\beta + \dot{e}_{it}$$

where (abusing notation to some extent), I am going to take $\dot{X}_{it}$ to include indicators for a particular time period and β to additionally include corresponding terms that are equal to $\dot{\theta}$.

It's also convenient to write down matrix versions the above expressions. Towards this end, let $\mathbf{1}_i$ denote a $T \times 1$ vector of 1's, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, and recall that $\bar{Y}_i = (\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i'\mathbf{Y}_i$; this holds because $\mathbf{1}_i'\mathbf{1}_i = \sum_{t=1}^T 1 = T$ and $\mathbf{1}_i'\mathbf{Y}_i = \sum_{t=1}^T Y_{it}$, and that

$$\begin{aligned}\dot{\mathbf{Y}}_i &= \mathbf{Y}_i - \mathbf{1}_i\bar{Y}_i \\ &= \mathbf{Y}_i - \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i'\mathbf{Y}_i \\ &= \mathbf{M}_i\mathbf{Y}_i\end{aligned}$$

where $\mathbf{M}_i = \mathbf{I}_T - \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i'$ is a $T \times T$ annihilator matrix. Similarly, it follows that

$$\dot{\mathbf{X}}_i = \mathbf{M}_i\mathbf{X}_i$$

and note that $\dot{\mathbf{Y}}_i$ is a $T \times 1$ vector and $\dot{\mathbf{X}}_i$ is a $T \times k$ matrix.

Then, we can estimate β by the least squares regression of $\dot{Y}_{it}$ on $\dot{X}_{it}$, so that

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^n \sum_{t=1}^T \dot{X}_{it}\dot{X}_{it}'\right)^{-1} \sum_{i=1}^n \sum_{t=1}^T \dot{X}_{it}\dot{Y}_{it} \\ &= \left(\sum_{i=1}^n \dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)^{-1} \sum_{i=1}^n \dot{\mathbf{X}}_i'\dot{\mathbf{Y}}_i \\ &= \left(\sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i\right)^{-1} \sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i\mathbf{Y}_i\end{aligned}$$

Side-Comment: The above expressions are useful for the theory that we derive below, but for estimation it is useful to have "data matrix" versions of all of the terms here. Towards that end, let's define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \qquad \mathbf{D} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix}$$

which are an $nT \times 1$ vector, an $nT \times k$ matrix, and an $nT \times n$ matrix, respectively. Similarly define $\mathbf{M}_D = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ which is an $nT \times nT$ matrix. Then, you can show that

$$\mathbf{M}_D\mathbf{Y} = \dot{\mathbf{Y}} = \begin{pmatrix} \dot{\mathbf{Y}}_1 \\ \dot{\mathbf{Y}}_2 \\ \vdots \\ \dot{\mathbf{Y}}_n \end{pmatrix} \qquad \text{and} \qquad \mathbf{M}_D\mathbf{X} = \dot{\mathbf{X}} = \begin{pmatrix} \dot{\mathbf{X}}_1 \\ \dot{\mathbf{X}}_2 \\ \vdots \\ \dot{\mathbf{X}}_n \end{pmatrix}$$

so that we can write

$$\hat{\beta} = (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_D\mathbf{Y}$$

This expression is convenient to use in computations.

Next, notice that

$$\hat{\beta} - \beta = \left(\sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i \right)^{-1} \sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i\mathbf{e}_i$$

Under strict exogeneity, it immediately follows that

$$\mathrm{E}[\hat{\beta} - \beta|\mathbf{X}] = \left(\sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i \right)^{-1} \sum_{i=1}^n \mathbf{X}_i'\mathbf{M}_i \underbrace{\mathrm{E}[\mathbf{e}_i|\mathbf{X}]}_{=0}$$

which implies that $\hat{\beta}$ is unbiased for β.

Asymptotic Distribution

H: 17.20

Under standard assumptions (see Assumption 17.2 in the textbook) that include (i) iid sample (across units), (ii) a positive definite condition, (iii) existence of moments, and (iv) strict exogeneity, notice that we can write

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{M}_i\mathbf{e}_i$$

From the weak law of large numbers, we have that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i \xrightarrow{p} \mathrm{E}[\mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i]$$

which is a $k \times k$ matrix and from the central limit theorem, we have that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{M}_i\mathbf{e}_i \xrightarrow{d} N(0, \mathbf{\Omega})$$

where

$$\mathbf{\Omega} = \mathrm{E}\left[\mathbf{X}_i'\mathbf{M}_i\mathbf{e}_i\mathbf{e}_i'\mathbf{M}_i'\mathbf{X}_i\right]$$

which is a $k \times k$ matrix, and the continuous mapping theorem implies that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{V})$$

where

$$\mathbf{V} = \mathrm{E}[\mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i]^{-1}\mathbf{\Omega}\mathrm{E}[\mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i]^{-1}$$

and this can be estimated in the usual way (i.e., replace population moments with sample averages and replace $\mathbf{e}_i$ with $\hat{\mathbf{e}}_i$ (the vector of residuals for unit i)).

Regressions under treatment effect heterogeneity

Next, let's consider whether or not this kind of TWFE regression is robust to treatment effect heterogeneity (or what exactly you are getting when there is treatment effect heterogeneity).

This is an interesting thing to consider because this kind of regression has been the dominant way that empirical researchers have implemented DID identification strategies over the past 30 years or so. Besides that, this kind of regression was robust to treatment effect heterogeneity; therefore, it seems reasonable to think (or at least hope) that this would continue to be the case in the present case where there are more periods. In order to make progress along these lines, let's define

$$\begin{aligned}\ddot{Y}_{it} &= Y_{it} - \bar{Y}_i - \mathrm{E}[Y_t] + \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_t] \\ \ddot{D}_{it} &= D_{it} - \bar{D}_i - \mathrm{E}[D_t] + \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_t]\end{aligned}$$

which are population versions of what's called **double de-meaning** Y_{it} and D_{it}. This sort of transformation removes the unit- and time-fixed effects from Y_{it} and D_{it}. In particular, notice that

$$\begin{aligned}Y_{it} &= \theta_t + \eta_i + \alpha D_{it} + e_{it} \\ \bar{Y}_i &= \bar{\theta} + \eta_i + \alpha \bar{D}_i + \bar{e}_i \\ \mathrm{E}[Y_t] &= \theta_t + \mathrm{E}[\eta] + \alpha \mathrm{E}[D_t] + \mathrm{E}[e_t] \\ \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_t] &= \bar{\theta} + \mathrm{E}[\eta] + \alpha\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[D_t] + \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[e_t]\end{aligned}$$

Combining these expressions implies that

$$\ddot{Y}_{it} = \alpha \ddot{D}_{it} + \ddot{e}_{it}$$

which has removed the unit- and time- fixed effects. This expression is easier to deal with, and we know that,

$$\alpha = \frac{\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_{it}\ddot{D}_{it}]}{\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[\ddot{D}_{it}^2]}$$

There are some useful properties of double de-meaned variables. First, it is straightforward to show that

$$\mathrm{E}[\ddot{D}_{it}] = 0 \qquad \text{and} \qquad \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[\ddot{D}_{it}C_i] = 0$$

where C_i is some generic random variable that is constant across time. I'm not going to provide the proof of these, but you can show them just by brute force (i.e., plugging in for $\ddot{D}_{it}$) and algebra.

In order to related α to underlying $ATT(g,t)$'s, it is helpful to notice that $\ddot{D}_{it}$ is fully determined by a unit's group. In particular,

$$\begin{aligned} D_{it} &= \mathbf{1}\{t \geq G_i\} \\ \bar{D}_i &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{t \geq G_i\} = \frac{T - G_i + 1}{T} \end{aligned}$$

and $\mathrm{E}[D_t]$ and $\frac{1}{T}\sum_{t=1}^{T} \mathrm{E}[D_t]$ are just numbers. Thus, we can write $\ddot{D}_{it} = h(G_i, t)$. Two more things that we use below are that

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{g \in \mathcal{G}} h(g,t)\mathrm{E}[Y_{it}|U=1]p_g = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_{it}|U=1] \underbrace{\sum_{g \in \mathcal{G}} h(g,t)p_g}_{=\mathrm{E}[h(G_i,t)]=\mathrm{E}[\ddot{D}_{it}]=0} \tag{4}$$

and that

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{g \in \mathcal{G}} h(g,t)\mathrm{E}[Y_{i1}|U=1]p_g = \mathrm{E}[Y_{i1}|U=1]\frac{1}{T}\sum_{t=1}^{T} \underbrace{\sum_{g \in \mathcal{G}} h(g,t)p_g}_{=\mathrm{E}[h(G_i,t)]=\mathrm{E}[\ddot{D}_{it}]=0} \tag{5}$$

Now, let's consider the numerator for α. It is given by

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_{it}\ddot{D}_{it}] &= \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_{it}\ddot{D}_{it}] - \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[Y_{i1}\ddot{D}_{it}] \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{g \in \mathcal{G}}\mathrm{E}[h(g,t)(Y_{it} - Y_{i1})|G=g]p_g \\ &= \frac{1}{T}\sum_{t=1}^{T}\sum_{g \in \mathcal{G}}\mathrm{E}[h(g,t)(Y_{it} - Y_{i1})|G=g]p_g - \frac{1}{T}\sum_{t=1}^{T}\sum_{g \in \mathcal{G}}\mathrm{E}[h(g,t)(Y_{it} - Y_{i1})|U=1]p_g \\ &= \frac{1}{T}\sum_{t=2}^{T}\sum_{g \in \bar{\mathcal{G}}} h(g,t)\Big(\mathrm{E}[Y_{it} - Y_{i1})|G=g] - \mathrm{E}[Y_{it} - Y_{i1})|U=0]\Big)p_g \\ &= \sum_{g \in \bar{\mathcal{G}}}\sum_{t=g}^{T}\frac{h(g,t)p_g}{T}ATT(g,t) \end{aligned}$$

where the first equality uses the properties of $\ddot{D}_{it}$ and that Y_{i1} doesn't vary over time, the second equality replaces $\ddot{D}_{it}$ with $h(G_i, t)$ and from the law of iterated expectations, the third equality uses Equations 4 and 5, the fourth equality rearranges terms, the fifth equality uses the parallel trends assumption (and that $ATT(g,t) = 0$ for $t < g$).

This implies that

$$\alpha = \sum_{g \in \bar{\mathcal{G}}}\sum_{t=g}^{T} w^{TWFE}(g,t)ATT(g,t)$$

where

$$w^{TWFE}(g,t) = \frac{\frac{h(g,t)p_g}{T}}{\frac{1}{T}\displaystyle\sum_{t=1}^{T} \mathrm{E}[\ddot{D}_{it}^2]} = \frac{h(g,t)p_g}{\sum_{t=1}^{T} \mathrm{E}[\ddot{D}_{it}^2]}$$

You can show that $\sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} w^{TWFE}(g,t) = 1$ (which is good; to see this you can notice that $\mathrm{E}[\ddot{D}_{it}^2] = \mathrm{E}[D_{it}\ddot{D}_{it}]$ and then follow the same arguments as above but with D_{it} replacing Y_{it}. When you do this, the term $ATT(g,t)$ will be replaced by 1.), but notice that in general $w^{TWFE}(g,t) \neq w^{O}(g,t)$ defined above. This means that, in the presence of treatment effect heterogeneity (so that $ATT(g,t)$ is not constant across g and t), it would not generally be the case that $\alpha = ATT^{O}$. This suggests a lack of robustness to treatment effect heterogeneity.

Besides that, $w^{TWFE}(g,t)$ can be negative for some values of g and t. Negative weights are particularly troubling as they open up the possibility that the effect of the treatment could be, say, positive for all units but you could get a negative estimate due to the estimation strategy.

Example:

I think it is useful to try out an example. One example that delivers negative weights is when $T = 3$, $p_2 = 0.4$, $p_3 = 0.4$ and the probability of being in the untreated group is 0.2. In this case, I get that

$$
\begin{aligned}
w^{TWFE}(2,2) &= \frac{5}{8} \\
w^{TWFE}(2,3) &= -\frac{1}{8} \\
w^{TWFE}(3,3) &= \frac{4}{8}
\end{aligned}
$$

It is interesting to contrast this with the weights on $ATT(g,t)$'s from ATT^O. These are given by

$$
\begin{aligned}
w^{O}(2,2) &= \frac{1}{4} \\
w^{O}(2,3) &= \frac{1}{4} \\
w^{O}(3,3) &= \frac{1}{2}
\end{aligned}
$$

You can see how this might go poorly. For example, suppose that the treatment takes one period to have any effect so that $ATT(g,g) = 0$ and $ATT(g,g+1) = 1$ for all groups. In this case $ATT^O = \frac{1}{4}$ (which is the average treatment effect across all post-treatment periods across all groups). On the other hand, $\alpha = -\frac{1}{8}$. Thus, even if the DID identification strategy were working exactly right, you would estimate the wrong sign of the effect of the treatment due to using a TWFE regression rather than the alternative approaches that we have been talking about.

A natural question to ask at this point is: what is going wrong here? It turns out that you can alternatively decompose α from the TWFE regression into a weighted average (with all positive weights) of comparisons between paths of outcomes of units that become treated relative to units whose treatment status does not change. The problem with this is that units whose treatment status does not change include (i) not-yet-treated units (these are "good" comparisons that are in the spirit of DID), and (ii) already-treated units (these are "bad" comparisons in that their paths of outcomes could be affected by the treatment).

If there is treatment effect homogeneity, then it does make sense to use already-treated units as part of the comparison group (in this case, treatment happens and increases the level of outcomes but units keep the same trend they would have had absent the treatment, so it is ok to use their post-treatment paths of outcomes in this case). However, if there is treatment effect heterogeneity (particularly dynamics), you start to get weighted averages of terms similar to the ones in Practice (*) above.

Covariates

Recall that our original unconfoundedness assumption also included observed covariates X. To conclude this part of the course, let's revisit that case. I'll consider the case where the observed covariates are time-invariant as I think this is a leading case; in the example on job displacement, the most important covariates that are likely to be observed are a person's demographic characteristics (typically time invariant) and a person's years of education (typically close to time invariant for at ages where people are at risk of being displaced). It is natural to consider a version of the parallel trends assumption that includes covariates

$$\mathrm{E}[\Delta Y_t(0)|X, G = g] = \mathrm{E}[\Delta Y_t(0)|X]$$

In other words, conditional on having covariates X, then paths of untreated potential outcomes are the same across all groups and time periods.

Moreover, in this case, you can show that

$$\begin{aligned} ATT(g,t) &= \mathrm{E}[Y_t - Y_{g-1}|G = g] - \mathrm{E}[Y_t(0) - Y_{g-1}(0)|G = g] \\ &= \mathrm{E}[Y_t - Y_{g-1}|G = g] - \mathrm{E}\Big[\mathrm{E}[Y_t(0) - Y_{g-1}(0)|X, G = g\big|G = g\Big] \\ &= \mathrm{E}[Y_t - Y_{g-1}|G = g] - \mathrm{E}\Big[\mathrm{E}[Y_t(0) - Y_{g-1}(0)|X, U = 1\big|G = g\Big] \\ &= \mathrm{E}[Y_t - Y_{g-1}|G = g] - \mathrm{E}\Big[\mathrm{E}[Y_t - Y_{g-1}|X, U = 1\big|G = g\Big] \end{aligned}$$

which is identified and is similar to expressions that we have seen before.

The above expression is very similar to the outcome regression approach that we discussed earlier this semester; you could imagine estimating $ATT(g,t)$ by imposing that $\mathrm{E}[Y_t - Y_{g-1}|X, U = 1] = X'\beta_t$ and estimating β_t from running a regression of $(Y_t - Y_{g-1})$ on X using observations from the untreated group.

You can also develop propensity score weighting and doubly robust expressions for $ATT(g,t)$, similar to what we've done before, in this case as well:

$$\begin{aligned} ATT(g,t) &= \mathrm{E}\left[\left(\frac{\mathbf{1}\{G = g\}}{\bar{p}_g} - \frac{U p_g(X)}{\bar{p}_g(1 - p_g(X))}\right)(Y_t - Y_{g-1})\right] \\ ATT(g,t) &= \mathrm{E}\left[\left(\frac{\mathbf{1}\{G = g\}}{\bar{p}_g} - \frac{U p_g(X)}{\bar{p}_g(1 - p_g(X))}\right)(Y_t - Y_{g-1} - \mathrm{E}[Y_t - Y_{g-1}|X, U = 1])\right] \end{aligned}$$

where we define $p_g(X) = \mathrm{P}(G = g|X, \mathbf{1}\{G = g\} + U = 1)$; which is a version of the propensity score – it is the probability of being in group g conditional on covariates and on being either in group g or the never-treated group.

These are very similar to what we talked about in the previous set of notes. Moreover, the aggregations that we talked about previously can continue to apply. Finally, you could consider using the following sort of TWFE regression here

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + X_i'\beta_t + e_{it}$$

This strategy would work fine under treatment effect homogeneity, but, in the presence of treatment effect heterogeneity would suffer from the problems that we talked about in the last set of notes (such as weight reversal) as well as the problems due to multiple periods that we have talked about above. To me, this strategy seems notably less attractive than the alternative approaches discussed above.