Online appendix for "An Introductory Guide to Event Study Models"

The supplemental materials for the paper contain Stata code that produces the Figures in this appendix.

A Data structures, and related designs

A.1 Connections to Difference-in-Difference models

Event study models fit within a family of related models that rely on a parallel trends assumption for identification of causal effects. All of these employ panel fixed effects (or a simplified version, such as dummies for "post" and "treated unit") as key control variables. In Table A.1, I summarize some related approaches within this family. The first column labels the approach; the second column indicates the relevant estimating equation, the third and fourth columns identify the relevant data structure.

Table A.1: **Collection of ES and related models**

	Model Name	*Estimation Equation*	*Event Date Variation*	*Never-treated group(s)*
1.	$2x2$ Difference-in-Difference	DiD	N/A	Yes
2.	$2xT$ Difference-in-Difference	ES, DiD	N/A	Yes
3.	NxT Difference-in-Difference	ES, DiD	Common	Yes
4.	NxT Generalized DiD	DiD	Varying	Optional
5.	Event Study, Timing based	ES	Varying	No
6.	Event Study, DiD style	ES	Common	Yes
7.	Event Study, Hybrid	ES	Varying	Yes

The first row is the basic 2×2 difference in difference model. Here we have two units, one treated and one control. And we have two time periods: one before treatment and one after. Row 2 is the generalization of this where we have multiple time periods for each unit. In this case, there is the possibility of creating an event-study type graph. The next extension is to have many (N) units, some treated and some control; and for the treated units to have a common event time. This is the $N \times T$ difference-in-difference setting. The essence of

the identification is the same as the $2 \times T$ DiD model; but the many units can allow for difference in calculating standard errors (we can now estimate standard errors by clustering on each unit).

The last four rows of the table are all characterized by settings where the event time varies across units. The Generalized Difference-in-Difference estimates a single "treatment effect" from this. This is the first model where it's possible to have only "ever treated" units, and to identify treatment effects based solely on the timing of the treatment. The three event study setups build from earlier data structures, and produce our typical ES graphs.

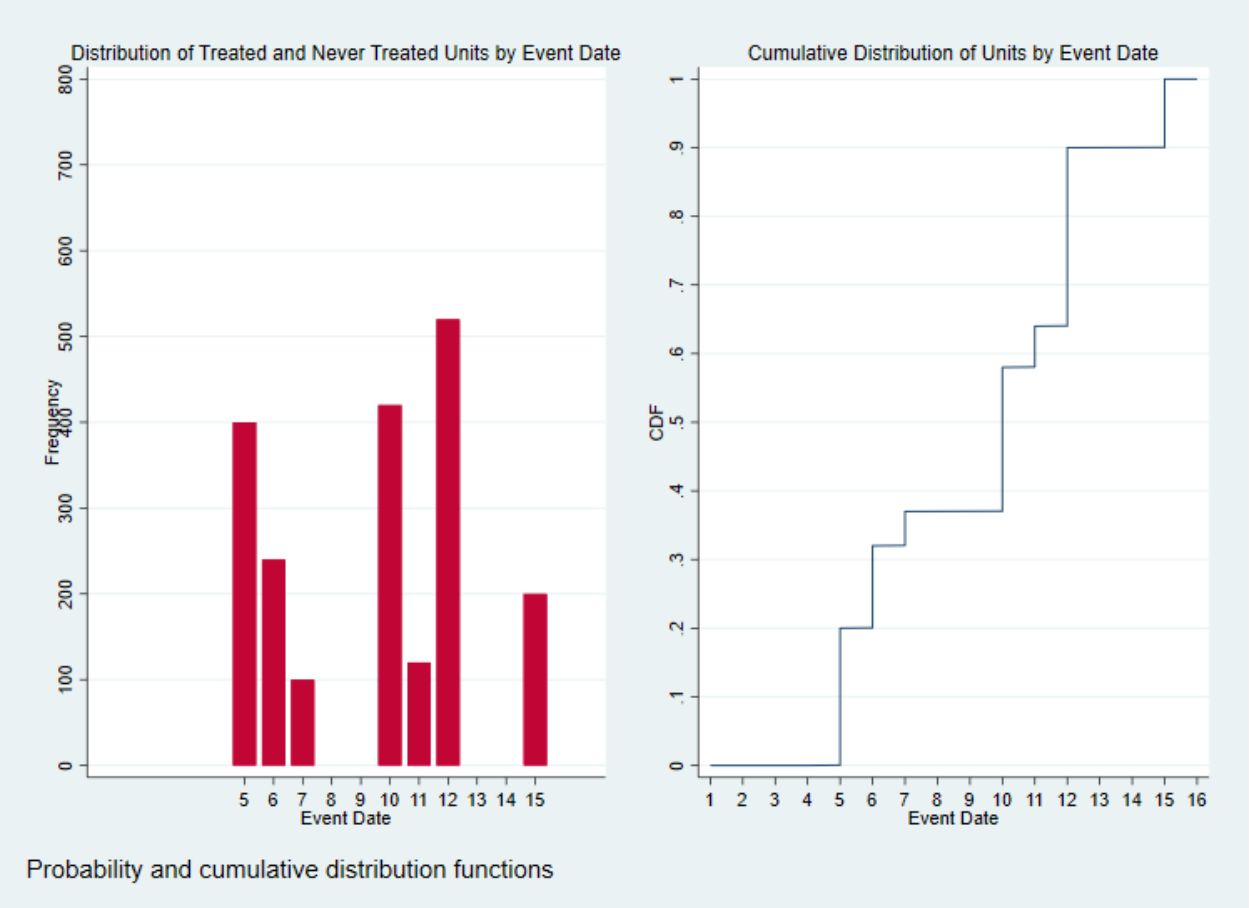
The list above is incomplete, and there are many variations. One common situation is when the variation in event dates E_i is neither cleanly "all at once" ($E_i = E, \forall i$), but there are important groupings of E_i across units. For example, a policy might be adopted by a handful of states at different times; and then a federal policy might bring along all of the remaining states all at once.

A.2 Showing the variation in your event dates

Because the data structure you are working with impacts specification choices, you should clearly let your reader know which structure you have. Also, if you are working with a timing-based or hybrid data structure, you should let your reader know the variation in the event dates in your sample. This can be done with a tabulation of event dates, or graphically as in the figures below. The figures represent a couple of different hypothetical data sets, and show two ways of illustrating the data structure and variation in event date. Each pair of graphs shows the same information in two different ways. For your paper, you can choose whichever format you think is most clear for your readers.

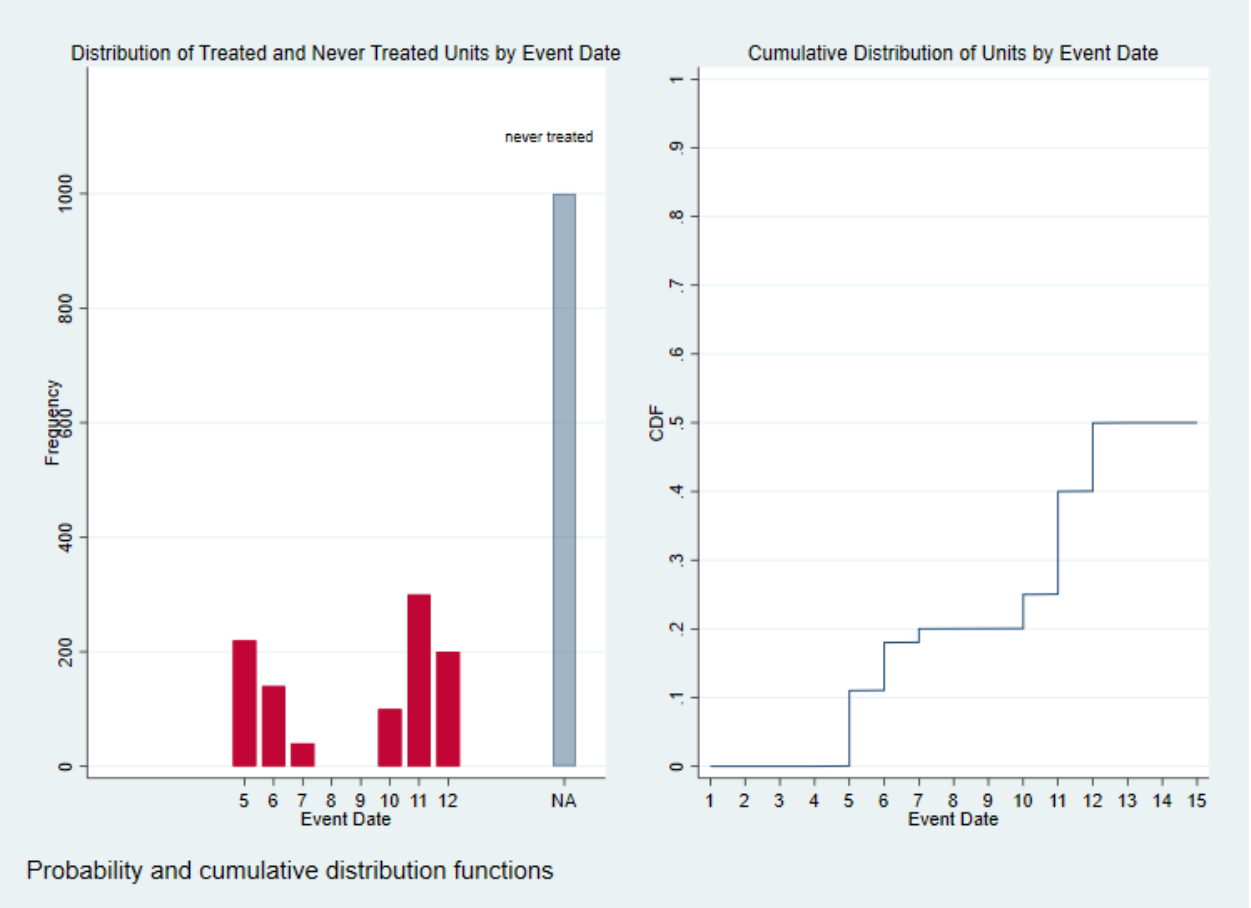
Figure A.1 illustrates this for a timing-based data structure. The earliest treated units have their event date in period 5, and the latest event date is period 15, by which point all units have been treated. The graph on the right shows the same information, in the form of a CDF across units of event dates.

Figure A.1: Two ways of showing the variation in event dates: Timing-based data structure



Note: The left panel shows a histogram of event dates, with one observation per unit. The right panel shows the same information as Cumulative Distribution Function. This data set has a timing-based data structure, with no "never treated" units.

Figure A.2: Two ways of showing the variation in event dates: Hybrid data structure



Note: The left panel shows a histogram of event dates, with one observation per unit. The right panel shows the same information as Cumulative Distribution Function. This data set has a hybrid data structure, with variation in event date among treated units, and many "never treated" units.

Figure A.2 shows a hybrid data structure. Here, half of the units are never-treated. Of those that are treated, there is an early-block, with event dates 5-7, and a later block, with event dates 10-12.

For each of the figures above, the two graphs on the left and right convey the same information about the data structure. I recommend presenting one of these, choosing the style that you think will be most informative to your readers.

B Parameter restrictions

B.1 Timing-based Data Structures and parameter restrictions required

In DiD based data structures, in models with no trend controls, three restrictions on the parameters are required. The regular panel fixed effects restrictions are typically (1) drop the intercept, and (2) drop a unit fixed effect. These "make sense" and are unobjectionable. The third restriction is (3a) the typical restriction to normalize an event time coefficient to zero (e.g. set $\gamma_{-1} = 0$, or (3b) normalize an average of the "reference period" coefficients to zero.

In timing-based data structures, things get more complicated. With two event dates, there are the same number of "effective limiting observations", but now one or more extra parameters (based on $E_{max} - E_{min}$) to be estimated (because we have more event-time parameters). So one or more extra restrictions are needed. In one sense, this seems worse. On the other hand, we can still identify the same number of parameters that we could have with the DiD structure. (What did the DiD structure have to say about the novel parameter? Nothing.) However, the restrictions we impose on the model will impact all of the estimated parameters. It's not like we can say "we ignore the extra parameter" like we do in the DiD structure; instead we have to say something like "we think its value is the same as its neighbor", and that assumption has implications for all of our estimated parameters.

When we add extra unit types with extra event dates (E_i), each one apparently brings with its T new limiting observations. However, there are lots of multicolliniearities; and so the extra information (as measured by the rank of the X matrix) typically grows by only 2 degrees of freedom. One of these is used to identify the level shift α_i for that unit type. And if our new unit type expands the event time parameter space (e.g. by increasing $E_{max} - E_{min}$) then we are left with the same number of total extra restrictions needed. This is still a situation of "good news"; for the same number of needed restrictions we can identify

more and more γ_j.

When there is a gap between E_{min} and E_{max}, the location(s) of other event dates within this gap are important for the amount of identifying information (as measured by the rank of the regressors X, including all the dummy variables and event dummies). The patterns here are complex; and while I would guess that there is a closed-form solution, I am not sure what it is. The themes appear to be: (1) information decreases (more parameter restrictions are required) as the minimum gap $min_{i,j}(E_i, E_j)$ grows; (2) In a data structure with three events, information jumps to "max" when the interior event time is just-barely-offset (by 1 time period) from the mid-point of the range; (3) more unit-types typically helps, as they add new event dates E_i to anchor event time around, and typically narrow gaps between event dates.

This is illustrated in Figure A.3 below. The setting here is based on a timing-based data structure, with no "untreated units", and a panel length of $T = 20$. One of the unit types is treated at $T = 4$, and a second unit type is treated at $T = 16$. If these were the only unit types, the model would need $E_{max} - E_{min} = 12$ extra parameter restrictions to be identified. Next, we consider having a third unit type, with treatment date somewhere in between 4 and 10. This doesn't change the number of parameters to identify; but it can add additional non-collinear observations. In doing so it can reduce the number of needed parameter restrictions. Depending on when the third unit's event date is, we can calculate the rank of the X matrix, and compare this rank to the number of parameters in the model. The gap between these two gives the number of additional needed parameter restrictions to identify the model.

The blue circles in the graph show how the number of needed restrictions changes when we add a third unit type, as a function of the timing of the event for that unit type E_3. When its event date is 4 (the same date as our first unit type), we are still in the case of really having only two unit types, and we need the full 12 parameter restrictions. With an event date of 5, we now need only 1 parameter restriction. The patterns of the blue circles

Figure A.3: Strange patterns in the number of needed parameter restrictions in timing-based data structures

Note: The y-axis show the number of additional parameter restrictions (beyond those that would be requred for a difference in difference data structure) that are required to identify the parameters of the model. For the blue circles ("units3") there are three unit types. One has an event date at $t = 4$, and the other at $t = 16$. The x-axis represents the event date of the third unit type. For the red triangles ("units4") there are four unit types, three of whom have event dates at $\{4, 10, 16\}$. The x-axis represents the event date of the fourth unit type.

are strange and non-monotonic. I think that explaining these is a puzzle for future research.

The red triangles expand the thought experiment to conisder four unit types. In this scenario, the fourth unit type receives treatment at the midpoint, $E_4 = 10$. The x-axis is based on the location of the third unit type, and the y-axis shows the number of additional parameter restrictions needed to identify the model. As before, the patterns are strange and intriguing.

B.2 Implementing parameter restrictions in Stata with cnsreg

One way to implement parameter restrictions $\gamma_j = 0$ is to drop the associated variable. The most common restriction used in event study models is $\gamma_{-1} = 0$, and this is implemented by excluding the -1 event time dummy variable. To implement equality of coefficients across event times, an easy way implement this is to create a pooled dummy variable. For example to impose $\gamma_0 = \gamma_1$, we can include a dummy variable for "event time is zero or one". This idea extends to the "end cap" variables that are often used.

In this subsection I discuss an alternative approach: the use of direct parameter restrictions in estimation. In Stata, this is implemented with the command cnsreg ("constrained regression"). This is the command I use to create the figures in the Online Appendix, and the supplementary materials for the paper include code which illustrates its use.

To use cnsreg, first you define the parameter restrictions in the form of linear constraints, and then reference the constraints when calling the command. For example to implement "set the reference period to be event times -1 and -2", we want to constrain $\gamma_{-1} + \gamma_{-2} = 0$. To implement this in Stata we do this as follows:

```
constraint define 1 Dm1 + Dm2 = 0
cnsreg y Dm3 Dm2 Dm1 Dp0 Dp1 Dp2 ibn.time i.id , constraints(1) collinear
```

One advantage of using cnsreg is that you can make sure that Stata is not dropping unexpected collinear terms. In order to do this, you need to use the "collinear" option. And if you are using Stata's factor notation for your time or unit-dummies, you need to use the no-base option: "ibn.time".

Another use of cnsreg is to implement the proposed trend normalization in section 4.4. of the paper. A third use can be used to implement a spline in the event time coefficients, by imposing a "no concavity" constraint, so that the slope is equal across two segments of the spline. For example: $\gamma_1 - \gamma_0 = \gamma_2 - \gamma_1$.

For an alternative approach in Stata to estimating event study modesl, see Clarke and Schythe (2020) who present a Stata add-on command.

C Illustration of alternative normalizations of the reference period

C.1 DiD Data Structure, alternative normalizations, and visual pre-trends

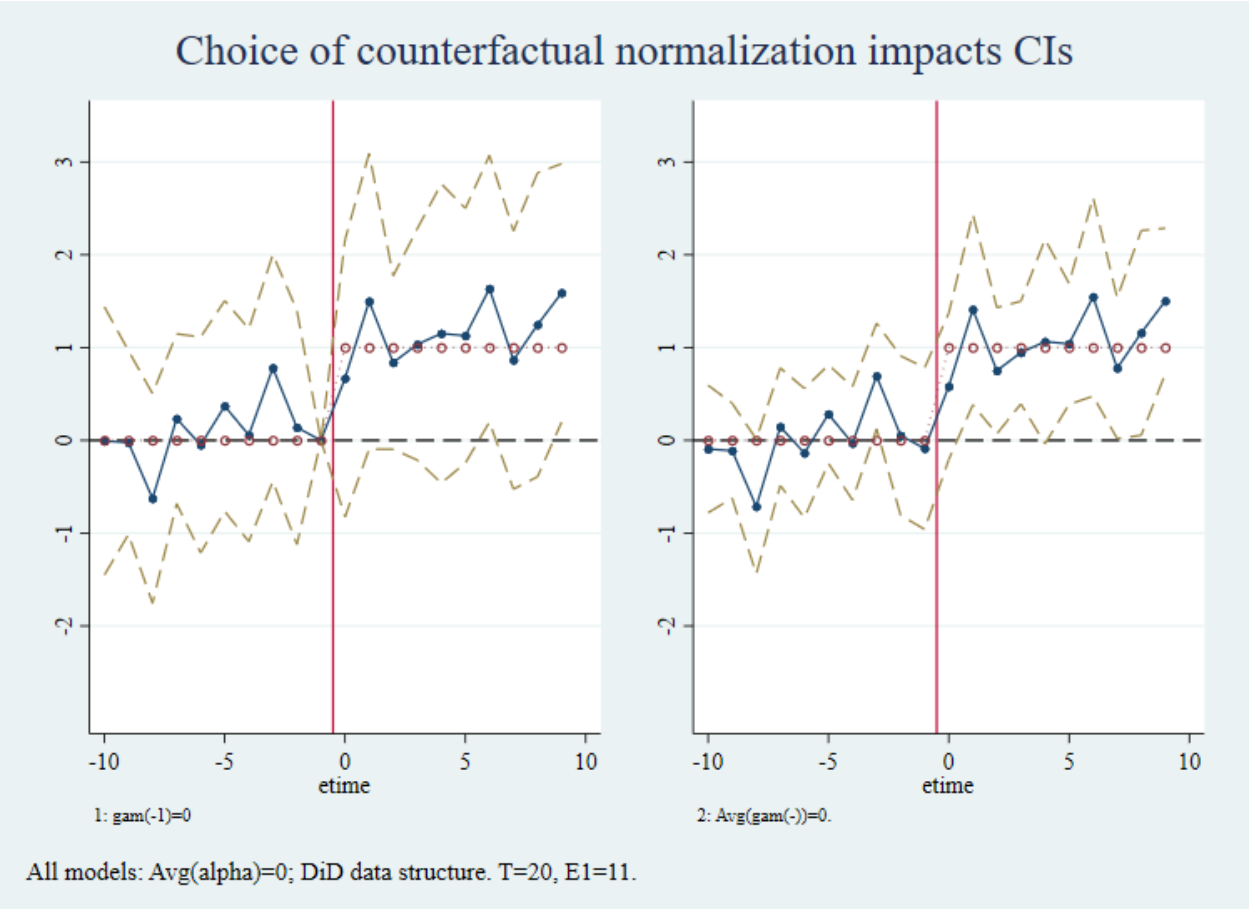
This section illustrates some issues from Section 3.1.

In figure A.4, both graphs are estimated on the same data. They both employ one panel-fixed-effects restriction in common: that the average unit-type coefficients are zero. The figure on the left uses the more common event study normalization, that the coefficient on the -1 term equals zero. The figure on the right uses the recommended event study restriction, that the average coefficient in the reference period is zero. Here I use event times -1 through -10 as the reference period. The difference in restrictions has the effect of shifting up or down the whole pattern of coefficients. In this example, the shift is very small, because the -1 coefficient is very close to the overall average for the pre-period. The other effect is on the estimated standard errors. They are larger when using the -1 restriction, reflecting the additional uncertainty driven by the noise in this term on its own. When the full reference period is used, the standard errors are noticeably smaller.[19]

If we normalize to a broader reference period, we can still examine the pre-event coefficients for a sign of a pre-trend. However, because we are normalizing these coefficients to average to zero, the pre-trend will manifest differently than if we had normalized the -1 coefficient to zero. We need to assess the overall trend in coefficients, rather than examine

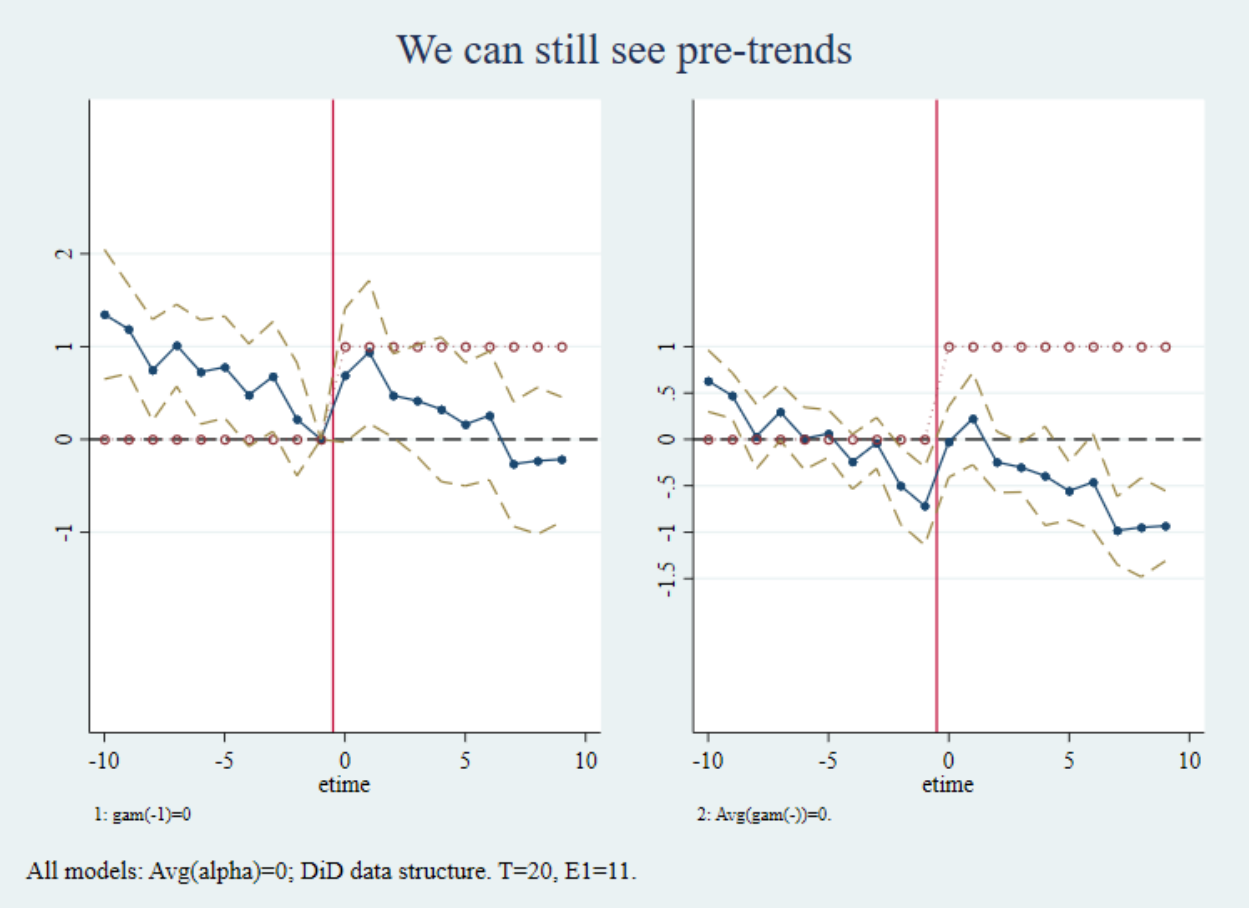
[19]The data in this example were selected so as to have results of statistical significance differ across the two graphs, as a rhetorical trick to emphasize the main point. The general lesson is that using the full reference period will (1) show increased precision, and (2) corresponds to our intuitive counterfactual, informed by difference in difference models.

Figure A.4: Different counterfactual normalizations



Note: The y-axis show the estimated treatment effects and 95% confidence intervals. The x-axis shows event time. The left panel normalizes event time -1 to zero; while the right panel normalizes the average of -10 to -1 to be zero.

Figure A.5: Different counterfactual normalizations and pre-trends

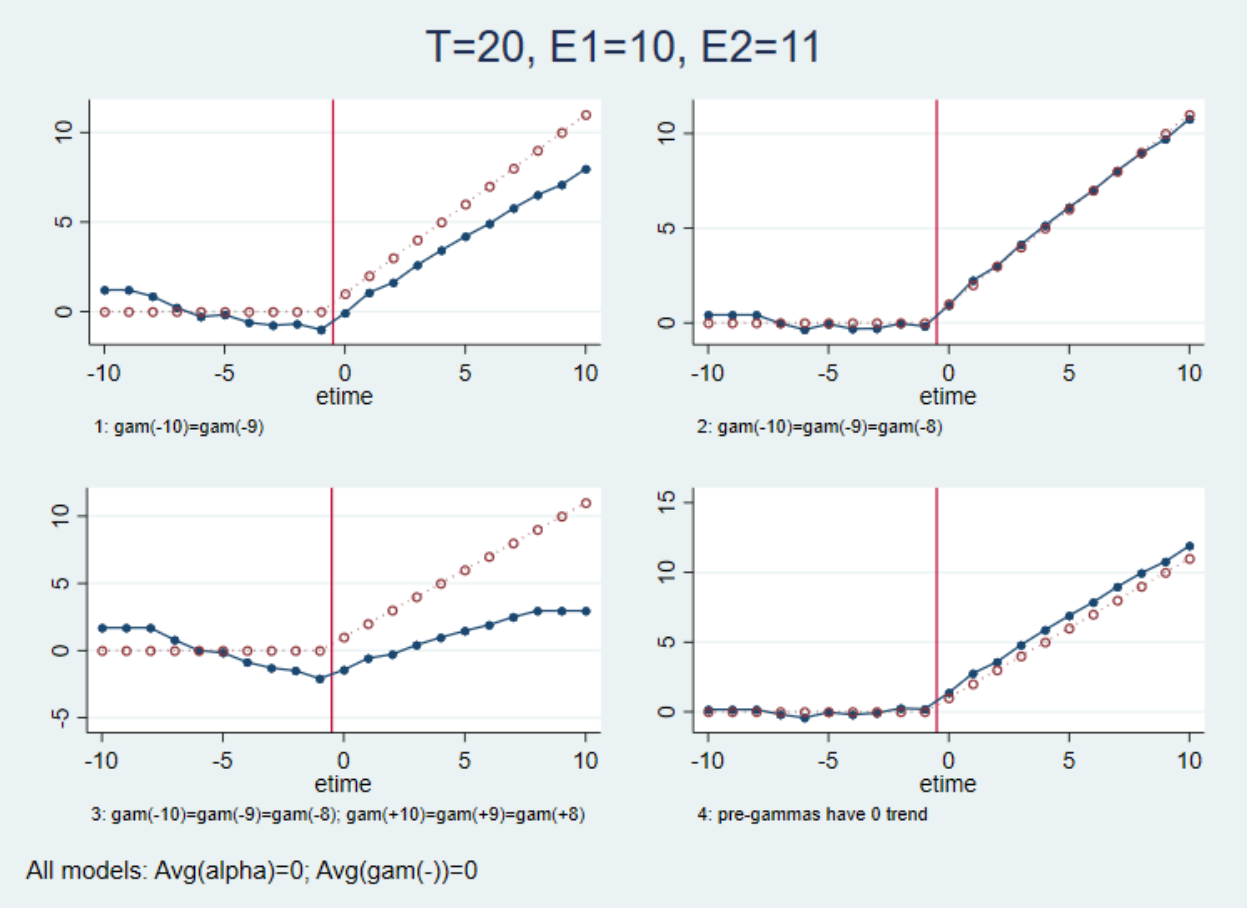


Note: The y-axis show the estimated treatment effects and 95% confidence intervals. The x-axis shows event time. The left panel normalizes event time -1 to zero; while the right panel normalizes the average of -10 to -1 to be zero.

point-wise coefficients and their difference from zero. This is illustrated in figure A.5. In this data generating process, I have added in a systematic time trend for the treated units.

The graph on the left of figure A.5 shows the expected visual evidence of this pre-trend. The graph on the right is shifted down (because it constrains the average pre-period coefficient to be zero). The trend is just as apparent if we examine the overall pattern of the pre-event coefficients. If we used tests of "are these coefficients different from zero", the graph on the right would reject less often. But this would be the wrong criterion to use for pre-tests. Instead we need to examine the overall pattern of the pre-event coefficients. There is a clear steady downward trend in these coefficients. Using this criterion, there is no loss in moving to the broader reference period normalization.

Figure A.6: Timing-based data structure and different restrictions



Note: The y-axis show the estimated treatment effects and 95% confidence intervals. The x-axis shows event time. The four panels are based on different parameter restrictions.

C.2 Timing-based Data Structure: $E_2 = E_1 + 1$

In this section, we consider a timing-based data structure with two unit types. The event dates for the two units are off-set by 1. Because it is a timing-based data structure with no control group, in addition to the basic constraints, we need at least one more. In figure A.6 I illustrate consequences for four different possibilities for the additional constraint(s). The first and last graphs are "just identified"; graphs 2 and 3 have extra constraints.

Model 1 uses a minimal "end-cap" constraint, on the pre-period end-cap only. It looks okay; but shows a lot of noise, which twists the estimates about the fulcrum of the two points in the end-cap. It might be made worse because γ_{-10} only comes from one unit-type. Model 2 extends the end cap to cover 3 periods. It looks much better, as it is much flatter.

Model 4 implements my recommended constraint that the pre-event terms have zero trend. It also looks good, and (like model 1) is "just identified". Model 3 looks awful; this would be a commonly estimated model using end-caps on both ends. This example is a cautionary tale for standard practice.

D Getting closer to raw data

This appendix illustrates how we can show both our event study estimates, and also provide additional context by showing results that are closer to the raw data. It illustrates some of the suggestions in section 3.2

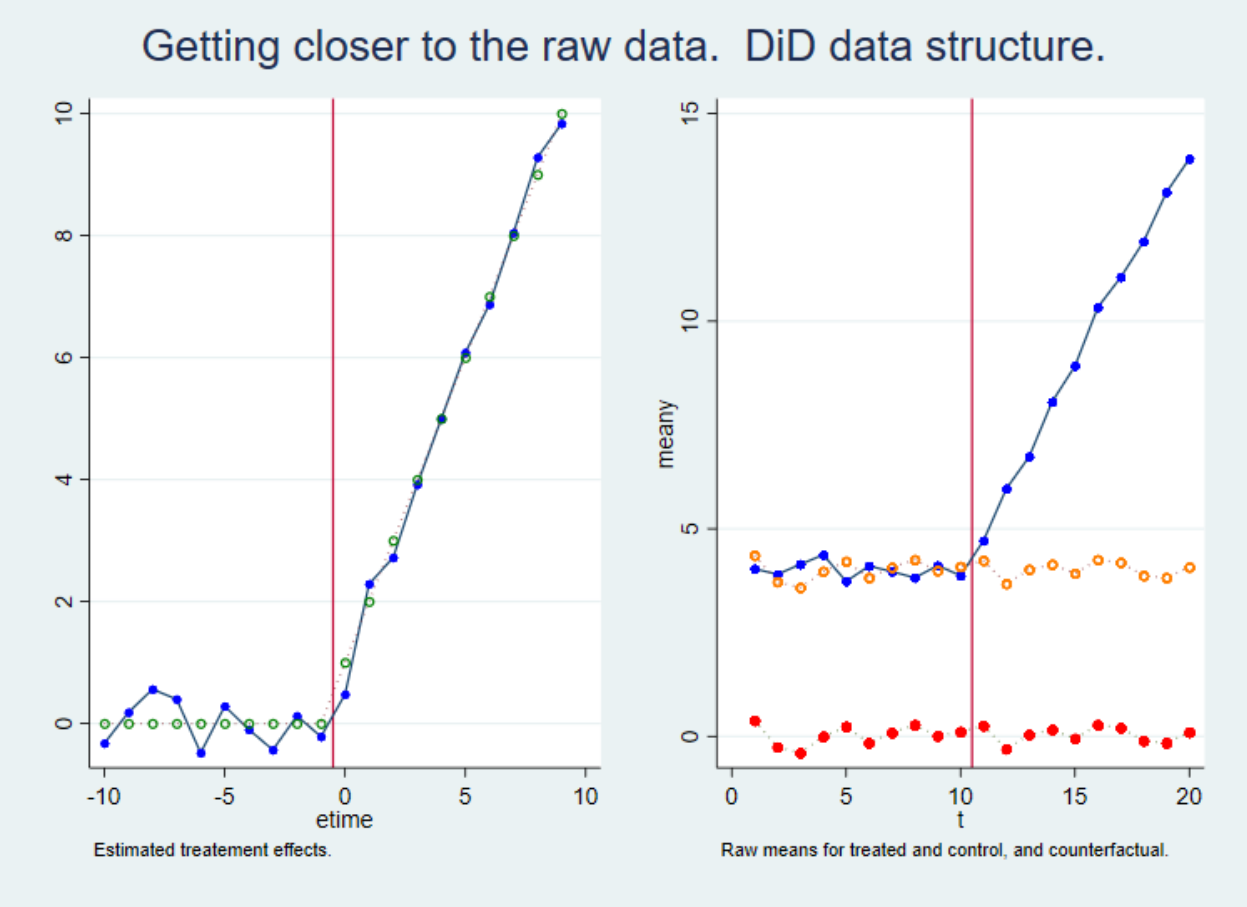
In figure A.7, we see an illustration of showing the counterfactual alongside the raw data. The data structure in the figure below is a Difference-in-difference data structure, with two unit types: (1) treated units sharing a common event date, and (2) control units. The first graph shows the estimated event study treatment effects; with the true treatment effects (the true γ_j from equation 1) superimposed in green hollow dots. The second graph shows the raw means for the treated (blue) and control (red) groups, and also shows the counterfactual untreated prediction for this group (orange hollow dots). The counterfactual is computed by subtracting off the estimated event-study effects ($\hat{\gamma}_j$) from the raw means for the treated group.

Next, figure A.8 shows a similar graph for a timing-based data structure. Here we have two treated groups, with an event date of 8 for group 1 and an event date of 12 for group 2. Here there are two counterfactuals, one for reach unit type.

E Pooling and Splines for event study coefficients

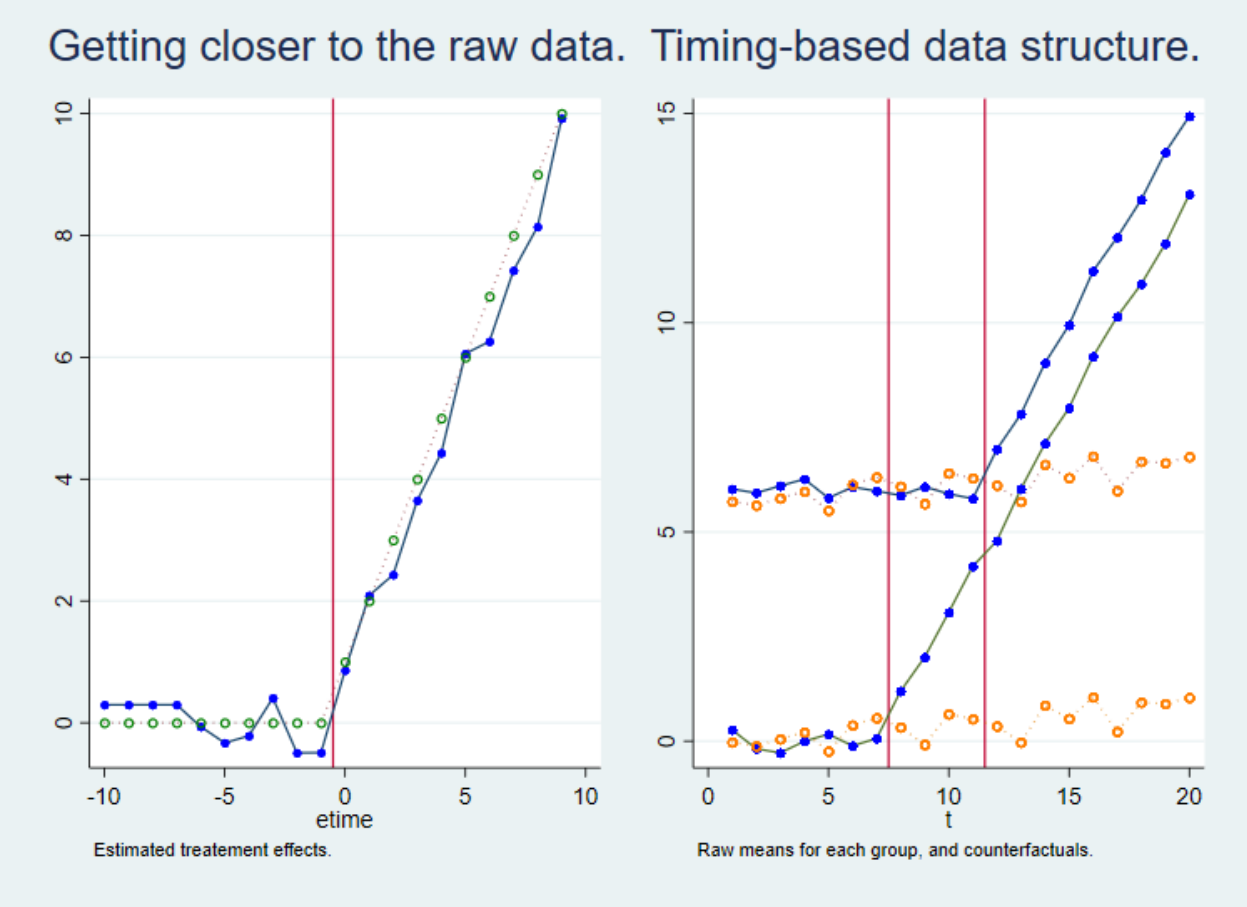
In this Appendix section I illustrate pooling event study coefficients, and imposing splines on event study coefficients for improved statistical power. These are discussed in section 3.6 in the paper.

Figure A.7: Event study coefficients vs. "getting closer to the raw data"



Note: In the left panel, the y-axis show the estimated (blue) and actual (green) treatment effects (γ_j). The x-axis shows event time. In the right panel, the x-axis shows calendar time. The red dots show the mean outcomes for the control unit. The blue connected line shows mean outcomes for the treated units. The orange dots show the counterfactual (untreated) outcomes for the treated units.

Figure A.8: Event study coefficients vs. "getting closer to the raw data"



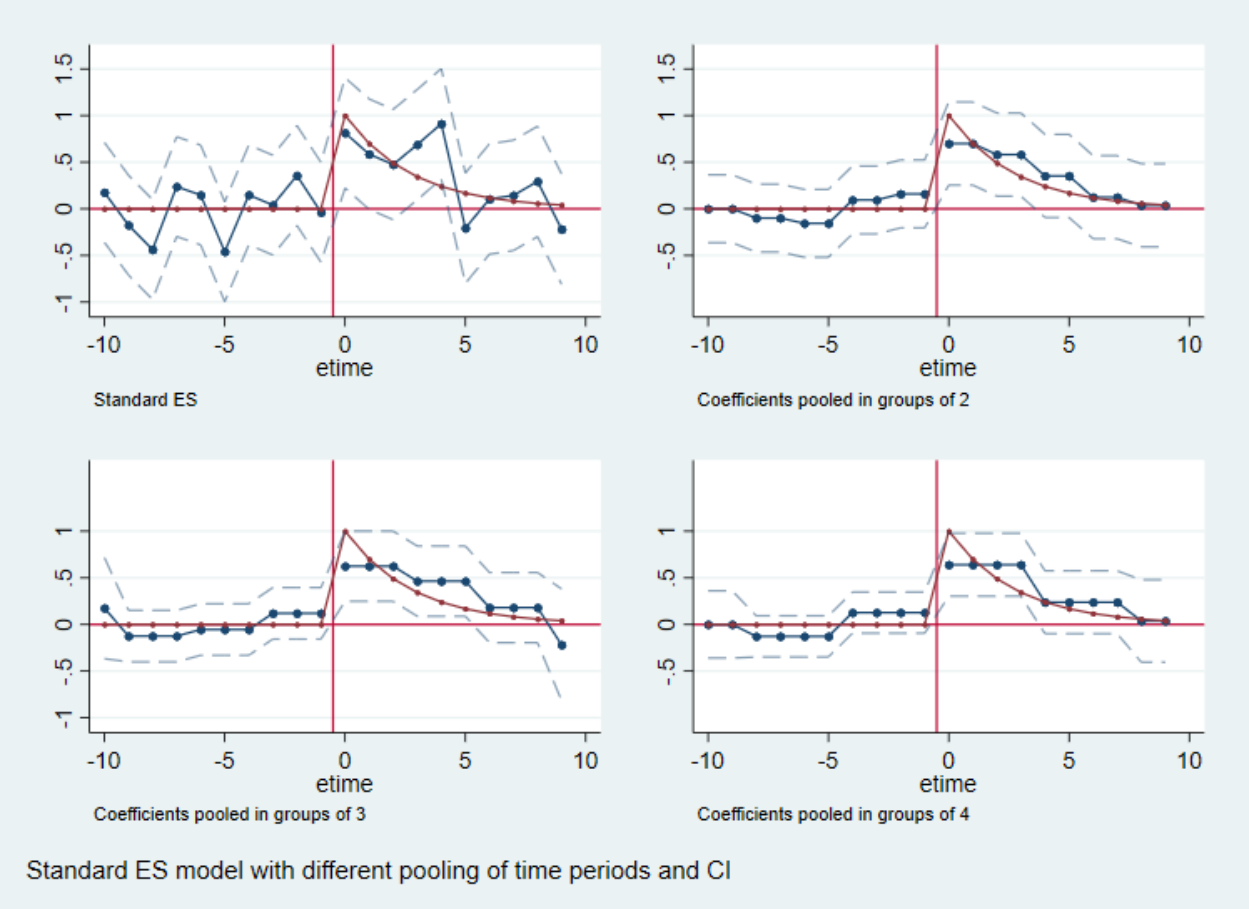
Note: In the left panel, the y-axis show the estimated (blue) and actual (green) treatment effects (γ_j). The x-axis shows event time. In the right panel, the x-axis shows calendar time. The blue connected lines shows mean outcomes for each of two types of treated units (who receive treatment at different dates). The orange dots show the counterfactual (untreated) outcomes for those treated units.

There are two way to implement pooling of event study coefficients. The first is to create pooled event time dummies, so that one dummy represents two or more adjacent event times. The alternative is to directly impose the pooling constraints at the point of estimation (e.g., using "cnsreg" in Stata). These two approaches are equivalent in standard cases. They could differ when other constraints are added in to the model: e.g. if imposing a "no pretrends" constraint, this could be implemented differently depending on how you are pooling.

Figure A.9 shows the impact of pooling constraints on the estimated results. For this illustration, the true treatment effects (shown in red) are designed to have a "jump, then decay" pattern. The top left graph shows (blue connected dots) a standard event study model, with no pooling. The top right model pools pairs of coefficients. For example, there is one estimate for "event time 0 or 1", and another estimate for "event time 2 or 3", and so forth. There is a noticeable shrinking of the width of the confidence intervals. The bottom left and right graphs pool sets of three and four coefficients, respectively. For example in the bottom right graph, there is one estimate for "event time 0 through 3", another estimate for "event time 4 through 7", and so on. In this example, greater averaging leads to improved statistical power (smaller confidence intervals), but worsening ability to capture the true dynamics of the treatment effects. To my eyes, pooling 2 or 3 event times together seems to be the best compromise for this data.

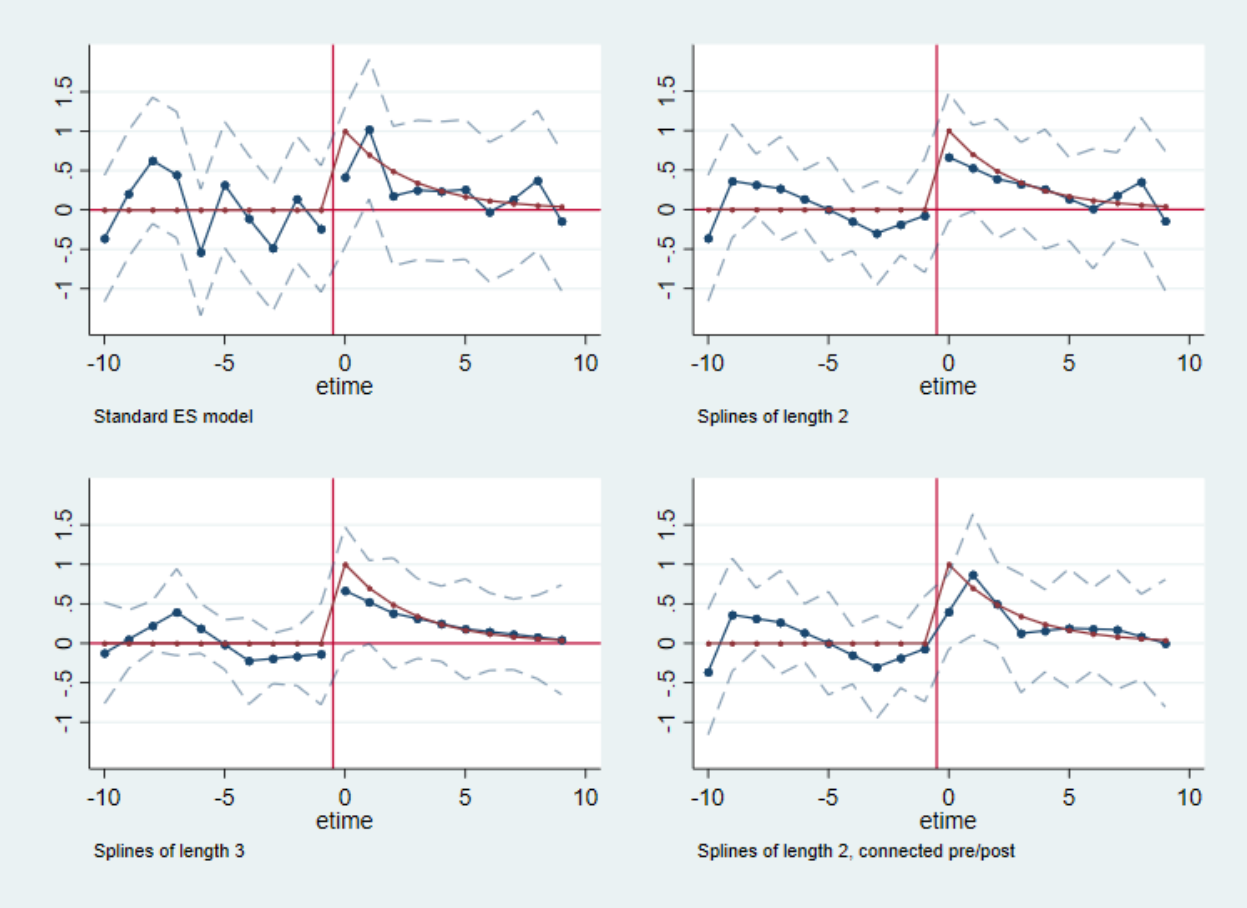
One alternative to pooling is to implement a spline model. This can be implemented by imposing "no concavity" constraints at the point of estimation. These constraints take the form of, e.g., $\gamma_1 - \gamma_0 = \gamma_2 - \gamma_1$ for connected segments of event time coefficients. Figure A.10 illustrates the use of splines to improve statistical power. The top left graph is the standard event study model with no splines. The top right graph imposes linear splines of length three. It allows for a break in coefficients between the pre-event and post-event coefficients. These splines improve statistical power moderately. The bottom left graph imposes splines of length four. The bottom right graph returns to splines of length three, but has the pre- and post-event time coefficients connected (the splines connect at the -1 segment). For this

Figure A.9: Pooling event study coefficients



Note: The top left panel shows a standard event study model with one parameter γ_j per event time. The blue dots show the estimated coefficients ($gam\hat{m}a_j$), and the red dots show the true treatment effects (actual $gamma_j$). The top right panel pools (groups) the event study coefficients into two-periods. The bottom left panel pools into groups of 3 periods, and the bottom right panel pools into groups of 4 periods.

Figure A.10: Splines in event study coefficients



Note: The top left panel shows a standard event study model with one parameter γ_j per event time. The blue dots show the estimated coefficients ($\hat{gamma}_j$), and the red dots show the true treatment effects (actual $gamma_j$). The top right panel constraints the event study coefficients to lie on a piecewise spline with segments of length 2. It allows for a break in the spine segments between the "pre" and "post" periods. The bottom left panel uses splines with length 3. The bottom right panel returns to splines of lenght 2, but forces the "pre" and "post" spline segments to connect.

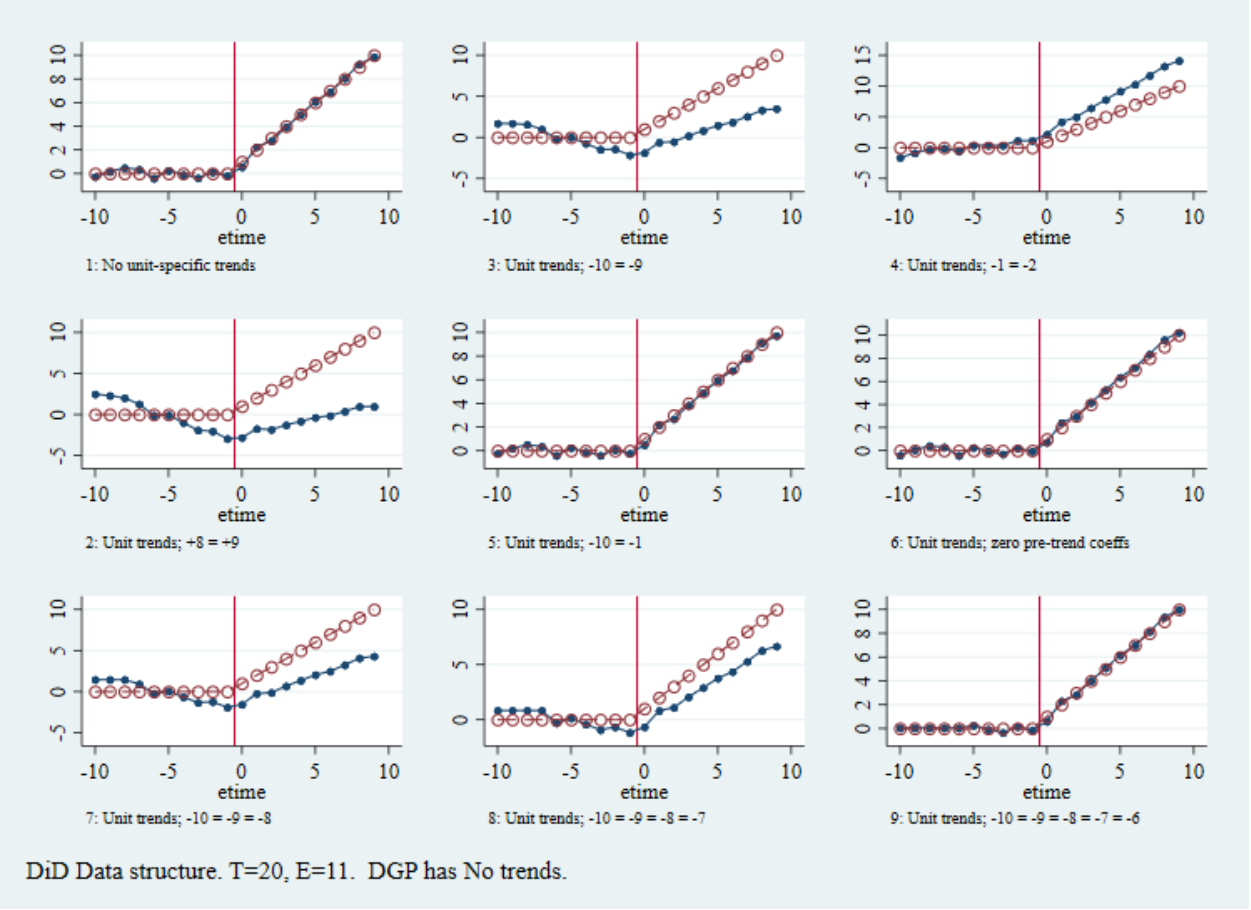
data generating process (DGP), this results in a mischaracterization of the effect at event time 0.

F Controlling for trends

F.1 DiD Data Structure

The DiD data structure is a good place to start, because it's easier to keep track of the possibilities for different terms to be multicollinear with each other. The simplest case to consider is one where we just add in one term: $time_t \cdot Treated_i$. However, this term is collinear

Figure A.11: Parameter restrictions when trend controls are included



Note: Each panel estimates an event study model on the same data, which are from a DiD type data structure. The true data generating process does not have any trends. The first graph does not include an estimated time trend for treated units, but the other 8 graphs do include this estimated time trend. Each panel employs different parameter restrictions in order to identify the model.

with the terms already in the model. So when we add this term some other constraint in the model will need to be added; there is no difference in the content of the specifications.

What can be tricky is that depending on what the restriction is, the estimated event study coefficients γ can look very different. To see this, consider Figure A.11. This shows 9 different models; many of which are equivalent.

The first graph has no trends included and serves as a baseline. Because the data generating process (DGP) here also has no trends, the event study coefficients (blue solid dots) match the true effects (red hollow dots). The remaining graphs add in a trend term for treated units; so each one needs one (or more) additional parameter constraints. The sec-

ond, third, and fourth graphs each impose those parameter constraints by equating the coefficients for adjacent terms. In the second graph we have an end-point for -10 and -9; in the third graph we equate the coefficients for -2 and -1; and in the fourth graph we have an end-point in the post-period, equating +8 and +9 terms. In each case, the event study coefficients have a zero trend through the terms that are equated; and the full pattern of coefficients pivots to reflect this normalization. As it happens, for none of these cases do the results look satisfactory.

For graphs 5 and 6, we impose constraints with the intention of having a flat pre-trend. Graph 5 equates the -10 and -1 terms. Graph 6 imposes a constraint that the pre-event coefficients have a zero average trend. Both of these restrictions give results that look good.

The last three graphs build on the idea of having an end-point in the pre-period, pooling terms. While graph 2 pooled only two terms (-10 and -9), graphs 7,8 and 9 each add in an additional term that gets pooled in. These produce results that look increasingly good. It might be the case that graph 9 is "too good"; once we've imposed that coefficients -10 through -6 are equal, and combine that with the pre-existing constraint that all the pre-event coefficients average to zero, this might have an implicit "zero trend" constraint.

F.2 Timing-based data structures and linear trend controls

F.2.1 Two unit types

Let's start from a data structure with two unit types, and $E_2 = E_1 + 1$. As noted above in section B.1, we now have an extra event-time coefficient we can in principle estimate, and so we need one additional restriction compared to the DiD data structure. Two common choices are to equalize two or more end-point coefficients at the beginning and/or end of possible event times; or to impose a flat pre-trend on event time coefficients.

Next we consider: what if we also want to add in trend controls? Suppose we want to control for $time \cdot 1(E_i = E_2)$, which allows for a different linear time trend for the unit-type with the later event date. It turns out that extra covariate is multicollinear with the

covariates already included in the model, in somewhat complicated ways. If we regress $time \cdot 1(E_i = E_2)$ on the RHS variables in (1), we will find that the event time γ_j parameters have a quadratic function in j; the δ_t have an opposite quadratic function in t; and the α_i parameters have a level shift based on unit-type. The result of this is for $(E_i = E_1)$ types, the γ_j and δ_t offset one another, leading to no trend. But for the $(E_i = E_2)$ types, their γ_j parameters are off-set, and so they have a linear time trend.

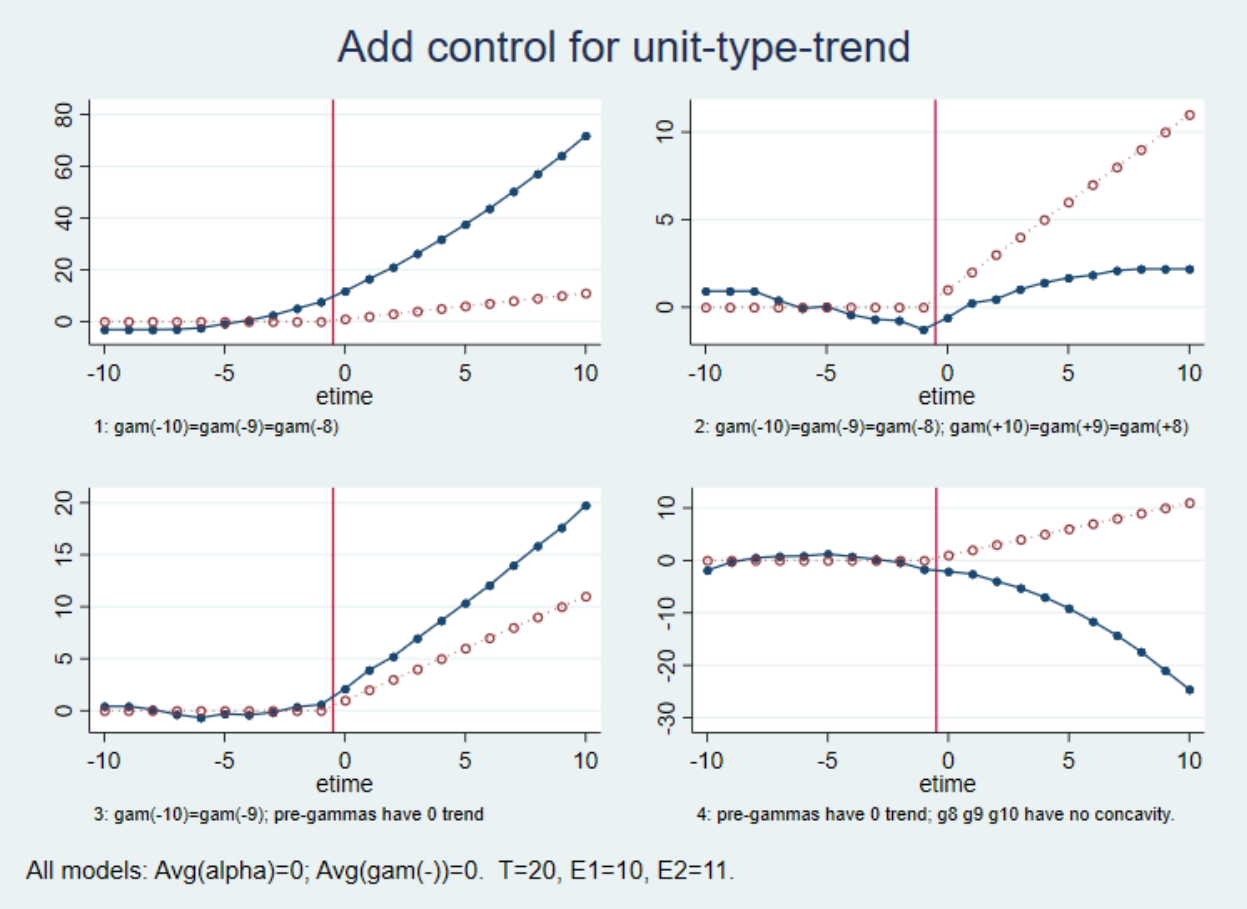
This complicated multicollinearity has two implications: (1) in order to get our model to be estimable, we will have to impose additional restriction(s); (2) these restrictions can interact with the complicated multicollinearity to produce unusual and unsettling results. Specifically, controlling for a unit-type linear trend can induce a quadratic relationship into the γ_j and δ_t parameters. This can interact with the additional parameter restrictions imposed to estimate the model in unsatisfactory ways. Even if the parameter restrictions are "true", the noise from the model errors will load on to the restrictions, and this can produce wildly incorrect counterfactuals. Figure A.12 shows estimated results from four seemingly reasonable parameter restrictions (indeed; the parameter restrictions in models 1, 3, and 4 are each consistent with the true model). None of these are very good. These weird results depend on the shape of the true treatment effect.

Next, figure A.13 shows results for the same restrictions as above, when the true treatment effect is a nice constant treatment effects step function. In this case, Model 2 is looking the best. But even there it's not so good. The take away message from this is to be extremely cautious when working with a timing based data structure and controlling for linear trends.

F.3 Getting closer to "raw data" when there are trends and trend controls

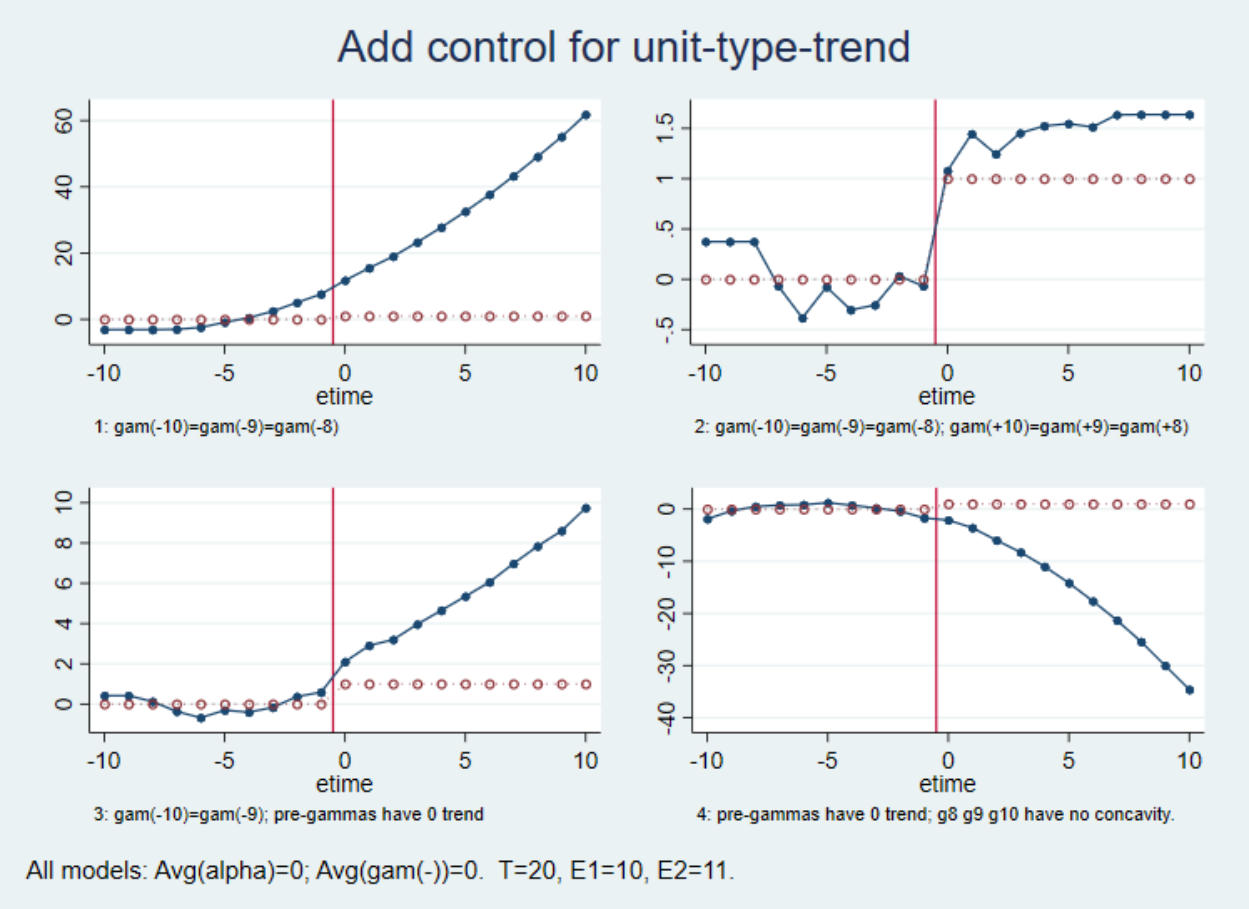
As in the case without trends, it is informative to show both the direct treatment effect estimates, as well as something that is closer to the raw data. Figure A.14 illustrates this, for three different models applied to the same data. Each model is in a different column,

Figure A.12: Timing based data structure and unit-specific time trends



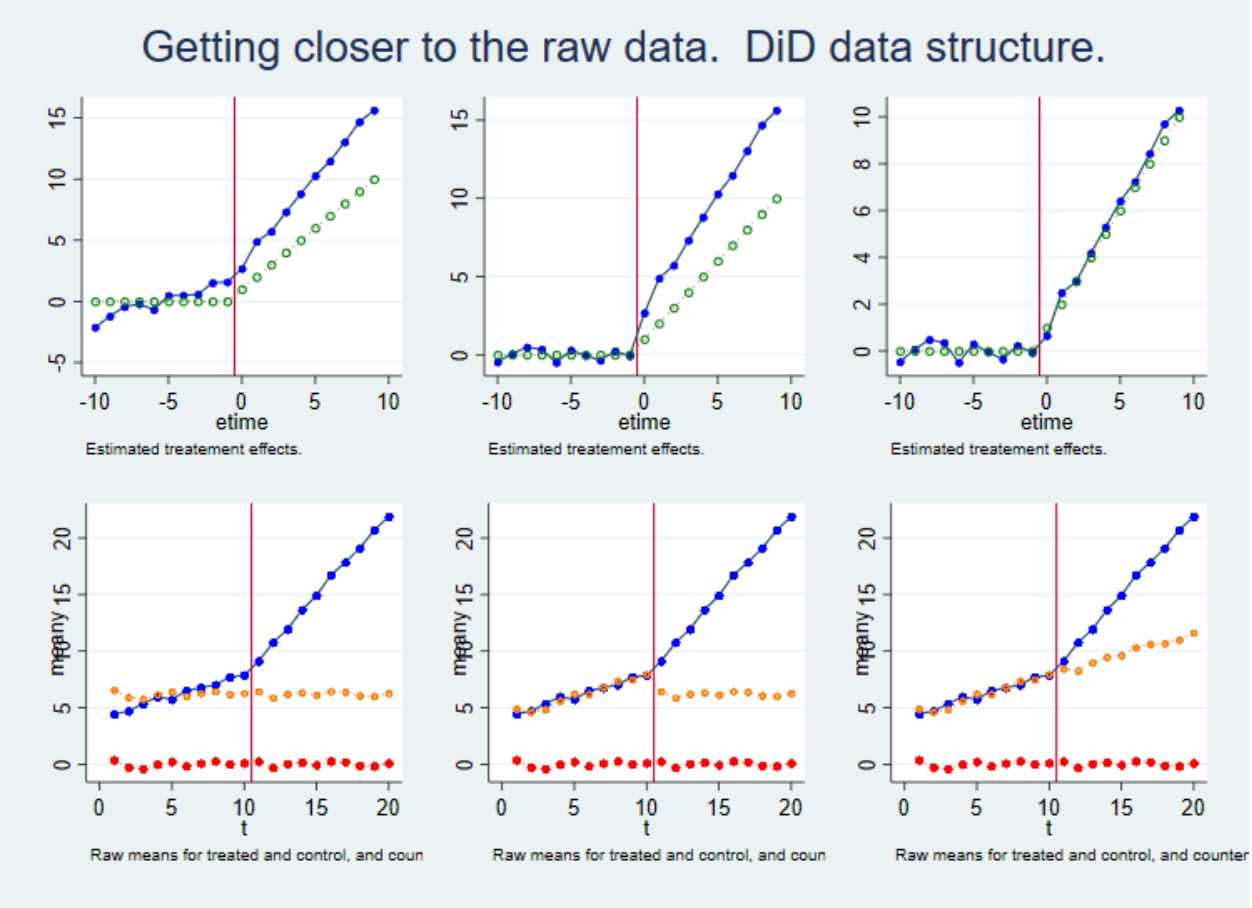
Note: Each panel estimates an event study model on the same data, which are from a timing-based data structure, with one unit treated at $t = 10$ and the other treated at $t = 11$. The true data generating process does not have any underlying trends. The true treatment effects (in red) follow a "ramp" pattern. Each panel includes an estimated unit-specific time trend, and employs different parameter restrictions in order to identify the model.

Figure A.13: Timing based data structure and unit-specific time trends



Note: Each panel estimates an event study model on the same data, which are from a timing-based data structure, with one unit treated at $t = 10$ and the other treated at $t = 11$. The true data generating process does not have any underlying trends. The true treatment effects (in red) follow a constant "step function" pattern. Each panel includes an estimated unit-specific time trend, and employs different parameter restrictions in order to identify the model.

Figure A.14: Closer to raw data with estimated trends in a DiD data structure



Note: The top row shows estimated (blue) and actual (green) treatment effects, and the bottom row shows corresponding raw data (and estimated counterfactuals). Each column relies on different parameter restrictions to identify the model. The blue dots show estimated treatment effects (top row) or raw averages (bottom row). The green dots in the top row show the true treatment effects (γ_j). The red dots in the bottom row show the raw averages for the control units, and the orange dots show the estimated "untreated counterfactual" for the treated units.

with the top graph showing the estimated treatment effects (in blue) along with the true treatment effects (in green), and the bottom graph showing the raw data (for treated and control units, in blue and red) and the counterfactual outcome implied by the estimated model (in orange).

In the data generating process, the treated units have a pre-existing time trend that is different than the control units. They additionally have a "ramp" treatment effect that increases in time once they are treated. The first model (top left and bottom left) are based on a model with no trend controls. This model shows the diagnostic pre-trend problem in its estimated coefficients; and that pre-trend translates into biased estimated treatment

effects. The second model imposes a "flat pre-trend" constraint on the estimated event study coefficients, but does not add in estimated unit-specific trend controls. This helps with the model fit in the pre-period; but the estimated treatment effects are just as bad as the first model. Without direct trend controls, the constraint on the event study coefficients does not fix the problem of trends. The third model adds in a unit-type trend variable, and imposes the "flat pre-trend" constraint on the event study coefficients. This is the preferred model, and it performs well. In each case, the bottom panel shows the raw data as well as the counterfactual implied by the model.

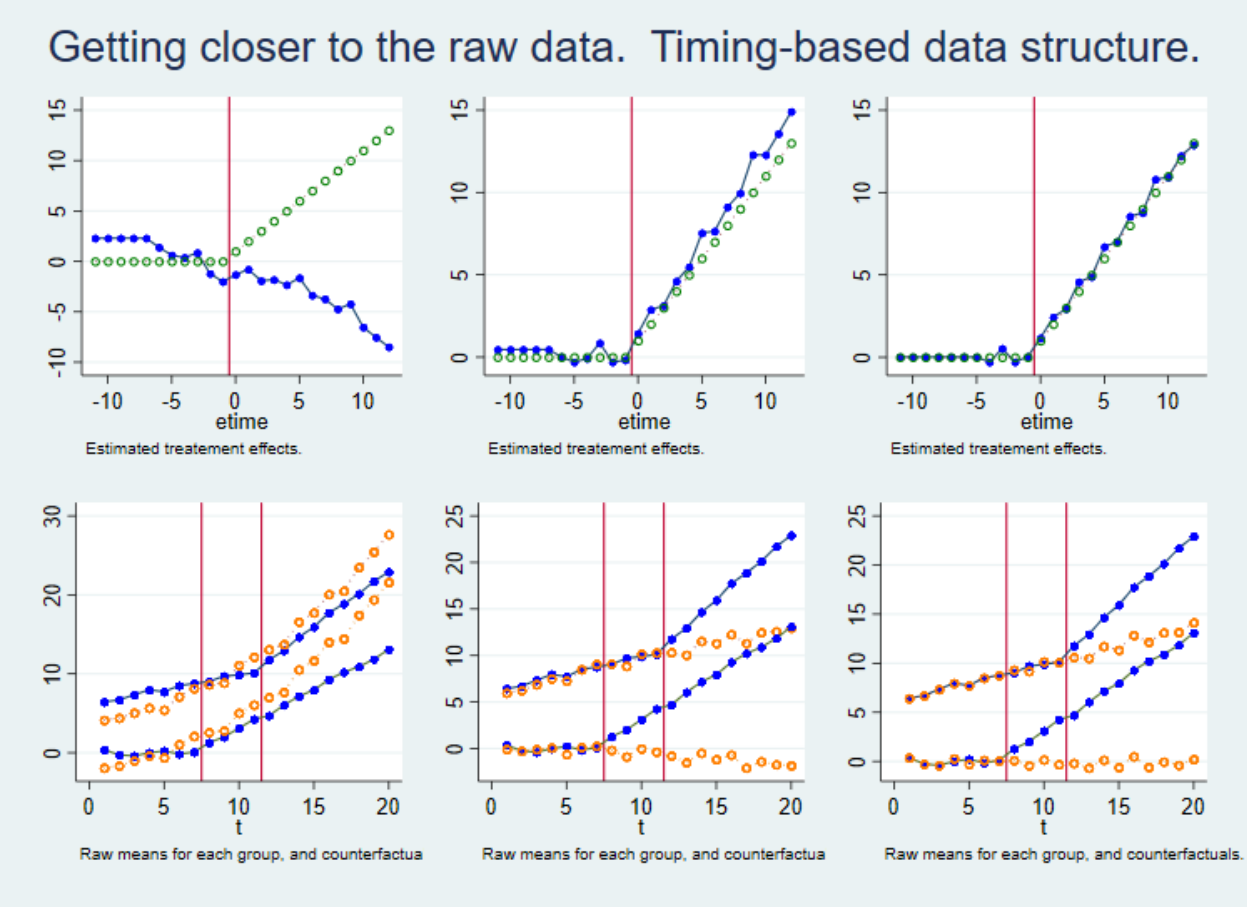
Next we consider the case of timing-based data structures and getting closer to the raw data. In the figure below, there are two unit types, one treated starting in period 8 and the other in period 12. The second unit type has a different underlying trend than the first. The figure shows three different models, one in each column. The top graph shows the estimated event study coefficients, and the bottom graph shows the raw data for the two groups, and the implied counterfactuals for each group.

In the first column, we do not control for any trends. The identifying restrictions are in the form of a pre-event pooled end point, and a normalization that the mean coefficient for the non-pooled pre-events is zero. We can see that (1) the model performs poorly, and (2) this could be diagnosed by examining the pre-trends. The second column adds in a unit-type specific trend shifter. Because this requires an additional constraint, it also imposes the "pre-event coefficients have zero trend" constraint. This constraint is applied to the same coefficients (-1 to -6) as the normalizing average-to-zero constraint. This model looks much better; although it is not perfect. The third column adds in additional two pre-event end-point constraints. This makes things look quite good.

F.4 Computational Issues with Unit-Specific Trends

In sistuations where it seems potentially useful to employ unit-specific trends, a useful approach is to "partial out" the unit-specific (or alternatively unit-type-specific) intercepts and

Figure A.15: Closer to raw data with estimated trends in a timing-based data structure



Note: The top row shows estimated (blue) and actual (green) treatment effects, and the bottom row shows corresponding raw data (and estimated counterfactuals). Each column relies on different parameter restrictions to identify the model. The blue dots show estimated treatment effects (top row) or raw averages (bottom row). The green dots in the top row show the true treatment effects (γ_j). The orange dots show the estimated "untreated counterfactual" for the two types of treated units.

trends. The partialing-out approach has the following steps. Step 1: For every variable z in the event study formulation—that is, all the characteristics of the unit variables as well as the indicator variables for when the event takes place, the dummy variables for each time period, and any added control variables, regress z on a constant and time t, only for observations in unit i. Then compute the residuals from this regression, $\tilde{z}$.[20] Step 2: Take the residuals from these regression equations, and then insert them into the event study model. Because you have already adjusted for time trends in the first step, you don't need to do any further adjustments for time trends in the second step—which means that the set of covariates will be modest in size. However, we need to take care in our second stage regression to impose the same parameter constraints that would apply to the one-step approach.

One limitation of this approach as described is that it controls for "overall trends" rather than "pre-trends," but this approach can be modified to partial out pre-trends only. To do so, in Step 1, estimate the model only on data up through the time period preceding the event. Then use this model to make predictions (and residuals) over the whole time period. For never-treated units, you can use the full time period. Step 2 is the same as described above. Goodman-Bacon (2021b) implements a version of this approach.

F.5 Beyond linear unit-specific trends

In general, unit-specific linear time trends allow for greater modeling flexibility. But even greater flexibility can be accommodated with more flexible unit-specific trends, like the use of higher-order polynomial trends. The greater flexibility can be good for avoiding interpreting secular time trends as a treatment effect. But it is a data-hungry approach, which requires adding additional parameter restrictions. The risks of over-controlling based on data from the post-period—and thus having estimates that are either biased, or less generalizable because they are based on idiosyncrasies in the data—can grow with increased modeling flexibility.

[20] To further save computational burden, z can be partialed out just once per unit-type. If our panel is balanced in calendar time, to save further computational burden, dummy variables for each time period can be partialed out only once, instead of once per unit.

An alternative approach is to control for covariates W_i that are defined at the unit level, interacted with linear or higher-order polynomial trends in time. For a somewhat extreme case, these covariates could be interacted with the calendar time dummies. I am not aware of guidance for assessing the value and risks of these alternative approaches.

G Comparing DiD models and ES models

G.1 Basic comparisons

Because the Event Study model can be written as a generalization of the Difference-in-Difference model, it is natural to compare the estimates from the two models. Roughly speaking, our intuition is that an average of the "post" ES coefficients, minus an average of the "pre" ES coefficients, should correspond to the DiD estimate. This lends itself to an informal diagnostic practice, which is to compare the ES coefficients and the corresponding DiD estimate. This can be done visually on your ES graph by plotting the DiD lines, with the pre-treatment line set as an average of the $s \leq -1$ coefficients, and the post-treatment line set to reflect the DiD treatment estimate. If the ES coefficients and the DiD estimates are meaningfully different, this can raise a warning flag for a potential problem, and is worth further investigation.

Although it feels intuitive that the DiD estimates and the ES estimates should line up, this is not necessarily the case. Several recent papers note how the two way fixed effects DiD estimate can be written as a weighted average of underlying 2x2 comparisons across units. In the presence of treatment effects that vary in time-since-treatment, the DiD averaging of these may not be what we would intuitively want at all. For example, Goodman-Bacon (2021a), Borusyak et al. (2022) and de Chaisemartin and D'Haultfoeuille (2020) all note that the some of the underlying treatment effects can get negative weight in the averaging, which can lead to strange results. Borusyak et al. (2022) and de Chaisemartin and D'Haultfoeuille (2020) each propose alternative estimators that can recover the treatment effects of interest

under some conditions.

G.2 Trending Treatment Effects can mess up a DiD specification, when we control for unit-specific trends

This subsection further develops the discussion in the main paper's section 4.3, which notes that the presence of trending treatment effects and controlling for unit-specific time trends can result in poor performance.

Figure A.16 considers a case where the treatment effect follows a "steady ramp" pattern. The basic DiD model (equation 2 in the main text) gives a sensible approximation; an average of the post-treatment effects. The ES model works well, as expected.

Suppose that we tried to control for unit-specific trends in our DiD estimation model. Because the treated units are trending up in the post-period, the trends will aim to partially capture that. This will narrow the estimated shift from pre-to-post; leading to downward biased estimates of the treatment effects in this version of the DiD model. This is shown by the unreasonably small estimates in yellow.

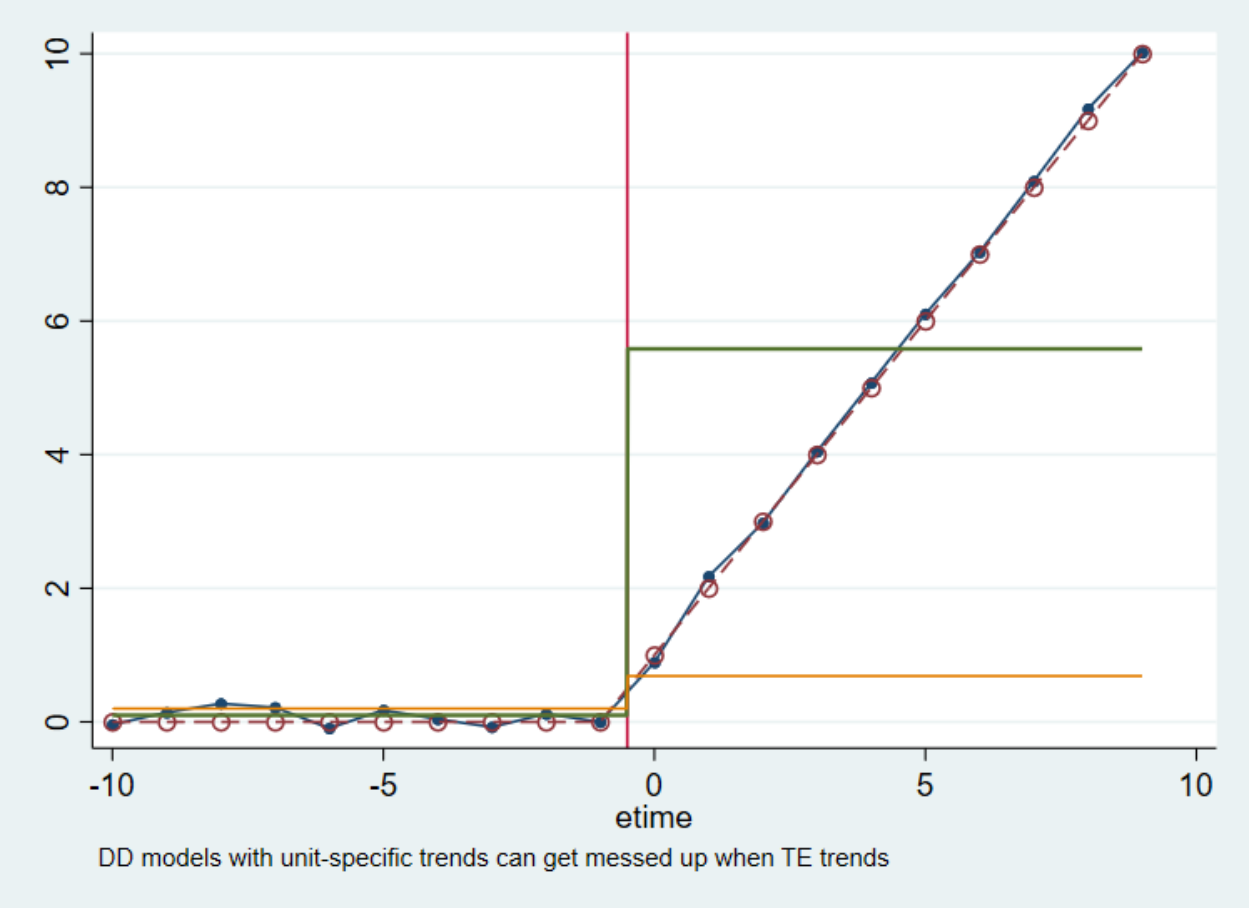
G.3 The Ben Olken Puzzle

This puzzle illustrates an example where DiD and ES coefficients give wildly different results. In this case, the ES coefficients are valid, and the DiD coefficient gives an unreasonable weight of zero to some of the ES terms.[21]

The simplest data structure to illustrate this puzzle is as follows: consider 2 units, each treated at a different time, with 3 calendar time periods. The Event Dates vary across the two units, $E_{i=1} = 2$ and $E_{i=2} = 3$. In the true DGP there are no calendar time effects or unit-specific effects: $y_{i,t} = 1 \cdot D_{i,t-1} + \gamma_2 \cdot D_{i,t-2} + \epsilon_{i,t}$. We consider the cases where $\gamma_2 = 1$ and where $\gamma_2 = 2$. We consider four estimation models, either an ES model or a DiD model,

[21] Many thanks to Ben Olken and Dan Fetter for conversations about this puzzle.

Figure A.16: Estimating a DiD model with trend controls when there is trending treatment effects can be problematic



Note: The hollow red dots are the true treatment effects (γ_j). The blue dots are the estimated treatment effects from an event study model. The green line gives the estimated treatment effect from a difference-in-difference model without unit-specific trend controls, and the yellow line gives the estimated treatment effect from a difference-in-difference model with estimated unit specific trend controls.

and either including or excluding calendar time dummy variables. To simplify, we omit unit-specific fixed effects. This produces results as follows:

		True DGP	
Estimation Model		$\gamma_2 = 1$	$\gamma_2 = 2$
ES Model (no δ_t)	$E[\hat{\gamma}_1]$	1	1
	$E[\hat{\gamma}_2]$	1	2
DiD Model (no δ_t)	$E[\hat{\gamma}]$	1	1.33 (OK)
ES Model (yes δ_t)	$E[\hat{\gamma}_1]$	1	1
	$E[\hat{\gamma}_2]$	1	2
DiD Model (yes δ_t)	$E[\hat{\gamma}]$	1	1 (uh-oh!!)

Here the Event Study models estimate coefficients that correspond to their true values. The DiD model does just fine when either $\gamma_2 = 1$ (constant treatment effects), or when there are no time fixed effects modeled (in this case, it averages a treatment effect of 1 with weight 2/3, and of 2 with weight 1/3).

The problem arises in the last row, when time fixed effects are included. Here the DiD model estimates a coefficient of 1, which places zero weight on the $\gamma_2 = 2$ ES impact. What is going on here? In the DiD model the "after*treated" coefficient is the same for both units for period 3; and the period 3 time dummy will make sure that the average is predicted correctly for period 3. So for period 3, two things are true: (1) there will be an unavoidable gap between the prediction and the realized values (with errors of $+0.5$ and -0.5 for the two units), and so (2) the treatment coefficient γ won't depend on the values of the period 3 realizations. So then γ is set to fit the unit-1 period-2 value ($\hat{\gamma} = 1$). There is a pathological collinearity between the time dummies and the model misspecification of the DiD model.

This example illustrates how a difference between the ES coefficients and the DiD coefficients can provide a nudge to dig deeper into the model, for a better understanding of what variation is driving the estimated effects. In this case, the ES estimates are valid, while the DiD estimate are distorted.