Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 203–230

An Introductory Guide to Event Study Models

Douglas L. Miller

The event study model is a powerful econometric tool used for the purpose of estimating dynamic treatment effects. One of its most appealing features is that it creates a built-in graphical summary of results. In one of the earliest papers in labor economics to use an event study model, Jacobson, LaLonde, and Sullivan (1993) sought to estimate the loss of income after being displaced from a job. Figure 1 reproduces a graph from that paper.

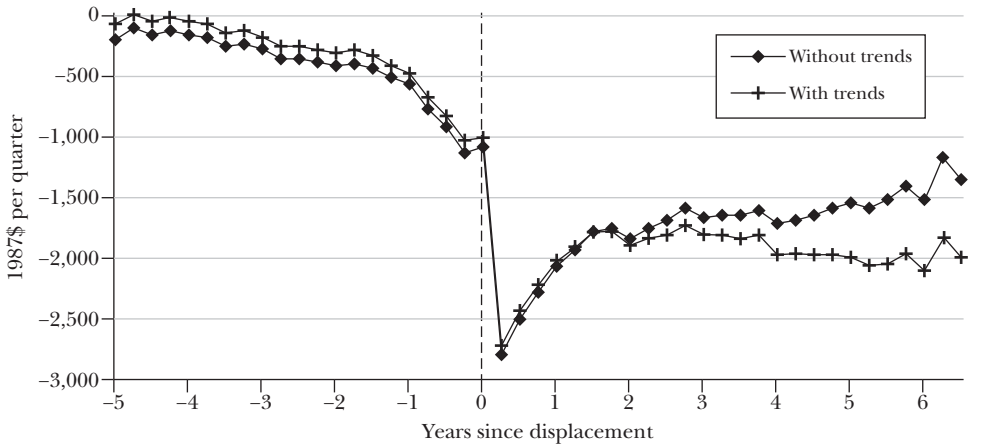
The *x*-axis is measured in "event time," meaning that for each person, the time of job displacement is treated as zero. The time-zero event is often referred to as the "treatment"—that is, the event or policy that changed what otherwise would have happened. The *y*-axis of the picture shows income for each period relative to a baseline comparison period. In this example, the baseline is more than five years prior to the job displacement.

The change after the event time of zero is the key takeaway from an event study picture, but the picture also reveals other rich patterns of behavior. For example, it also shows patterns before the event. Ideally, we hope that the line before the event is trendless, and deviations from that pattern alert us to a potential problem with our model; in particular, a trend suggests that the treatment may have been expected or that other factors are in play. In Figure 1, we see a modest deterioration in earnings in advance of the layoffs. This may reflect the presence of third factors—say, perhaps declines in demand for the output of a certain industry—that affect earnings prior to the event and that ultimately contribute to the displacement.

■ *Douglas L. Miller is Professor of Economics and Public Policy, Cornell University, Ithaca, New York. His email address is dlm336@cornell.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.203.

Figure 1
An Event Study Example: Loss of Income after Being Displaced from a Job



Source: Jacobson, LaLonde, and Sullivan (1993).
Note: Figure reproduced from Jacobson, LaLonde, and Sullivan (1993). The x-axis is measured in "event time." The y-axis show income for each period relative to a baseline comparison period more than five years prior to the job discplacement.

Alternatively, if displacement was anticipated and resulted in discouragement, this could lead to a pre-event trend through reductions in labor supply. The figure shows a modest pre-event trend, but also shows a sharp drop in earnings at the time of displacement, followed by a bounce back over the next two years that levels off at an earnings decline of about $500 to $1,000 per quarter compared to the pre-event level.

Event study models in economics started with finance applications: for a survey of earlier event studies in finance, see MacKinlay (1997). His earliest example is Dolley (1933a, b), who examines the effect of stock splits on trading activity, dividend payout rates, and market returns. In recent years, event study models have been growing in popularity. Currie, Kleven, and Zwiers (2020, Figure 4c) summarize trends in working papers from the National Bureau of Economic Research (1980–2018) and papers published in top economics journals (2004–2019). They document a sharply increasing share of papers using event study approaches, with an inflection point around 2012. Typically, event study models are estimated in a reduced-form "treatment effects" context.[1] Applications of event study models vary broadly, from job displacement (as in Figure 1), to school finance reform

[1] They can also be used to estimate statistical moments, which in turn can be used to estimate a structural model, as in Finkelstein et al. (2022).

(Lafortune, Rothstein, and Schanzenbach 2018), to the effect of trade liberalization (Braun and Raddatz 2008).

Behind the scenes of the easily digestible event study picture, a researcher needs to make a number of choices. Some choices are as obvious as the question of how (or when) to deal with pre-existing trends like those shown in Figure 1—and indeed that figure shows different estimates if the pre-event trend is taken into account—and some are more subtle, but researchers are often insufficiently clear about the choices they have made. In this essay, I discuss the range of decisions that go into an event study model, and in this way I aim to improve the understanding of these models for researchers, teachers, and consumers of this research.

For those who wish to dig a few layers deeper, a set of online Appendices provide more detail along with graphic examples and underlying code on related topics, such as connections from event study to difference-in-difference models, showing event study results in a way that is closer to raw data, pooling event study coefficients or using splines over event times to improve efficiency, additional considerations when controlling for pre-event trends, and other topics.

Core Features of Event Study Models

An event study model has two key elements: the estimating equation and the structure of the data.

Estimating Equation

The traditional approach to estimating an event study model is shown in this equation. We have units i and calendar time periods t; in the original example, the units are workers and the time period is calendar time (for example, earnings in the first quarter of calendar year 1982).

$$y_{it} = \underbrace{\left(\sum_{j \in \{-m,\ldots,0,\ldots,n\}} \gamma_j \cdot D_{i,t-j} \right)}_{\text{Event Study Terms}} + \underbrace{\alpha_i + \delta_t}_{\text{Panel Fixed Effects}} + \underbrace{\beta \cdot X_{it}}_{\text{(Optional) Control Variables}} + \epsilon_{it}.$$

On the left-hand side, the y variable shows the outcome. On the right-hand side, $D_{i,t-j}$ is an indicator variable for event time j, meaning that the event took place j periods before this observation's calendar time. A separate term is included for each event time. The key features of this specification are the $\gamma_j \cdot D_{i,t-j}$ terms. The coefficients after the event has occurred (γ_j for $j \geq 0$) capture the dynamic effects of the treatment as these effects manifest over time since the event. The terms γ_j for before the event has occurred (for $j < 0$) provide a placebo or falsification test. In the absence of anticipation effects, model misspecification, or omitted confounding variables, these pre-event terms should not have a trend in j. Together, this part of the regression equation terms will trace out a graph similar in appearance to Figure 1, measured in event time.

The index t represents the "calendar time" in which we observe the outcomes. The index j represents time-since-event, or "event time." In Jacobson, LaLonde, and Sullivan (1993), event time would be interpreted as, for example, "two quarters after job displacement" (for $j = 2$). In many applications, with at most one event per unit, we can designate the "Event Date" E_i, which is the date that the event occurs. The connection between these three variables is $j = t - E_i$. However, names and labels for the $\{t, j, E_i\}$ variables are not standardized across the literature. As you read event study papers, take care to check your understanding of what names and labels are used for each time concept. The constants m and n determine the endpoints for the estimated event study terms.[2]

Event study models are estimated on data that have a panel structure. It is conventional to add two sets of fixed effects, α_i and δ_t, for unit and time fixed effects. These serve the role of controlling for confounding omitted variables that vary at the unit or time level. Using this two-way fixed events approach helps to isolate the effect of the event. The outcome variable $y_{i,t}$ may also be influenced by other underlying factors. Thus, some event studies add other control variables X_{it}.

Sometimes our events occur at a different level of aggregation than our data. For example, perhaps an event occurs at the state-year level and we are working with a repeated cross section of individual-level data. It is okay to define the event dummies based on the state-year variation and to keep our regressions at the individual level (incorporating cluster- robust standard errors so that inference accounts for the more aggregated level of the event study dummies and their correlation over time within a state). This approach can be useful if we want to control for individual-level covariates; that is, even though we are working with a repeated cross section of individual-level data, we still conceptually have a panel at the state-year level. It can also be okay to first aggregate our data up to the state-year level and then run the model at that level. This makes the dataset more manageable. If we do this, I think it makes sense to weight our aggregated observations by the population represented in each state-year cell in order to get closer to results we would have obtained from the micro data.

Event Study Data Structures

In the panel data used by event studies, units may have an event (in the basic model) or else multiple events (in a more complex model) that occur at certain dates. An event study data structure can be defined based on an understanding of the unit types in your dataset. Two key questions are: (1) Are there "never treated" units or not? (2) Is there (a lot of) variation in the treatment date across units? A researcher needs for the answer to one or both of these questions to be "yes." The combined answers to these two questions represent different data structures, with corresponding differences in the thought experiment behind identification of the treatment effect coefficients. A key theme in this paper is that the options,

[2] In some applications, the time variable t is based on birth cohorts instead of calendar time. This possibility is discussed further later in the paper.

Table 1
Data Structures for Event Study Estimation

	Only Ever-Treated Units	*There are Never-Treated Units*
Common Event Date	N/A	DiD-type
Varying Event Date	Timing-based	Hybrid

Note: Author's proposed labels for event study data structures, based on whether the analysis data sample uses never treated units or not, and on whether treated units have a common event date or varying event dates. "DiD-type" = "Difference in Difference type."

guidance, and conclusions for an event study can depend on the data structure with which we are working.

Table 1 lists the possibilities. In the top-left corner, if we answer both questions with "no" we have only treated units and they share a common event date. In this setting, we cannot separate the effects of the event from other confounders that occur in calendar time, and so cannot identify treatment effects.

If we answer "yes" to the first question and "no" to the second question, then the data include both treated and untreated units, while all treated units share a common event date. The never-treated units help to identify the change in counterfactual outcomes across calendar times. Then the treatment effects can be estimated. In the canonical event studies graph, the treatment effects line represents allowing for over-time changes in the treated group and over-time changes in the untreated group and then looking at the differences between these changes.[3]

In the timing-based data structure, there are only treated units and the event dates vary. A leading example is when different geographic units (or perhaps individuals) all experience the same policy change or treatment, but they experience the change at different (event) dates. Here, the underlying thought experiment is that the timing of the event is as good as random, and so those treated earlier or later can serve as controls for one another. Dobkin et al. (2018) have a timing-based data structure in their study of the effects of hospitalizations on expenditures and labor supply. All of the individuals in their study experience a hospitalization, but they do so at different times.

Sometimes with this data structure, researchers make descriptive event study graphs that omit calendar-time fixed effects and unit fixed effects. For example, Card, Heining, and Kline (2013) track German workers who transition jobs across firms, based on the quartile of wages at the old and new firm. In a

[3]Indeed, the event study specification is a generalization of a standard two-way fixed effects difference-in-difference specification:

$$y_{it} = \gamma \cdot Treated_i \cdot Post_{i,t} + \alpha_i + \delta_t + \beta \cdot X_{it} + E_{it}.$$

Here $Treated_i$ is a binary variable for units that ever receive treatment, and $Post_{i,t}$ is a binary variable that indicates that treatment has occurred. If we restrict the pretreatment coefficients from the earlier equation to be zero, ($\gamma_j = 0$ for $j < 0$), and restrict the post-treatment coefficients to have the same value ($\gamma_j = \gamma$ for $j \geq 0$), then the traditional equation approach shown earlier and this regression here are equivalent.

separate example, Chetty et al. (2014) track Danish workers who transition to jobs with greater defined contribution pension shares. The event study graphs shown in these two examples are essentially expanded pre-post designs. Their credibility comes from three factors: (1) an a priori expectation that the pre-move outcome provides a reasonable counterfactual, (2) the visibly flat pre-trend in the raw data, and (3) the stark jumps at the time of job transition. These graphs have no unit or calendar-time fixed effects and are based on balancing the dataset in event time rather than calendar time.

The data structure might combine variation in event dates and both treated and untreated units. I label this the "hybrid" data structure, and it will include both sources of identification: the comparing of treated and control units and timing-of-event. This data structure is common in event studies. One application that employs a hybrid data structure is the Jacobson, LaLonde, and Sullivan (1993) study mentioned above. They pool data on workers who were displaced at different dates from their jobs as well as workers who were never displaced. Another example is Lafortune, Rothstein, and Schanzenbach (2018), who examine the impact of state-level school finance reforms on funding and test scores. They have 26 states which implement reforms, across a wide range of implementation dates spanning 1990–2011. They also incorporate states without reforms in this period into their analysis.

Estimates from the hybrid data structure can be (informally) compared to estimates relying solely on the timing-based subset of the data (estimated using only ever-treated units) to see whether the different sources of variation are producing similar estimates. So far, I have not seen a formal approach or recipe for making this type of comparison, but I think it could be a useful addition to our standard practice.

When carrying out or interpreting an event study, it is important to be explicit with your reader about the data structure. It is also best practice to show your reader the distribution of observations across event times in your sample. In Appendix A, I place these data structures in the context of related difference-in-difference models. I also illustrate a couple of graphical ways of showing the variation in your unit types and other key aspects of your data structure.

Parameter Restrictions

The basic event studies model includes more parameters to estimate than is possible. Remember, the total number of parameters comes not just from the γ parameters for the treatment effects over each time period, but also from the α and δ fixed effects parameters on units and times and potentially from more parameters if the researcher decides to include additional control variables. More important than a simple count of parameters is the fact that the event-time dummies are multicollinear with the combination of unit (for example, state-level) and time (for example, calendar year) fixed effects.[4] To proceed, we need some restrictions

[4]This multicollinearity is due to the fact that event time, calendar time, and event date are connected by $j = t - E_i$, and that event dates can be defined by unit dummies.

on these parameters. It is useful to group these restrictions into three (potentially overlapping) categories: (1) standard restrictions on panel fixed effects parameters; (2) restrictions that help to define our desired counterfactual; and (3) potential additional restrictions that are required to address concerns about multicollinearity.

In an event study model, the event-time coefficients γ_j in the traditional equation approach shown earlier are our main coefficients of interest. They estimate the treatment impact j periods after receiving treatment. This treatment impact needs to be defined in reference to a specific counterfactual. That definition is embodied in parameter restrictions. For example, we might think of a difference-in-difference-type counterfactual as "compared to a pretreatment period, how much change we would have expected to have occurred in the absence of treatment." Thus, probably the most common normalization is to choose a specific pretreatment event time and normalize the corresponding coefficients to average to zero. For example it is a common choice to set $\gamma_{-1} = 0$, by excluding the dummy variable for the $j = -1$ event time from the regression. Alternatively, we might have experimental assignment to treatment and control unit types. In this case, our normalizing assumption might be that those in the untreated group can serve as a control group for those who are treated. We would, therefore, have all of the event-time dummy variables but omit the unit fixed effects, setting $\alpha_i = 0$.

Multicollinearities abound in event study models. At a basic level, the sum of the unit dummies is equal to one, and the sum of the calendar-time dummies is equal to one. This introduces a multicollinearity between these two sets of dummies as well as the intercept, typically requiring dropping one from each set. There is also an additional multicollinearity between the event-time dummies $D_{i,t-j}$ and the unit and calendar-time dummies. Sometimes, once we have made basic restrictions on fixed effects and to define the counterfactual, the remaining parameters in the model can be identified and we are good to go. But this is not always the case. The problem of multicollinearity is especially prevalent in a "timing-based" data structure, where all units are treated but their event date E_i varies. In this data structure, the event-time dummy variables, unit dummy variables, and calendar-time dummy variables will have one or more additional multicollinearities, and so additional restrictions are needed in order to proceed.[5] The problems of multicollinearity also compound when we directly add in unit-specific time trend controls.

How should we implement our additional required parameter restrictions? In current practice, a common approach is to let the software (like Stata) automatically choose some collinear variables to drop, with unknown and possibly problematic implications. This approach should be avoided, and my recommendation is to check your regression output carefully to ensure that no variables are being unexpectedly dropped.

[5] See Proposition 1 in Borusyak, Jaravel, and Spiess (2022) and section 2.4.2 of Schmidheiny and Siegloch (2023). I discuss the number of needed parameter restrictions further in online Appendix B.1 and different examples are illustrated in online Appendix C.2.

Another common approach is to pool some of the data by grouping several of the treatment variable γ's in the tails to be equal. In the traditional event study equation, this would mean including an "end-cap" dummy variable such as $D_{i,t \leq E_i - m}$, indicating "the event will happen m or more periods in the future." This approach can sometimes be okay, but it can be problematic if there are uncontrolled-for underlying trends or (for a posttreatment end cap) if the treatment effects themselves are trending. It should only be used if these concerns seem unlikely to be important. As an alternative, it is possible to apply milder but still-useful constraints. For example, you can focus your parameter restrictions on the pre-event coefficients. I discuss "end caps" more in the next section.[6]

It's not always obvious when our model is okay as is or when additional restrictions are needed. When researchers need to impose additional restrictions to identify the model, we should keep in mind the following: (1) these are not merely formalities—the treatment effect coefficients γ_j we estimate are directly dependent on the restrictions imposed; (2) these restrictions are untestable, at least in part; and so (3) we want for these to be as uncontroversial and "obviously true" as possible. Indeed, (4) because of the "multicollinearities abound" nature of some event study data structures, our main estimates of interest can be unexpectedly sensitive to these extra restrictions. This can result in (5) "small bits of noise" propagating through our model in unexpected ways. This last fact can sometimes argue for employing additional restrictions beyond what would be minimally necessary.

My main recommendation is to be clear and explicit about what restrictions are being imposed. Going forward, it would be useful if all event studies would clearly report (1) the number of categories for each group (time, unit-type, or unit) of dummies and/or event study coefficients, both the total possible as well as those that are included in our actually estimated specification (after variables are dropped due to collinearity); (2) the constraints we (or our statistical package) impose on the estimation, either directly or through dropped terms; and (3) a direct assessment of the identifying variation in your data structure (for example, by computing the rank of the relevant proportion of your X matrix).

Event Study Specification Choices

This section outlines some of the main specification choices to be made when estimating event study models and discusses the trade-offs involved.

Choice of Pre-event Reference Period

When estimating an event study model, a common choice is to use "one period before treatment" as a normalization, so that the γ_{-1} coefficient is set equal to zero

[6] In online Appendix B.2, I offer more detail and examples for parameter restrictions, including some discussion of useful Stata commands.

in the time period immediately before the event. In the traditional event study equation presented earlier, this is implemented by dropping the -1 event-time dummy variable. But instead of blindly choosing the period immediately before the event for the normalization, it is better practice to make a judgment call as to what is a reasonable pre-event window, balancing considerations of "close enough to be the appropriate counterfactual baseline" and "more data allows for more precision." Then all of the event dummies can be included and the γ_j coefficients constrained to average to zero within the pre-period window.

How long of a pre-event window should a researcher choose? There is no hard-and-fast rule. I think it is useful to consider the pre-event window you would choose if you were estimating a simple difference-in-difference model. If you chose just one pre-event-time period, you might be worried about the extra statistical noise this would bring. As your pre-event window extends farther back, at some point you might get increasingly worried that those time periods become less appropriate for your counterfactual. In the end, for your difference-in-difference model you would make a judgment call, trading off these two considerations. It seems sensible to have this same judgement call inform your choice of the pre-event reference period.

Normalizing to zero over several event times, rather than just the period immediately before the event, has two effects on the canonical event studies graph. Choosing a longer time period has the effect of shifting the whole pattern of coefficients up or down—while retaining the same shape. The other effect is that when a more extended reference period is used for normalizing to zero, the standard errors can be noticeably smaller. The reason is that when using a single reference time period there is additional uncertainty driven by the noise in this term on its own, which tends to make the standard errors larger.[7]

If we normalize to a broader reference period, our search for a trend before the event will manifest itself differently than if we had normalized the –1 coefficient to zero. We need to assess the overall trend in coefficients rather than examine pointwise coefficients and their difference from zero. (This is also illustrated in online Appendix C.1.)

When we suspect (or see) a dip in outcomes shortly before the event, we might speculate that this is driven by some process that is bundled with the event and which is playing out shortly before the event as it is recognized in our dataset. In this case, we probably do not want to use the period of the dip as our counterfactual baseline because it is actually part of the treated period, even though nominally it's before treatment.[8] Instead we could define our baseline counterfactual to be a period prior to the beginning of the dip. For example, in Figure 1 we see a dip in

[7] This is illustrated in online Appendix C.1. Also, online Appendix C.2 illustrates the potential impact of different normalizations within a timing-based data structure.

[8] A dip that occurs just before the event is sometimes called "Ashenfelter's dip," after Ashenfelter (1978), who studied the impact of job training on earnings. Ashenfelter's models were not presented in the now-traditional event study graphical format, but his table's results have an event study framing, including showing a pretraining drop in earnings.

earnings prior to the layoff. Inspection of the figure suggests that we would want to have our reference period be at least one year prior to the layoff. In Jacobson, LaLonde, and Sullivan (1993), the authors chose "5 or more years prior to the layoff" as the reference period.

Show More than the Estimated Treatment Effects

An event study provides a treatment estimate as a single set of numbers. However, it is good practice to get closer to the raw data by also reporting a combination of actual and counterfactual average outcomes separately for each unit type. These graphs will complement each other in terms of the information provided. For example, when the event study allows comparison of treated and untreated units, this presentation allows readers to assess whether the unit types experience parallel trends during periods when treatment status is unchanging. Both the difference in levels and in the trends can provide important context for interpreting the treatment effects.

We can also add to this plot a line for the counterfactual untreated prediction that applies to the treated units. To generate this, here are the appropriate steps: (1) estimate the event study model; (2) "zero out" the event-time dummies and make predictions; (3) average these predictions within calendar time for the treated units; and (4) plot out this counterfactual. This calculation lets us see both the raw data and the estimated treatment effects. It also implicitly shows the content of the normalizing restrictions of the model. For example, if we are normalizing the pre-trend in event studies coefficients to be zero and are controlling for unit-type trends, this will show up in a trending counterfactual line. For timing-based or hybrid data structures, this lesson is slightly more complicated to apply. However, the researcher can still plot the average time series for each unit type and then supplement this by adding the counterfactuals for each unit type. (These ideas are illustrated in Appendix D.)

Choices with Control Units: Selection and Re-weighting

Suppose that we are carrying out an event study that includes both treated and untreated units (for example, individuals or states), with untreated units as the control group. However, sometimes we might worry that the never-treated units could be problematic comparisons for the treated units. For example, Krolikowski (2018) reconsiders the Jacobson, LaLonde, and Sullivan (1993) example that generated Figure 1 presented earlier. In the 1993 paper, the event is "first observed layoff"; never-treated units are therefore individuals who never experienced a layoff. However, subsequent layoffs can only occur for the treated group. Thus, there is a mechanical difference in the future earnings potential of the treated group compared to the control observations, above and beyond the effect of the first layoff under consideration. In addition, the control group may be positively selected with regard to unobservable skill, labor force attachment, and/or job match quality. In this setting, those who are never laid off may not provide a good counterfactual for outcome for treated individuals; indeed, the use of this control group could make the impacts of the layoff look worse than they actually are.

There are a range of options to have the control units (for example, individuals who did not experience job displacement) offer better counterfactuals, with the overall goal of making the assumption that "the control units tell us the counterfactual over-time changes" more plausible. First, a researcher might exclude a subset of the control units because they are in some way unrepresentative or because they experienced unusual shocks. For example, if you are working with a city-year panel, and your treated cities are all medium- or large-sized, then you might consider excluding small cities from the control units that you use. Second, for the time periods before the event, it is possible to check for parallel trends between the control and treated units. Third, one can look at the degree of similarity between treated and control units along a number of dimensions, using covariates.

Finally, the researcher might use a reweighting or matching procedure prior to estimation of the event study. A reweighting procedure would apply different weights to the never-treated units so that the covariates match the treated units. In a study of the impact of the introduction of the Legal Services Program (during the 1960s) on demographic outcomes, Goodman-Bacon and Cunningham (2019) observe that untreated counties are different in their observables compared to treated counties. To address this, they estimate a cross-county first stage model to obtain propensity scores (specifically, the probability of being a treated county). They then re-weight the control counties to be more representative of those treated.[9]

An alternative is to choose one or more never-treated "matches" for each treated unit. These matches would typically be made based on observable covariates, possibly including some values of pre-event outcomes. Some practitioners choose to combine these approaches with assigning a pseudo-event time to each control unit, in an effort to present a more plausible counterfactual outcome path. I am not aware of a systematic look at possible trade-offs involved in the choice to use pseudo-event times for the control units.

If an event study has a hybrid data structure, a researcher has the option of discarding the data from untreated units and focusing instead on a timing-based strategy. This approach that would be based on the belief that "among those treated, timing of treatment is as good as random" is more believable than the assumption that "control units tell us the counterfactual over-time changes." On the other side, using never-treated units will bring in more data, usually improving statistical power and requiring fewer parameter restrictions in order to identify the model. The trade-off between these two considerations will vary on a case-by-case basis. Whatever approach is chosen for dealing with never-treated units, it is useful to show sensitivity of the results to alternate approaches.

[9]This approach could in principle also be used in situations that use only ever-treated units. If there is reason for concern over possible differences between, say, earlier-treated and later-treated events, We could use reweighting to balance covariates across "early event date" and "late event date" units prior to estimating the model.

Choice of Event Window

In some cases, data availability will limit what endpoints m and n can be used for the event-time window; otherwise, you need to make an explicit decision. On one hand, making the event window as wide as possible allows us to see a long path of dynamic treatment effects, and for the pre-event coefficients it gives us a long window to detect troublesome patterns. This consideration pushes toward including as many event-time lags as possible.

The main competing consideration is that we would ideally like for the event-time coefficients γ_j to all be estimated off of the same set of units. For example, in Jacobson, LaLonde, and Sullivan (1993) the events (job displacements) occur between 1980 and 1986, and the outcome (earnings) data are observed for the period 1974–1986. The event-time coefficients for "zero years since displacement" in Figure 1 are based off of all displacements. But the coefficients for "five years since displacement" can only be estimated for displacements that occur in 1980 or 1981. This means that the event-time coefficients post-displacement are estimated off of different sets of individuals. If there is something systematically different about the early- or late-displacement individuals, that could challenge interpretation of the coefficients. Even without a systematic difference, there will be a loss of statistical power as fewer units are available to identify the more remote coefficients further from the event itself. These considerations suggest if possible choosing the endpoints of the event window so that most or all coefficients are identified off of a balanced set of units. It also reinforces the need to show your reader the distribution of data across event times (as illustrated in Appendix A).

Depending on your data setup, there may be a straightforward resolution of these competing concerns. Suppose that the span of event dates lies within a ten-year window and that you have data for at least 20 years on either side of that window. Then it might be easy to focus on event-time endpoints that are within 20 years and have a fully balanced set of units for each event-time coefficient. But even in this case, if you observed that the interesting dynamics in terms of treatment effects are resolved within the first five years of treatment, it might make sense to limit the event window to eight to ten years, to bring more visual attention to the period of interest and show the leveling off.

If your data setup does not allow for a straightforward resolution, then you need to make a judgment call. In this case, it will be useful to offer a "robustness check" specification, in which you choose an alternate approach (such as a wider event window).

Finally, it will be important for readers of an event study to know the degree of balance or imbalance in the number of units available to identify the event coefficients. This could be discussed in the text or presented as an appendix table showing the count of units, by event time j.

Special Attention for the Endpoints?

In event studies, it will be common to have data for some units that occur before or after the event window. In the notation of the traditional event study

equation, these would be observations for which $j \leq -m$ or $j \geq n$. We need to decide how to address this issue.

One natural option is to create and include as many event dummies as possible. By directly estimating a γ_j for each event time, this removes the problem. This approach is natural and appropriate when the data structure has both treated and untreated units that are balanced in calendar time (for example, all US states are observed over the period 1980–2020).

A second option involves creating "end-cap" variables in the traditional event study equation. For example, the data before and including the "pre" endpoint might be given a common dummy variable, $D_{i,t \leq E_i - m}$. Similarly the data points after the "post" endpoint can share a common dummy variable. I think this choice is the most common one, and often it is a good one. Schmidheiny and Siegloch (2023) recommend this approach (which they call "binning"). They note that it can provide a natural identifying restriction for timing-based data structures, that it creates a natural connection to distributed lag models, and that it can lead to a straightforward way to model multiple events per unit.

The main risks to creating "end caps" arise with trending counterfactuals or trending treatment effects. These risks are discussed further in the next main section of the paper on trends. That section argues that we might be hesitant about including "post" end caps if we think that there is a chance that treatment effects are changing over event time.

Another possible approach is just to drop observations that have event dates outside of our main window of interest. This option keeps the specification simple and creates a balance in event time in our analysis sample (for example, all US states are observed from three periods before their event to five periods after). One possible trade-off is that the loss of data can weaken statistical power. An additional consideration arises when using only ever-treated units: with this data structure you can be balanced in calendar time or balanced in event time, but not both. If you limit your sample to be balanced in event time, then this creates an imbalance in calendar time. This in turn means that the time dummies at the extremes will be estimated off of strangely selected units. Because the time dummies play a fundamental role in the identification of treatment effects, this approach seems risky to me.

A final option is not to include an event-time variable that is turned on for these faraway observations. In this way, the faraway observations are pooled together as part of the reference group for when the event did not happen. For example, the reference group in Figure 1 appears to be "more than five years before job displacement." This choice can be acceptable, but you should not include both "before the first endpoint" ($j \leq j_{min}$) as well as "after the final endpoint" ($j \geq j_{max}$) in the same reference group. Also, you should not combine "before the first endpoint" with the time period before the event in the same reference group. For example, in Figure 1 event time –1 is not part of the reference group.

A related choice when presenting a graph of the event-time coefficients concerns whether and how to plot endpoint coefficients. When the endpoint has its own dummy variable, it will capture different averaging than the "interior" terms and will

sometimes appear to be offset a bit from the rest of the graph. This can distract the reader from the main story about what is going on closer to event time zero. On the other hand, including such endpoints in the graph gives a fuller picture of the model; indeed, including them can sometimes help to diagnose problems with the specification or the data. I think best practice should typically be to plot the endpoint coefficients and to indicate in the figure (whether with a distinct symbol and/or in the figure notes) that these are differently estimated from the other event study coefficients.

Overall, you need to explicitly decide how you will deal with the endpoints and inform your readers about your decision.

Pooling Event Times for Statistical Power

With so many "key coefficients" to estimate, event study specifications can ask a lot of the data. Many event study models have pretty wide confidence intervals around each of the main γ_j coefficients. One strategy to regain some statistical power is to estimate models that pool together two or more adjacent event-time dummies, and then include these pooled variables in the model instead of the single-year event-time dummies. This approach strives for a balance between flexibility and statistical power. The main risk is that the pooling might obscure features of the empirical results. If you do this pooling, it is probably best to also show results from the unpooled model as a robustness check.

There are a variety of ways to pool event-time data. For example, one can restrict the model so that the coefficients will be the same in, say, periods 1 and 2, periods 3 and 4, and so on. Goodman-Bacon (2018) uses pooled event-time dummies to present results in table format. A more complex alternative is to restrict the event study coefficients to lie on a spline function between the points—essentially forcing a kind of averaging across points, but allowing for a flexible functional form (in a piecewise linear spline, the event study coefficients are forced to lie on a connected set of straight lines). For example, Bailey et al. (2020) and Lafortune, Rothstein, and Schanzenbach (2018) use spline restrictions for improved statistical power. However, this approach comes with some risk of mischaracterizing the pattern of treatment effects, in particular if the imposed model is not flexible enough to reflect reality. When using splines, it can make sense to allow for a jump or break in the splines in the transition from pretreatment to posttreatment periods. Lafortune, Rothstein, and Schanzenbach (2018) implement a model with a linear trend in event time, a jump at event time 0, and then a separate linear trend for event times after the event. As with pooling, it is best practice to also show the unconstrained model as a robustness check. Appendix E offers examples of pooling coefficients and spline models.

The Problem of Trends in Event Studies

Trends can cause problems for event studies in two distinct ways. First, treated unit types might follow a different trend than untreated types in terms of their

untreated potential (and unobserved) outcomes, which can confound the estimated treatment effects. As illustrated by Figure 1 at the start of the paper, if a trend is already apparent before the event, it calls into question how to interpret patterns after the event. Second, treatment effects themselves may be trending in time-since-treatment. This second possibility is not necessarily a problem for event study models: after all, the point of these models is to allow for treatment effects that vary over time. But trending treatment effects can cause problems for the estimates from certain specification choices. In this section, I lay out these issues and some possible approaches in more detail. (Appendix F has an expanded discussion and graphical illustrations for several of the main points.)

Pre-event Coefficients as a Diagnostic Tool

The estimated pre-event terms can serve as a tool for diagnosing trends. This is often done informally by inspecting the graph of the pre-event coefficients. This tool is most appropriate when working with difference-in-difference or hybrid data structures, which include never-treated units.

An additional consideration arises if we are working with a timing-based data structure with no control units. In this setting, Borusyak, Jaravel, and Spiess (2022) show that due to the multicollinearity of event-time, calendar-time, and unit fixed effects it is impossible to identify a linear trend in the set of treatment effects (or in the pretreatment coefficients). In this case, the best we can do is to look for nonlinear pre-trends. For this data structure, they recommend the normalization of setting an additional pretreatment γ_{-a} coefficient to be zero. This step imposes a zero pre-trend, and allows for visual or statistical inspection of the other pretreatment coefficients as a check for nonlinear pre-trends. Schmidheiny and Siegloch (2023) argue that using end caps can provide identification of the event-time coefficients in a timing-based data structure. This would restore the ability to examine pre-event trends.

A separate difficulty is that if you have too few pretreatment periods, it can be hard to distinguish between actual pre-trends and statistical noise. This limits the comfort a researcher can take from "passing" a test of no visible pre-trend. There is no hard and fast rule for "how few is too few." When looking at a graph of event study coefficient estimates, I find it useful to mentally visualize the range of possible pre-trends that could be consistent with the pretreatment estimates. Across papers that I see, this approach often leaves me feeling skeptical if I see three or fewer pre-event terms. But this depends on both the variability and the apparent trends among those pretreatment coefficients. If you are concerned about this issue, a simple additional step here is to add more pretreatment periods, extending further back in event time. In a more structured approach, Dobkin et al. (2018) plot the linear pre-trend from a parametric model on the figures that show event study coefficients.

Recent econometric work identifies some potential problems with using pre-trends as a diagnostic tool. Roth (2022) notes that the widespread, informal practice of "rounding insignificant pre-trends to zero" can lead to "pre-test bias."

Even a mild pre-trend, which cannot be visually or statistically detected, can still meaningfully influence the estimated posttreatment impacts. Roth argues that if we are confident of the functional form of the trends (for example, that the trends are linear in time) we should plan always to control for trends regardless of whether or not there is not a strongly apparent pre-trend. He also presents more sophisticated extensions to methods of controlling for trends, based on Freyaldenhoven, Hansen, and Shapiro (2019) and Rambachan and Roth (2023), that allow a researcher to proceed under weaker assumptions about the functional form of the trends.

Separately, pre-trends can be biased if our underlying model is misspecified. Sun and Abraham (2021) examine the case where the misspecification arises from different unit types having different treatment effects—say, if those treated earlier in calendar time have larger treatment effects than those treated later in time. For example, suppose that in Jacobson, LaLonde, and Sullivan (1993) the individuals facing job displacement early in the sample (1980–1982) have greater impacts than those displaced later in the sample (1983–1986), perhaps due to changes in the macroeconomic environment. This difference in treatment effects can lead to a (spurious) apparent trend in the estimated pre-event coefficients. In this case, the appearance of a pre-trend is an indication that something is wrong with the specification of our model.[10] De Chaisemartin and D'Haultfœuille (2022) propose alternative pre-trend estimators that are robust to different unit types having different treatment effects.

Controlling for Unit-Specific Trends

Rather than focusing on pre-trends as a diagnostic measure, an alternative is to control for unit-specific trends, by including a (continuous) time variable interacted with unit (for example, state) dummies. This approach is suitable if we believe that pre-trends reflect trending omitted variables that could bias the main estimates of the treatment. Controlling for unit-specific trends aims to eliminate this omitted variables bias.

For example, Alsan and Goldin (2019) use an event study to examine the historical introduction of clean water and sewer projects across municipalities during 1880–1920. In their specification they control for municipality-specific time trends. In a separate example, Bostwick, Fischer, and Lang (2022) study the impacts of a university switching from a quarter-based to a semester-based schedule. They want to make sure their estimates are not confounded by outcomes trending differently across universities, so they control for university-specific time trends.

What are the main trade-offs between using pre-trends as a diagnostic and controlling for unit-specific trends? First, controlling for unit trends changes the counterfactual thought experiment. Our treatment effect estimates now have an

[10] The specification error here is the assumption of treatment effects that are the same for both early and late treated units. This contrasts with our usual interpretation of pre-trends as indicating anticipation effects or different underlying trends in untreated potential outcomes.

interpretation of "my outcome compared to the reference period, and net of underlying linear trends in the counterfactual between that period and now." Second, because the counterfactual now controls for these trends, our pre-trends should look flat by construction. Thus, we lose the basic falsification test that the pre-trends provide in the basic model. However, one can still use the pre-event coefficients to look for nonlinear violations of the parallel trends assumption.

Adding unit trend controls may also interact uncomfortably with the "too many variables" and "multicollinearities abound" challenges of event studies. A unit-specific time trend term will be multicollinear with the unit dummies, time dummies, and event-time variables. This means that it will be necessary to add (at least) one more additional parameter restriction. As noted earlier, it is imprudent to just let our software address the collinearity problem by dropping a variable on its own, because what it drops might undermine our interpretation of the resulting estimates. The choice of which restriction to apply is guided by the same principle as before: the additional restrictions are untestable, but estimated coefficients on the trend terms will be affected by the restrictions we place, so we want it to be as "obviously true" as possible. Our additional restriction should be applied to the event study coefficients γ_j so that we do not undermine the panel fixed effects controls in the estimation. A common choice is to impose a restriction that two event coefficients are equal. The trend estimates will be estimated in the context of those restrictions; in particular, the later treatment estimates may "pivot" as a result of imposing this restriction. (This is illustrated in online Appendix F.)

Trends versus Pre-trends

The estimated parameters for unit-specific trends will seek to capture trending behavior both before and after the event. But what if treatment effects are also trending? Suppose that in our traditional event study equation, treatment effects are increasing in event time: $\gamma_0 < \gamma_1 < \gamma_2 \ldots$. Then, an estimate of unit-specific trends might try to fit both pre-event trends *and* the treatment effect pattern. This in turn can bias the estimated event study coefficients. This problem can occur if our parameter restrictions include post-event terms, such as a post-event end cap. The post-event end cap (for example, "six or more periods after the event") forces all the estimated event-time effects γ_j within the end cap to be the same. If instead they are truly trending, this can cause problems. An extreme version of this is a difference-in-difference specification, which restricts all post-event treatment effects to be the same.

This is the main argument in the Wolfers (2006) critique of prior difference-in-difference approaches examining the impact of unilateral divorce laws on divorce rates. Much of this work used state trend controls. Wolfers argues that these laws will have dynamic impacts—that is, trending γ_j for $j \geq 0$. Because of this, the estimated state-specific time trends will be contaminated by trying to also fit the trending treatment impacts. This in turn will bias the main estimates. Wolfers proposes as an improvement a variation of an event study specification for the post-event periods. (I present a stylized illustration of this phenomenon in online Appendix G.2.)

One option to prevent this problem is to focus on controlling for "pre-trends" only. However, this may require custom programming (online Appendix F.4 presents one approach). Another option is to model your event study terms (the $\gamma_j \cdot D_{i,t-j}$ terms in the traditional event study equation) so that the unit-specific time trends will not be confounded by trending treatment effects. For example, one can drop all constraints on the event study parameters for posttreatment by including all "post" event-time dummies and having no "post" end cap. This step will ensure that the trend coefficients are estimated based only on pre-event data.

Recommendations in Controlling for Trends

These considerations lead to guidelines for researchers who are controlling for trends. First, don't let post-event parameter restrictions influence your estimated trends, unless you are highly confident that the treatment effects are not trending in that range. Otherwise, your control for trends may be picking up part of the trending treatment event.

Second, if one of your extra parameter restrictions is in the form of equality of two event coefficients, consider spacing those coefficients further apart, because the impact of any statistical "noise" between the two coefficients will be larger if they are closer together. As an alternative, focus the restrictions on event study parameters that allow for more averaging across units.

One restriction that accommodates the considerations above—at least for the difference-in-differences and the hybrid data structures—is to constrain the trend in the "reference period" event study coefficients to be zero. This approach has the advantage of averaging across several coefficients, and reducing the impact of noise from any one or two of them. It also respects the notion of having a reference period embodied in the normalizing restriction (as discussed earlier), and offers a natural counterfactual interpretation: "Compared to the level and trend in the reference period, and the over-calendar-time changes from control units, what would my expected outcome be?"[11]

Finally, if we are working with a timing-based data structure, controlling for trends has potential to create surprising and severe problems. Adding *linear* trend controls (and the required additional parameter restriction) can induce *quadratic* trends into our estimated event study coefficients. This arises from a subtle way in which the event dummies are collinear with the other variables in the model (as illustrated in online Appendix F.2). The key lesson is to be extra cautious about the combination of trend controls and a timing-based data structure.

There is one way in which adjustments for trends can often be simplified compared to common practice: we can focus on unit-type trends, rather than

[11] Given a reference period (k_1, k_2), this is implemented with the following linear restriction on the event study parameters: $\sum_{j=k_1}^{k_2} \left(j - \frac{k_1 + k_2}{2}\right) \cdot \gamma_j = 0$. To derive this, consider a bivariate regression $\gamma_j = \phi_0 + \phi_1 \cdot j$. To impose a zero trend, we want $\hat{\phi}_1 = 0$. The left hand side of the proposed restriction is the numerator for the coefficient $\hat{\phi}_1$.

unit-specific trends. That is, we can allow for one shared trend parameter for each group of units that share an event date. This is because the event study variables depend only on unit type and time. Once we condition on unit-type trends, any remaining unit-specific trends will be orthogonal to the event study dummies and will not influence their coefficients.

Statistical Inference for Event Study Models

For researchers, the usual primary concern is to have unbiased point estimates. However, we also need to be able to conduct statistical inference to test hypotheses about the true state of the world.

Cluster-Robust Inference

For event study models, the current practice appears to be to calculate cluster-robust standard errors, with clusters defined as the i-level units. This starting place is sensible. The key right-hand-side variables in an event study have some degree of autocorrelation and it is plausible to think about the model error term also being positively autocorrelated over time within a unit. Taken together, this argues for clustering at the unit level. The general rule of thumb is that we want to cluster at a level when there is correlation in the scores ($X_i e_i$, driven by correlation in the model errors e) across units in that cluster (Cameron and Miller 2015, section II.C). If the underlying event is shared across units, then this argues for clustering at a higher level. For example, if our dataset is a panel of individuals, but the event is a state-level policy change, then we likely want to cluster at the level of the state.

However, one potential concern is that standard cluster-robust methods provide accurate standard errors only if the number of clusters is "large enough," with no hard and fast rule for what that means. Folk wisdom and some simulations offer rules-of-thumb like 42 or 50 clusters, but in some settings this is not enough, and in other settings a smaller number will suffice. When there are too few clusters, traditional cluster-robust methods may over-reject. If we are facing too few clusters, we need to take account of this in our inference procedures (Cameron and Miller 2015, section VI).

The problem of few clusters is exacerbated when the clusters are of asymmetric size or when there are very few treated units. In these settings, our conclusions need to be more tentative. But it is not so bad that you just have to give up. In these settings, the adjustments offered in Imbens and Kolesar (2016), Carter, Schnepel, and Steigerwald (2017), and MacKinnon and Webb (2017) might be a good choice.[12]

[12] Permutation tests are an alternative approach to conducting inference. The idea is to randomly reassign pseudo-event dates across units, and re-estimate the model. Repeat this procedure many times, to construct a distribution of "estimated treatment effects, when there is no actual treatment." A distribution of test statistics can be constructed from these permutations. The main estimates can be compared

The Spatial Correlation Problem

The basic premise of cluster-robust inference requires that clusters are independent from one another. Spatial correlation in event dates undercuts this premise, and doing so may result in over-rejection of the null hypothesis. When events are the result of a political process or influenced by economic circumstances, neighboring units (say, neighboring states) can be closer in event date than more distant units. Often, economic outcomes are also spatially correlated.

For the most part, the current empirical literature ignores this concern, but there are some potential ways to address it; for example, see Conley (1999) on spatial robust standard errors. However, there is little guidance on how to on measure "distance" across some combination of space and time.

Another possibility is to allow for arbitrary correlations in observations within a cluster, and also allow for correlation that decays in calendar time across observations in nearby time periods, regardless of the unit to which they belong (as in Driscoll and Kraay 1998). This approach allows for greater dependence across observations than the current standard. But there is no "button to push" for implementation of these approaches, so it will require custom programming. In addition, allowing for spatial autocorrelation is likely to make the "few clusters" problem even more salient.

For the near term, a basic precaution is to examine your data for the possibility of spatial correlation in event dates, although currently there is no hard-and-fast guidance for what levels of spatial autocorrelation should be a matter of concern. For now, the standard practice of "cluster on the underlying event" seems likely to continue. However, researchers should probably start to pay more attention to spatial correlation in the future.

Extensions and Challenges

My discussion has focused on a basic version of event study models. In this section I briefly note a few of the additional extensions and challenges that may arise.

Events with Variable Intensity

What should researchers do when events can vary in their magnitude? For example, suppose the event is a cigarette tax hike or an increase in the state minimum wage. We might want to allow for the event to scale proportionally to the size of the shock. This issue can be handled in a straightforward way by pre-multiplying the

against these distributions, and if they fall in the tails of the distribution, this is evidence against the null hypothesis of no impact. MacKinnon and Webb (2019) study this randomization approach in a difference-in-difference setting with few units, very few treated units, and clusters having different sizes (for example, larger and smaller US states). Young (2019) also shows that randomization procedures (based on t-statistics) perform well. I think it likely that such results would carry over to the event study setting.

event dummy by the magnitude of the event. For example, the event variable $D_{i,s}$ could be "by what percentage were cigarette taxes hiked?"

One interesting variation is found in Goodman-Bacon (2018), who examines the introduction of Medicaid in the 1960s across US states. The impact of introduction varied state-by-state as a function of the fraction of population that was receiving assistance from the Aid to Families with Dependent Children welfare program at the time when Medicaid was introduced. This setup combines both timing-based and variable-intensity variation in treatment. Concerned about event dates being correlated with preexisting trends, Goodman-Bacon (2018) isolates the variation from variable-intensity of treatment from the timing by including dummies to control for event date by calendar year.

A group of recent papers center event study models within "mover" strategies. These include consumers changing purchase patterns as they move across locations (Bronnenberg, Dubé, and Gentzkow 2012) and either doctors (Molitor 2018) or patients (Finkelstein, Gentzkow, and Williams 2016) moving from one region to another with different patterns of health-care practice. These mover designs often pair with variable intensity of treatment. For example, Finkelstein, Gentzkow, and Williams (2016) track Medicare patients who move across regions with different intensities of medical usage. In this case the event is the move, and the variable intensity reflects the difference in medical usage between the destination and origin locations. Molitor (2018) examines cardiologists' patterns of practice as they move across regions. Again, the variable intensity of the event is given by the difference in regional patterns of practice across destination and origin locations.

More than One Event Per Unit

What if there is a possibility of multiple events per unit? For example, the data might include repeated layoffs or repeated state minimum wage hikes.

For the basic case, this can be straightforward to implement. We define the event $D_{i,s}$ to be one in any period where an event occurs, and we allow this to happen in different time periods for the same unit i. Thus, more than one of the event-time dummies can be turned on simultaneously. Sandler and Sandler (2014) suggest this approach.[13] However, this approach provides a specific interpretation of the estimated coefficients—a "partial effects" interpretation, holding constant the potential impact of subsequent events (including those whose existence might in turn be impacted by the current event). This can be different from the "total effect," which includes the impact of today's event on the likelihood of future events happening. Krolikowski (2018) explores this issue by using a simulation to

[13] A related but alternative approach is to duplicate data around the event. In this approach, each observation is "split" into multiple new observations based on unique combination event-by-underlying-unit. However, Sandler and Sandler (2014) show using Monte Carlo simulations that in some settings this can lead to biased estimates.

propose a weighted average of the partial effect estimates as well as a "first event only" model.[14]

Another approach is to adjust the definition of an event so as to have only one per unit. For example, in the Jacobson, LaLonde, and Sullivan (1993) paper behind Figure 1, the focus is on the first layoff, and subsequent layoffs are not modeled. This approach could be implemented based on "biggest event" or "first big event." Again, issues will arise in interpreting the resulting coefficient. By bundling subsequent events (and their dynamic impacts) into the definition of "treatment," we have a potentially nonintuitive definition of treatment—a version of the partial-versus-total effects problem just mentioned. In some cases, this approach is combined with using the "never treated" group as a control group. This combination can introduce the selection issue mentioned earlier—that is, the control group may now differ in unobserved ways, like stronger skills or labor market attachment, so comparisons with them will give biased counterfactuals for the treated. This can make the estimated effects of the displacement look worse than the true causal effects. Separately, it can raise concerns about external validity of the findings to the broader population.

The possibility of multiple events raises the question of whether the effect of an event depends on the history of other events. A first layoff is one thing, but we can imagine that subsequent layoffs are possibly worse (increasing fragility) or not quite as bad (either toughening up or "floor effects") as the first. In principle, the basic model could be modified to estimate sensitivity in treatment effects directly, based on the history of prior events, but I have not yet seen this implemented.

Heterogeneous Treatment Effects

What if the effect of treatment does not just vary in "time since event," but also depends systematically on the unit type, time, or context? For example, the treatment effect might depend on observable variables or on the date of adoption of the event.

Suppose we are interested in how a treatment effect varies across men and women. We can then include one set of event dummies for men and another set of event dummies for women. More generally, we can include a set of interaction terms, based on the covariates that we believe influence the treatment effects. In taking this step, it is important to follow usual best practice for interaction terms in regression models, such as including direct controls for the covariates if they are time-varying. These interactions can use up a lot of variation in the data, and in response, it may be useful to impose a parametric simplification on the interaction terms. For example, the Jacobson, LaLonde, and Sullivan (1993) example

[14] Basso, Miller, and Schaller (2022) label the partial effect "Y channel only (YCO)" and contrast the Event Study approach with a Local Projections approach to estimating dynamic treatment effects. They observe that Local Projections can directly recover the "total effect" (corresponding to the impulse response) and show that it can be transformed into the YCO. Cellini, Ferreira, and Rothstein (2010) also address this distinction in the context of a dynamic regression discontinuity model and estimate both effects. They label the partial effect "Treatment on the Treated" and the total effect "Intent to Treat."

from Figure 1 used a parsimonious approach of having three periods of treatment effect: the "dip" (the 13 quarters prior to job displacement), the "drop" (the quarter of displacement), and the "recovery" (six quarters following displacement). They allow covariates to produce different slopes (in event time) for these three periods.

In a setting with varying event dates, we might want to model the possibility that the dynamic treatment effects depend on the timing of adoption. For example, US states that are early to adopt a policy might be the ones that benefit the most from that policy; late adopters might have less or even opposite-signed effects. One approach is to treat the actual event date as an observable variable and estimate treatment effects based on the date, or perhaps using an "early"/"late" adopter dummy variable. This approach should work, so long as there are control units or enough variation in event dates.

There is a recent, active, and promising literature on how event study models perform when treatment effects differ across units and when we do not know the functional form of how they differ. This raises issues analogous to those of local average treatment effects (LATE) in the instrumental variables context, in which our main estimates are a weighted average of the underlying treatment effects. Typically these weights might not correspond to our common-sense intuitions or to our desired weighting.[15] This literature typically considers the case where there is (at most) a single event per unit. Sun and Abraham (2021) point out that the overall effect will be a weighted average of the heterogenous effects for different unit types. They show that an auxiliary regression can calculate the implied weights and also propose an alternative estimation method that works to recover a target average treatment effect. Using a different strategy, de Chaisemartin and D'Haultfœuille (2022) propose relying on using not-yet-treated units and the parallel trends assumption to recover estimates of the treatment effects for each treated unit type, which can then be averaged together.

When "Time" Is Not Calendar Time

What if the time variable is not calendar time? This situation can arise in cohort studies: for example Duflo (2001) studies the life-course impact of childhood exposure to school availability in Indonesia based on district and year of birth, and Bailey, Sun, and Timpe (2021) examine the long run impacts of childhood exposure to Head Start in the United States, with treatment based on county and year of birth. These are standard event study analyses, only the time variable is "year of birth" instead of calendar time.

The challenge here is how best to deal with cohort, age, and time (of survey) effects. The basic event study specification requires fixed effects for cohort. Often the outcomes of interest—such as labor market or demographic outcomes—depend

[15] This theme is addressed for ordinary least squares in Angrist (1998) and Sloczynski (2022); for one way fixed effects models in Gibbons, Suárez Serrato, and Urbancic (2019) and Miller, Shenhav, and Grosz (2021); and for difference-in-difference in Goodman-Bacon (2021a), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020), and Borusyak, Jaravel, and Spiess (2022).

importantly on age in nonlinear ways. There can also be important calendar-time effects (for example, if some data is collected in a recession). This raises the challenge of age-cohort-time multicollinearity (as discussed in Deaton 2018, pp. 123–127).

There is no avoiding the fact that the analysis becomes complicated here. If theory suggests all three factors—age, cohort, and time—may be important, I recommend including all three sets of dummies.[16] One leading alternative is instead to include a set of two-way interaction dummies: either *age–by–cohort* fixed effects, *cohort–by–time* fixed effects, or *age–by–time* fixed effects. Any one set of these two-way-interactive fixed effects controls for more, but also "uses up" more variation in the data, which raises its own issues. But it seems like good practice to at least include a specification with these two-way-interactive fixed effects as a robustness check.

A final issue is that in producing an overall estimate, we might want to make sure that each cohort is weighted proportional to its population, because the thought experiment of the model centers on the cohorts. However, our data might not naturally reflect those weights, perhaps especially when we are combining data from different-sized datasets.[17] This can set up a choice between improved statistical power (weighting based on the data in our sample) and improved representativeness (weighting based on size of cohorts), and I do not think there is currently a settled "best practice" for these issues. But we should think carefully about how to weight our observations, and not simply take the weights as they are given by the datasets we are using.

Conclusion

Event study models are great! But behind that attractive interpretive graph, researchers are necessarily making decisions. This raises risks of bias due to systematic (if perhaps unconscious) model selection processes, committed by either the researcher or the journal review process. Despite these risks, these decisions are unavoidable. There is no "button to push" that can automate the necessary judgment calls. For now, best practice should be to increase transparency through bringing clarity about the specification decisions made (and the reasons for those decisions) and to discuss robustness to alternative decisions, along with providing both estimation code and (whenever possible) data for replication.

[16] Sometimes researchers include only two sets of these three possible sets of dummy variables, such as dummies for age and for cohort. Presumably this is motivated by the collinear relationship: age = cohort + time. However, among the many coefficients for the three sets of dummy variables, there is only one degree of collinearity. So omitting a full block of dummies—say, leaving out the time dummies—imposes many more restrictions than are necessary.

[17] For example, the outcome data in Bailey, Sun, and Timpe (2021) pool observations from the long-form 2000 Census and the 2001–2018 ACS. The relatively large number of observations in the census means that birth cohorts from 1950 to 1965 are weighted more heavily than later cohorts, relative to their population size.

■ *I thank Gaetano Basso, Colin Cameron, Hilary Hoynes, Giulia Olivero, Marianne Page, Zhuan Pei, Jessamyn Schaller, my students in my PhD Applied Econometrics classes, and the editors, for many helpful comments and conversations. Special thanks to Liz Cascio, who first explained event study models to me in such a way that I felt like I started to "get it."*

References

Alsan, Marcella, and Claudia Goldin. 2019. "Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880–1920." *Journal of Political Economy* 127 (2): 586–638.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." The Review of Economics and Statistics 60 (1): 47–57.

Bailey, Martha J., Hilary W. Hoynes, Maya Rossin-Slater, and Reed Walker. 2020. "Is the Social Safety Net a Long-Term Investment? Large-Scale Evidence from the Food Stamps Program." NBER Working Paper 26942.

Bailey, Martha J., Shuqiao Sun, and Brenden Timpe. 2021. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." *American Economic Review* 111 (12): 3963–4001.

Basso, Gaetano, Douglas L. Miller, and Jessamyn Schaller. 2022. "Dynamic Treatment Effects for Empirical Microeconomists: Local Projections and Quasi-experimental Research Designs." Unpublished.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. "Revisiting Event Study Designs: Robust and Efficient Estimation." Unpublished.

Bostwick, Valerie, Stefanie Fischer, and Matthew Lang. 2022. "Semesters or Quarters? The Effect of the Academic Calendar on Postsecondary Student Outcomes." *American Economic Journal: Economic Policy* 14 (1): 40–80.

Braun, Matias, and Claudio Raddatz. 2008. "The Politics of Financial Development: Evidence from Trade Liberalization." *Journal of Finance* 63 (3): 1469–1508.

Bronnenberg, Bart J., Jean-Pierre H. Dubé, and Matthew Gentzkow. 2012. "The Evolution of Brand Preferences: Evidence from Consumer Migration." *American Economic Review* 102 (6): 2472–2508.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–72.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of West German Wage Inequality." *Quarterly Journal of Economics* 128 (3): 967–1015.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity." *Review of Economics and Statistics* 99 (4): 698–709.

Cellini, Stephanie R., Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *Quarterly Journal of Economics* 125 (1): 215–61.

Chetty, Raj, John N. Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. 2014. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark." *Quarterly Journal of Economics* 129 (3): 1141–1219.

Clarke, Damian, and Kathya Tapia-Schythe. 2021. "Implementing the Panel Event Study." *Stata Journal* ►21 (4): 853–84.

Conley, T. G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92 (1):

1–45.

Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods." *AEA Papers and Proceedings* 110: 42–48.

Deaton, Angus. 2018. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Reissue Edition with a New Preface*. Washington, DC: World Bank.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." NBER Working Paper 29873.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo. 2018. "The Economic Consequences of Hospital Admissions." *American Economic Review* 108 (2): 308– 52.

Dolley, J. C. 1933a. "Characteristics and Procedure of Common-Stock Split-Ups." *Harvard Business Review* 11: 316–26.

Dolley, J. C. 1933b. "Common Stock Split-Ups: Motives and Effects." *Harvard Business Review* 12: 70–81.

Driscoll, John C., and Aart C. Kraay. 1998. "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *Review of Economics and Statistics* 80 (4): 549–60.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Finkelstein, Amy, Matthew Gentzkow, Dean Li, and Heidi L. Williams. 2022. "What Drives Risky Prescription Opioid Use? Evidence from Migration. NBER Working Paper 30471.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence from Patient Migration." *Quarterly Journal of Economics* 131 (4): 1681–1726.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. "Pre-event Trends in the Panel Event-Study Design." *American Economic Review* 109 (9): 3307–38.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2019. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8 (1): 20170002.

Goodman-Bacon, Andrew. 2018. "Public Insurance and Mortality: Evidence from Medicaid Implementation." *Journal of Political Economy* 126 (1): 216–62.

Goodman-Bacon, Andrew. 2021a. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Goodman-Bacon, Andrew. 2021b. "The Long-Run Effects of Childhood Insurance Coverage: Medicaid Implementation, Adult Health, and Labor Market Outcomes." *American Economic Review* 111 (8): 2550–93.

Goodman-Bacon, Andrew, and Jamein P. Cunningham. 2019. "Changes in Family Structure and Welfare Participation since the 1960s: The Role of Legal Services." NBER Working Paper 26238.

Imbens, Guido W., and Michal Kolesar. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98 (4): 701–12.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *American Economic Review* 83 (4): 685–709.

Krolikowski, Pawel. 2018. "Choosing a Control Group for Displaced Workers." *ILR Review* 71 (5): 1232–54.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Economic Journal: Applied Economics* 10 (2): 1–26.

MacKinlay, A. Craig. 1997. "Event Studies in Economics and Finance." *Journal of Economic Literature* 35 (1): 13–39.

MacKinnon, James G., and Matthew D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2): 233–54.

MacKinnon, James G., and Matthew D. Webb. 2019. "Wild Bootstrap Randomization Inference for Few Treated Clusters." In *The Econometrics of Complex Survey Data: Theory and Applications*, Vol. 39, *Advances in Econometrics*, edited by Kim P. Huynh, David T. Jacho-Chávez, 61–85. Bingley, UK: Emerald Publishing.

Miller, Douglas L. 2023. "Replication data for: An Introductory Guide to Event Study Models." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185943V1.

Miller, Douglas L., Na'ama Shenhav, and Michel Grosz. 2021. "Selection into Identification, with Application to Head Start." *Journal of Human Resources*. doi: 10.3368/jhr.58.5.0520-10930R1.

Molitor, David. 2018. "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration." *American Economic Journal: Economic Policy* 10 (1): 326–56.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies*. https://doi.org/10.1093/restud/rdad018.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4 (3): 305–22.

Sandler, Danielle H., and Ryan Sandler. 2014. "Multiple Event Studies in Public Finance and Labor Economics: A Simulation Study with Applications." *Journal of Economic and Social Measurement* 39 (1–2): 31–57.

Schmidheiny, Kurt, and Sebastian Siegloch. 2023. "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization." *Journal of Applied Econometrics*. doi: 10.1002/jae.2971.

Sloczynski, Tymon. 2022. "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights." *Review of Economics and Statistics* 104 (3): 501–09.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.

Wolfers, Justin. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 96 (5): 1802–20.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics* 134 (2): 557–98.

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.

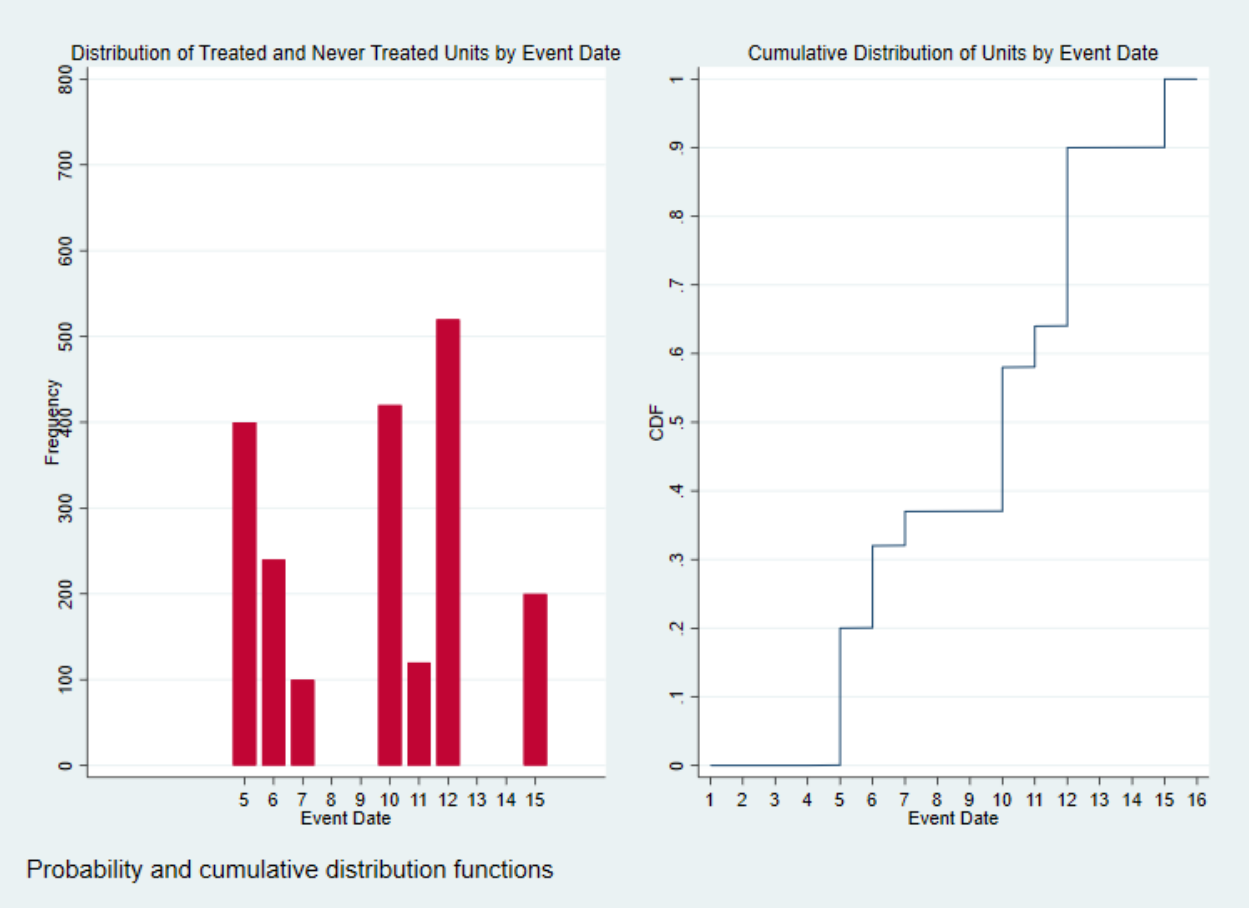
The list above is incomplete, and there are many variations. One common situation is when the variation in event dates E_i is neither cleanly "all at once" ($E_i = E, \forall i$), but there are important groupings of E_i across units. For example, a policy might be adopted by a handful of states at different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

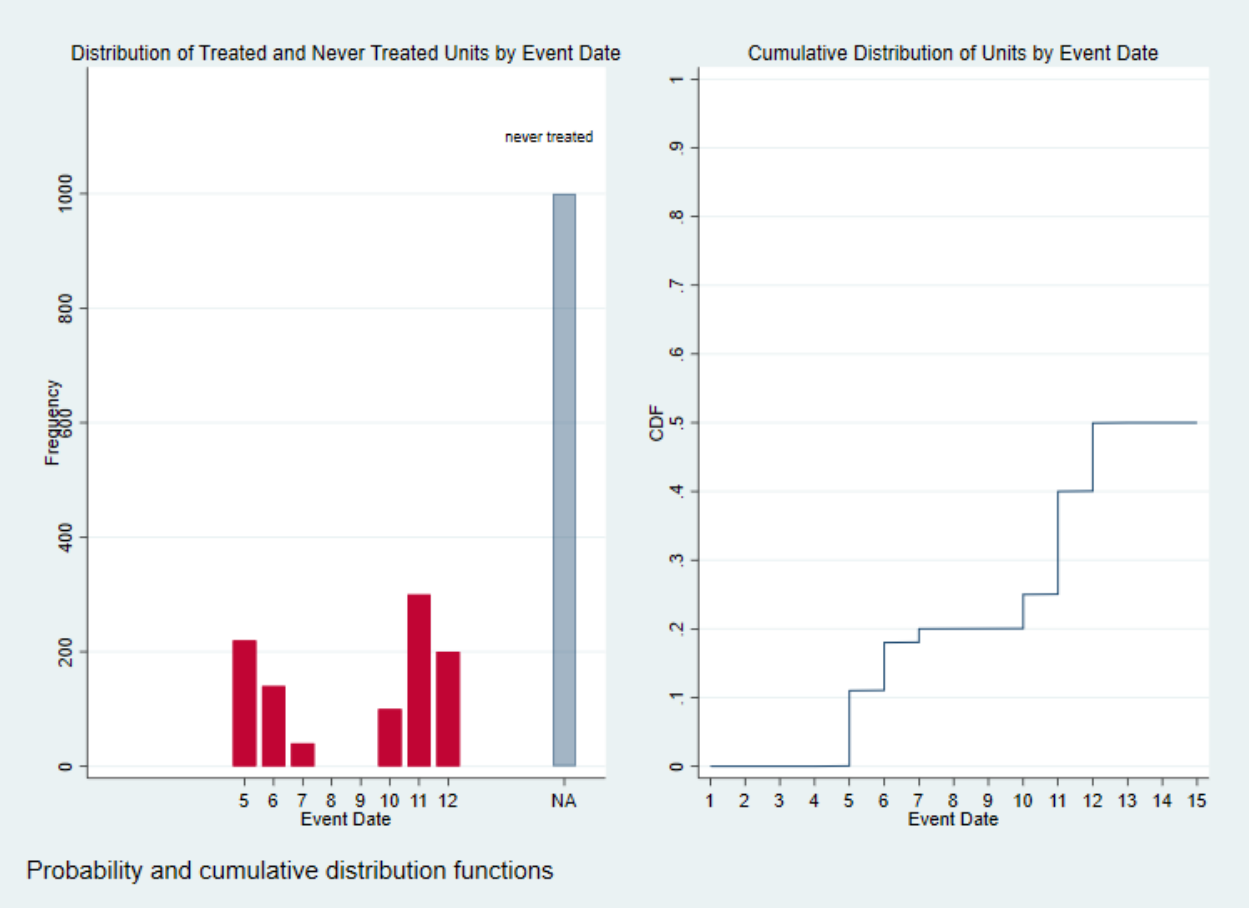
Figure A.1 illustrates this for a timing-based data structure. The earliest treated units have their event date in period 5, and the latest event date is period 15, by which point all units have been treated. The graph on the right shows the same information, in the form of a CDF across units of event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

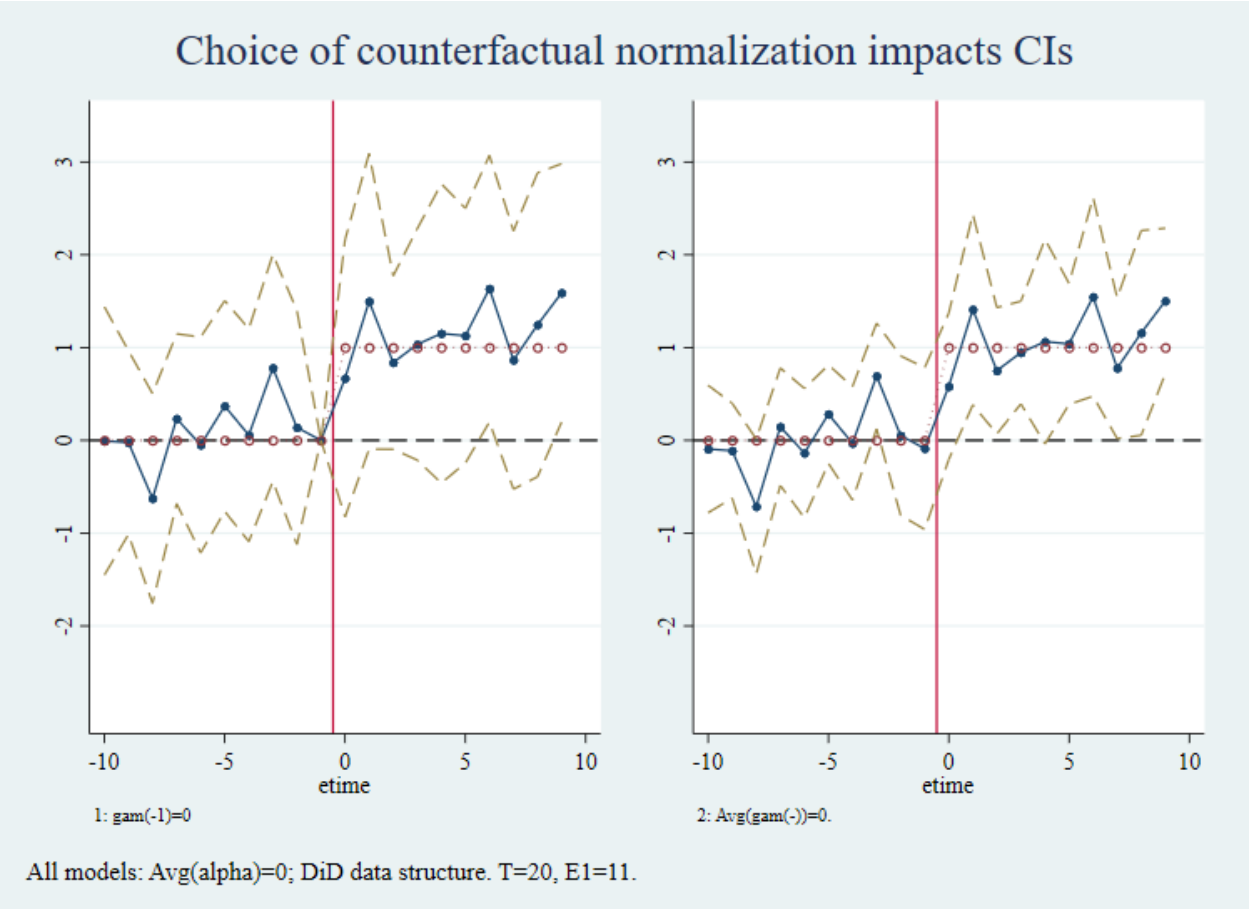
This section illustrates some issues from Section 3.1.

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

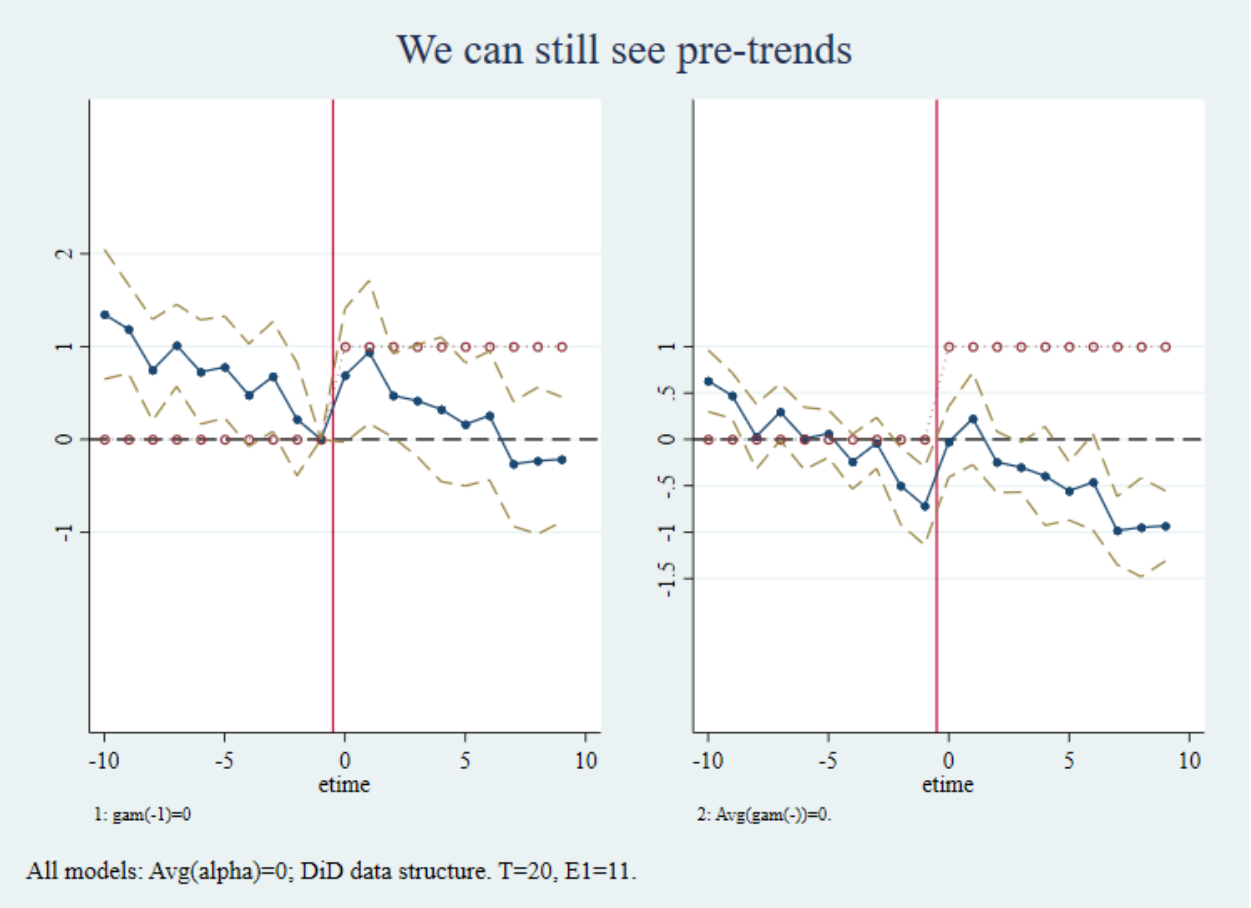
[19]The data in this example were selected so as to have results of statistical significance differ across the two graphs, as a rhetorical trick to emphasize the main point. The general lesson is that using the full reference period will (1) show increased precision, and (2) corresponds to our intuitive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

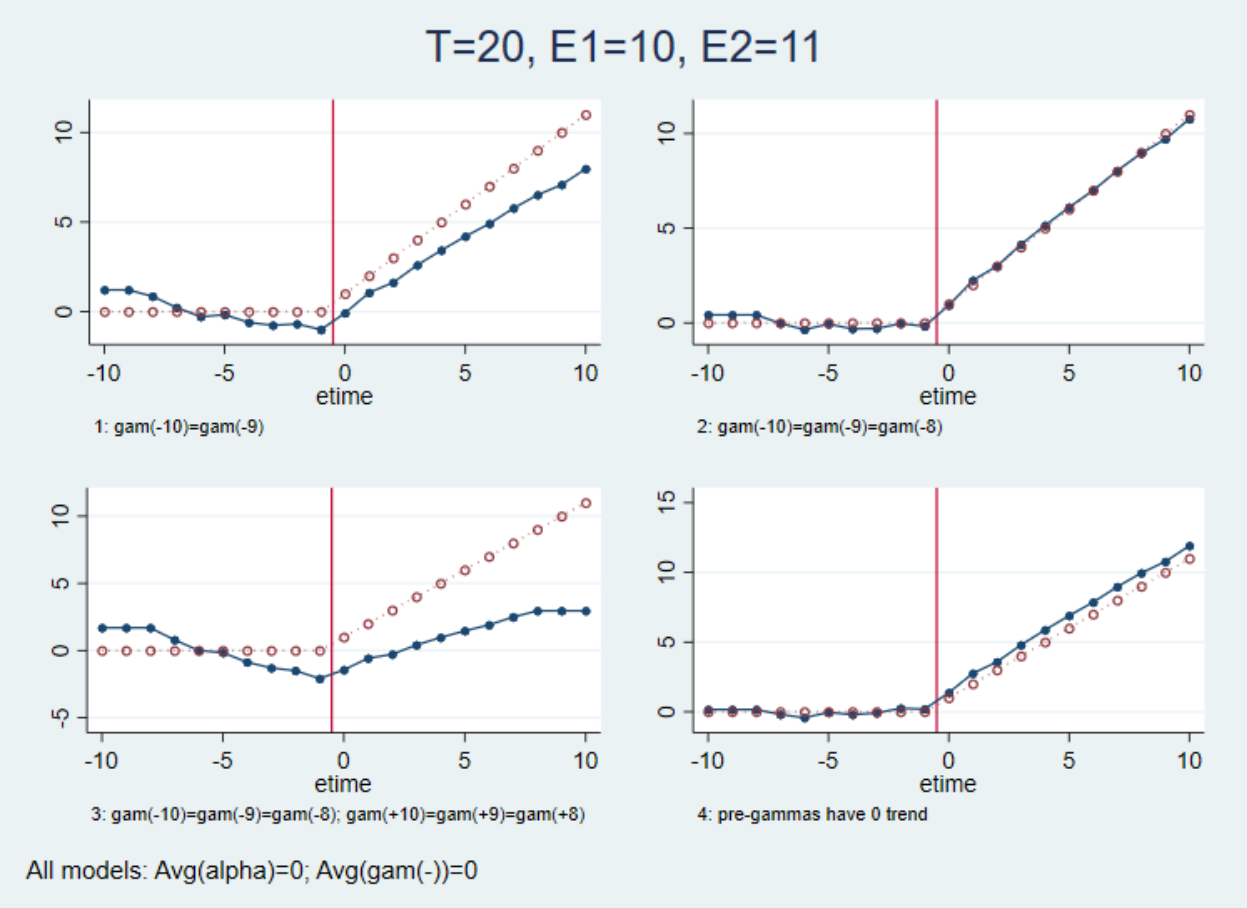


Note: The y-axis show the estimated treatment effects and 95% confidence intervals. The x-axis shows event time. The left panel normalizes event time -1 to zero; while the right panel normalizes the average of -10 to -1 to be zero.

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

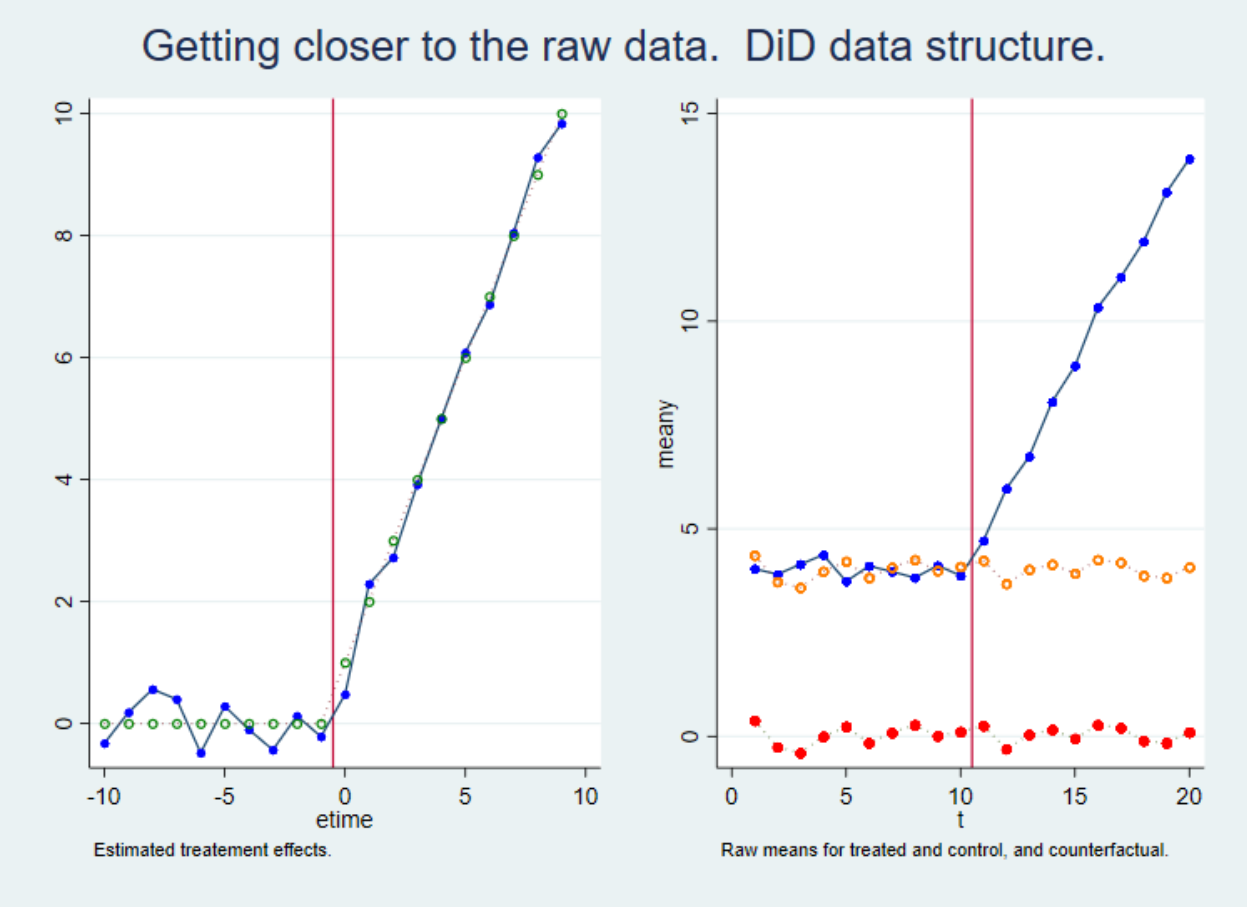
In figure A.7, we see an illustration of showing the counterfactual alongside the raw data. The data structure in the figure below is a Difference-in-difference data structure, with two unit types: (1) treated units sharing a common event date, and (2) control units. The first graph shows the estimated 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

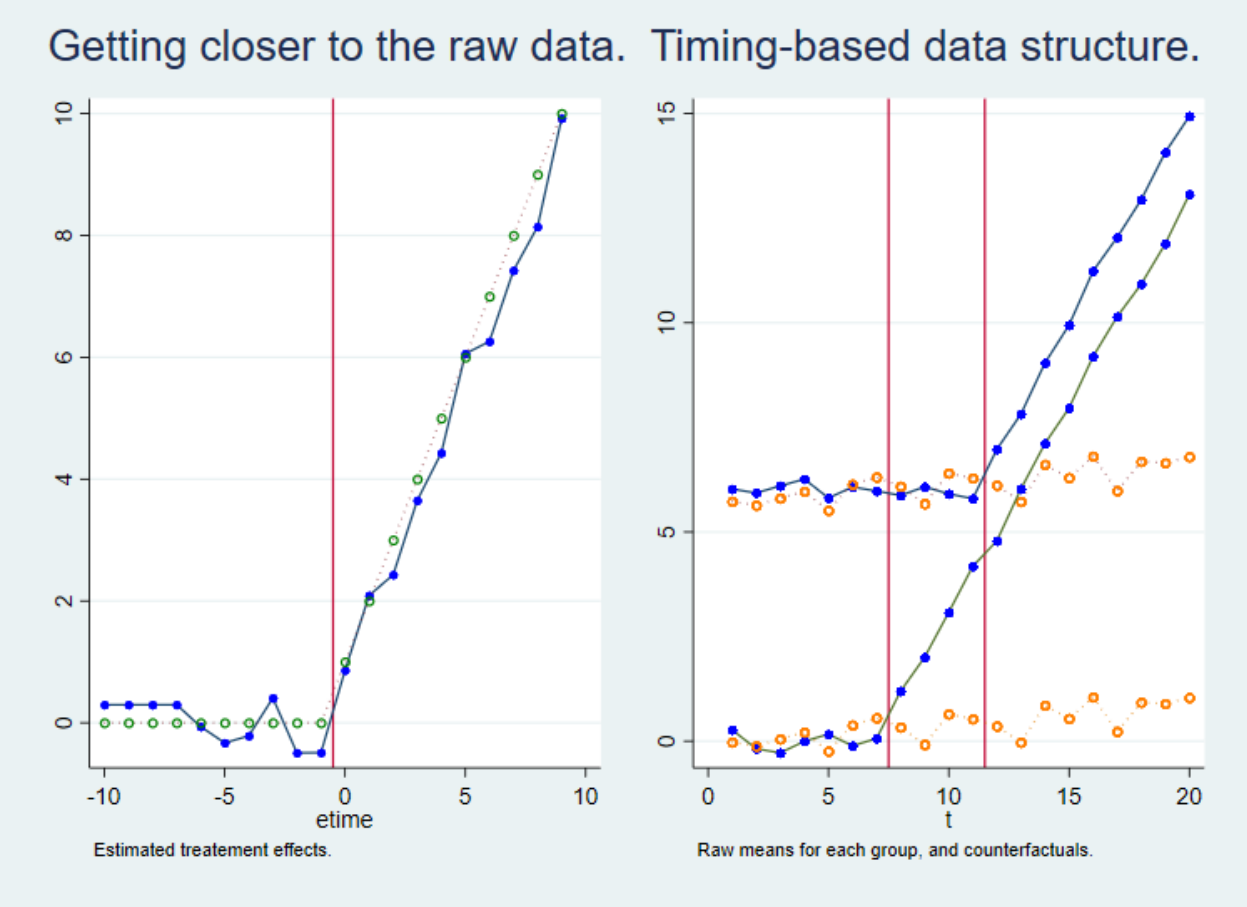
In this Appendix section I illustrate pooling event study coefficients, and imposing splines on event study coefficients for improved statistical power. These are discussed in section 3.6 in the paper.

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"



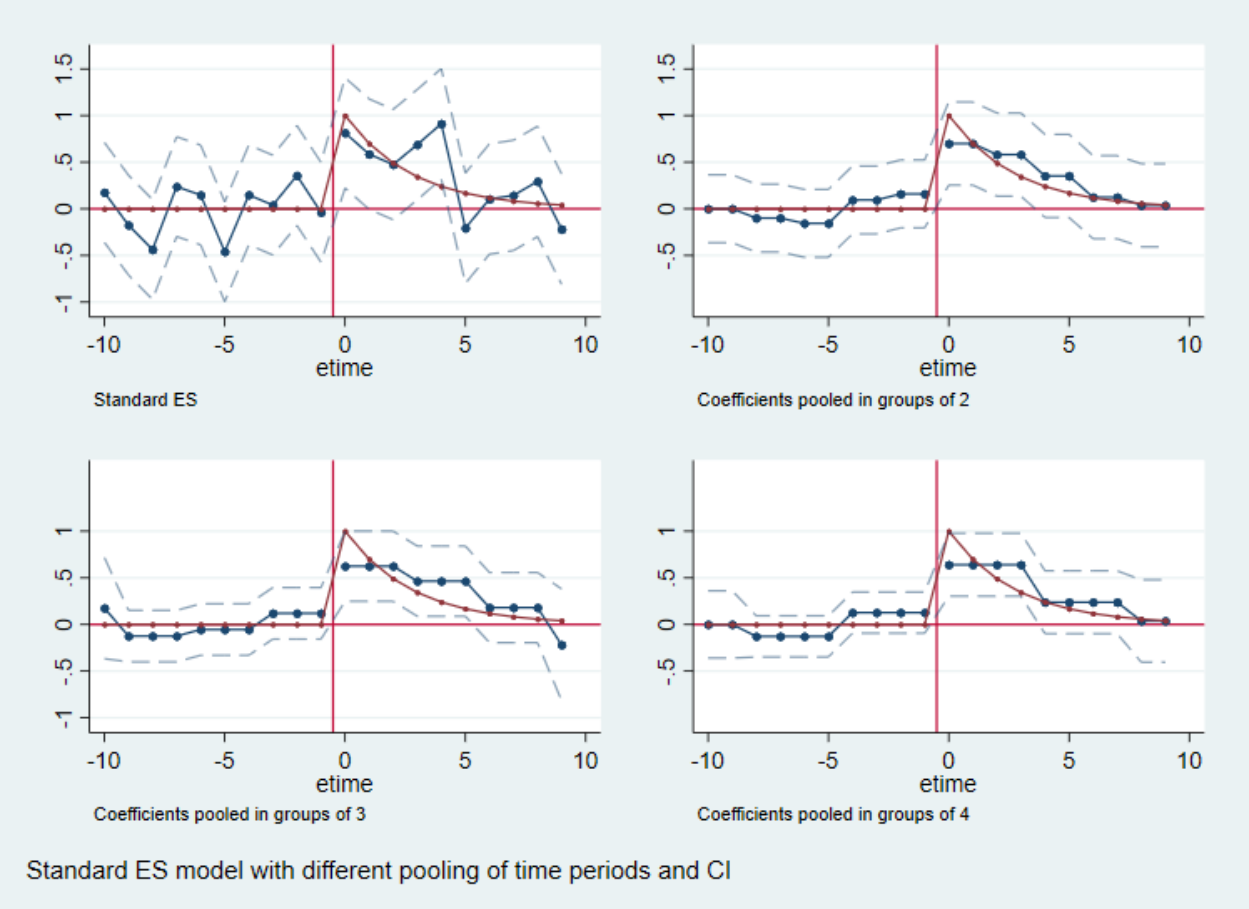
Note: In the left panel, the y-axis show the estimated (blue) and actual (green) treatment effects (γ_j). The x-axis shows event time. In the right panel, the x-axis shows calendar time. The blue connected lines shows mean outcomes for each of two types of treated units (who receive treatment at different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

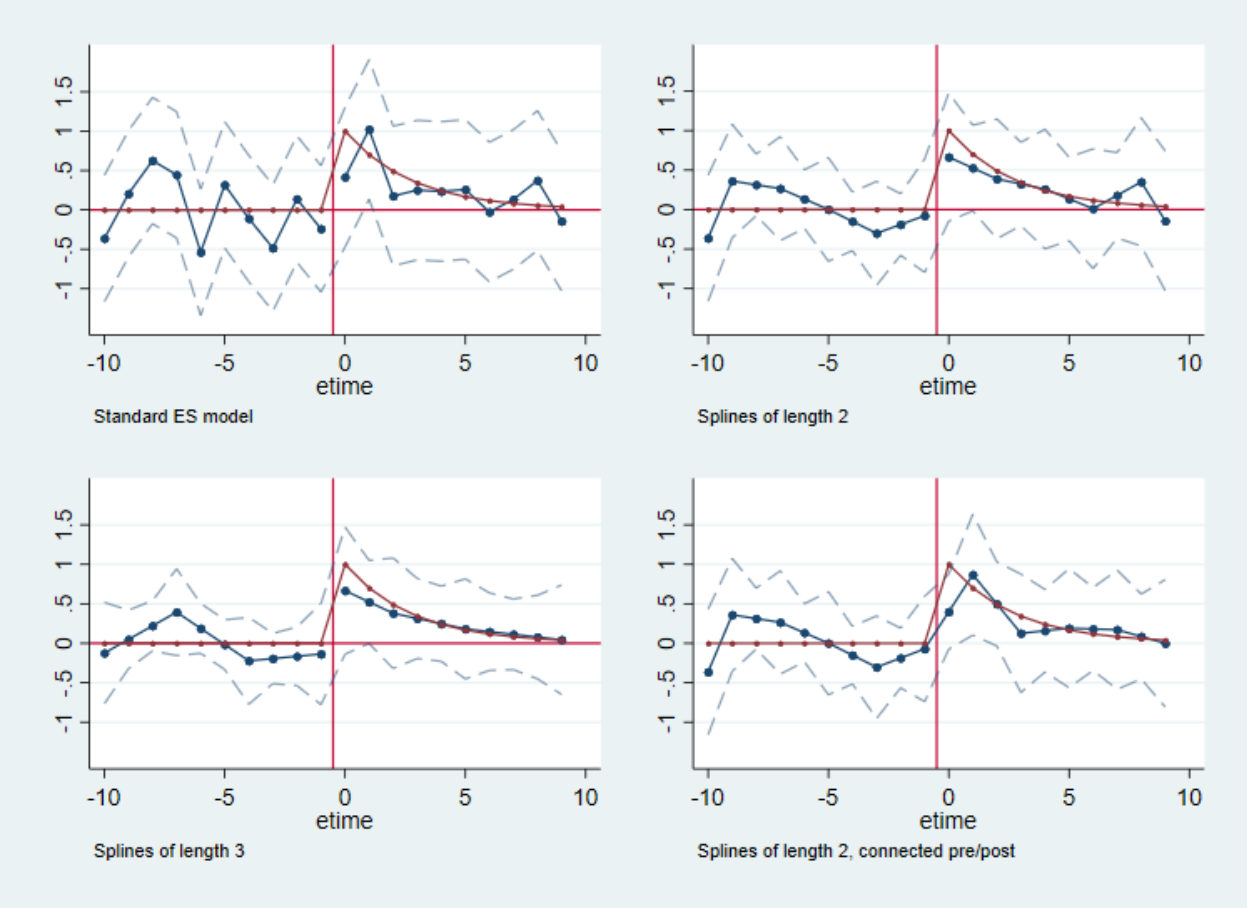
One alternative to pooling is to implement a spline model. This can be implemented by imposing "no concavity" constraints at the point of estimation. These constraints take the form of, e.g., $\gamma_1 - \gamma_0 = \gamma_2 - \gamma_1$ for connected segments of event time coefficients. Figure A.10 illustrates 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.

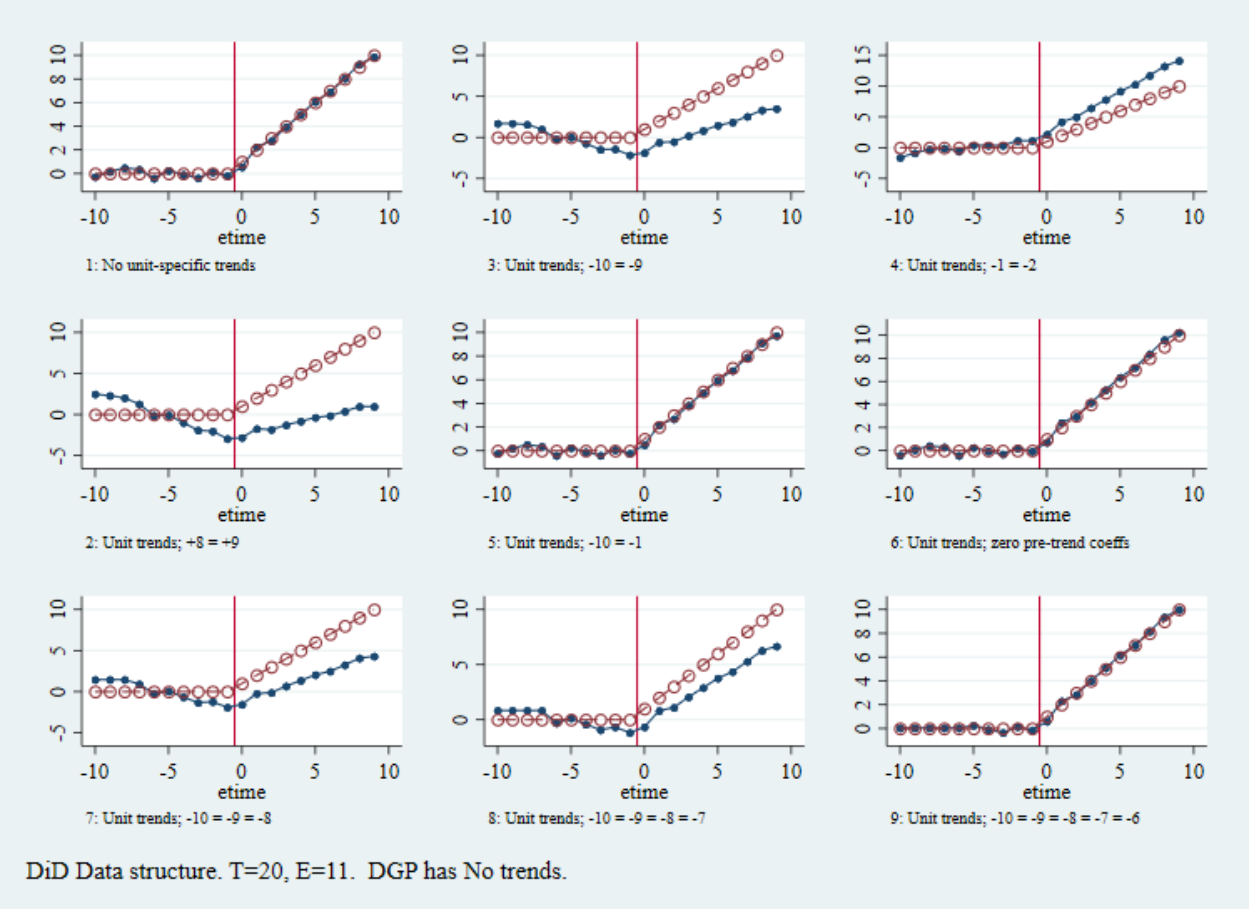
data generating process (DGP), this results in a mischaracterization of the effect at event time 0.

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.

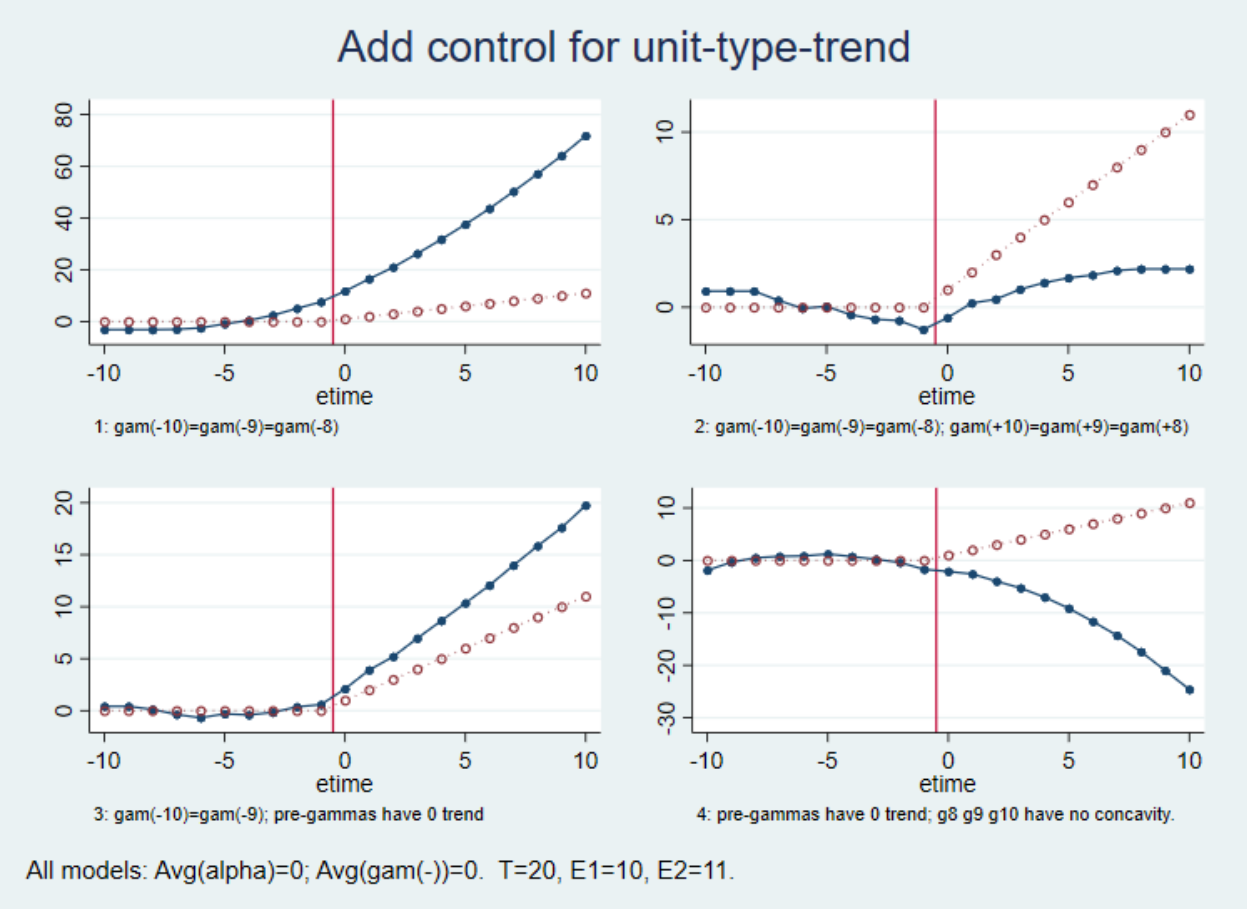
This complicated multicollinearity has two implications: (1) in order to get our model to be estimable, we will have to impose additional restriction(s); (2) these restrictions can interact with the complicated multicollinearity to produce unusual and unsettling results. Specifically, controlling for 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

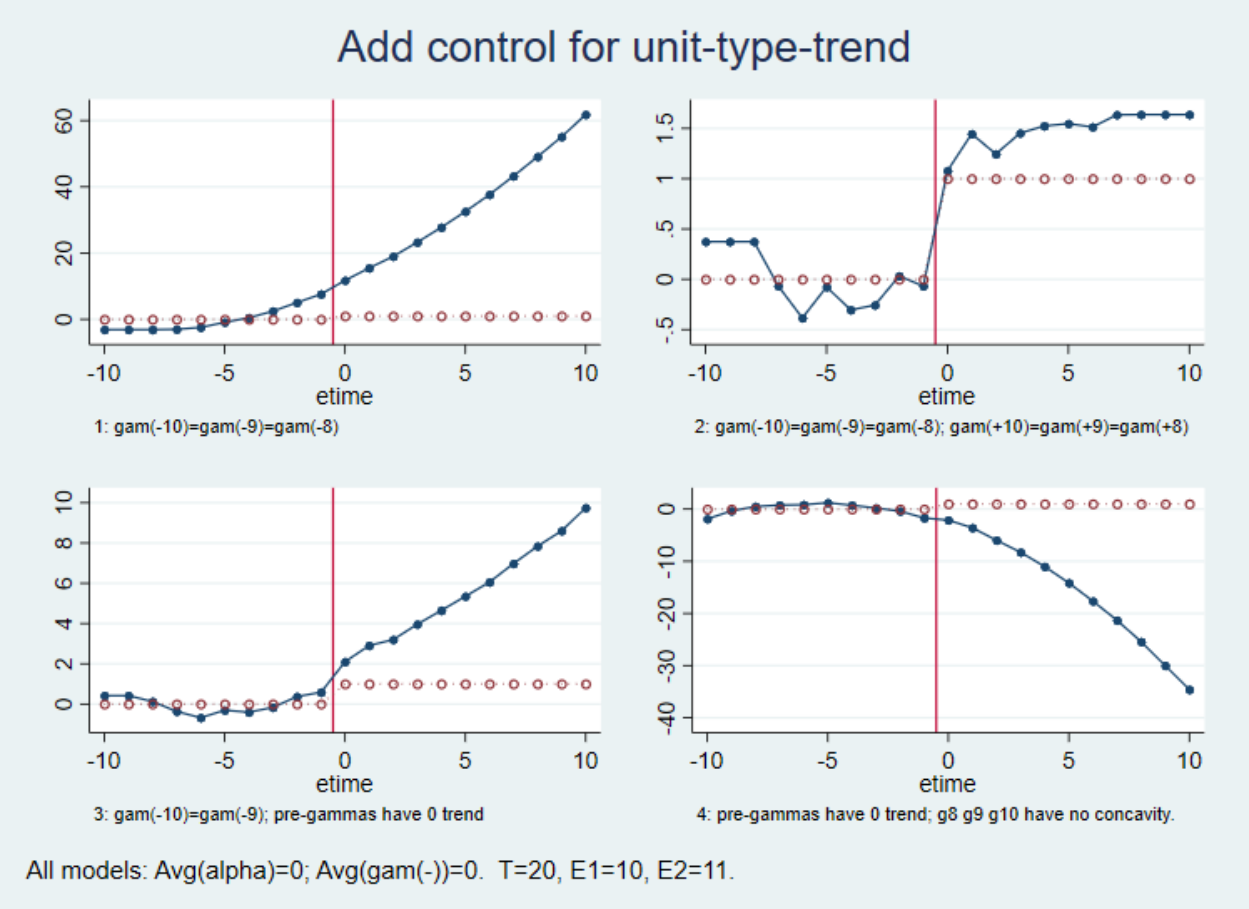
As in the case without trends, it is informative to show both the direct treatment effect estimates, as well as something that is closer to the raw data. Figure A.14 illustrates this, for three different models applied to the same data. Each model is in a different column,

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



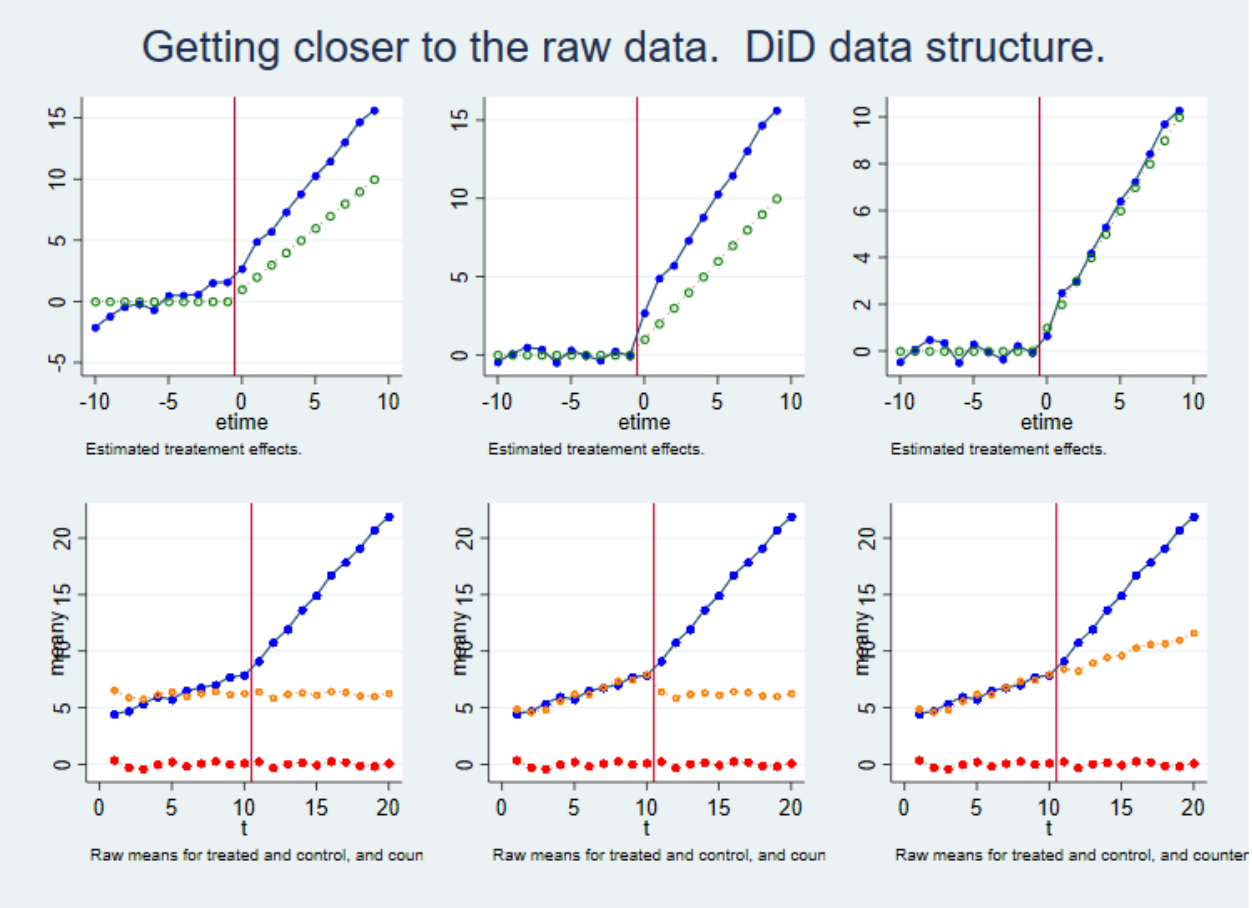
Note: Each panel estimates an event study model on the same data, which are from a timing-based data structure, with one unit treated at $t = 10$ and the other treated at $t = 11$. The true data generating process does not have any underlying trends. The true treatment effects (in red) follow a "ramp"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

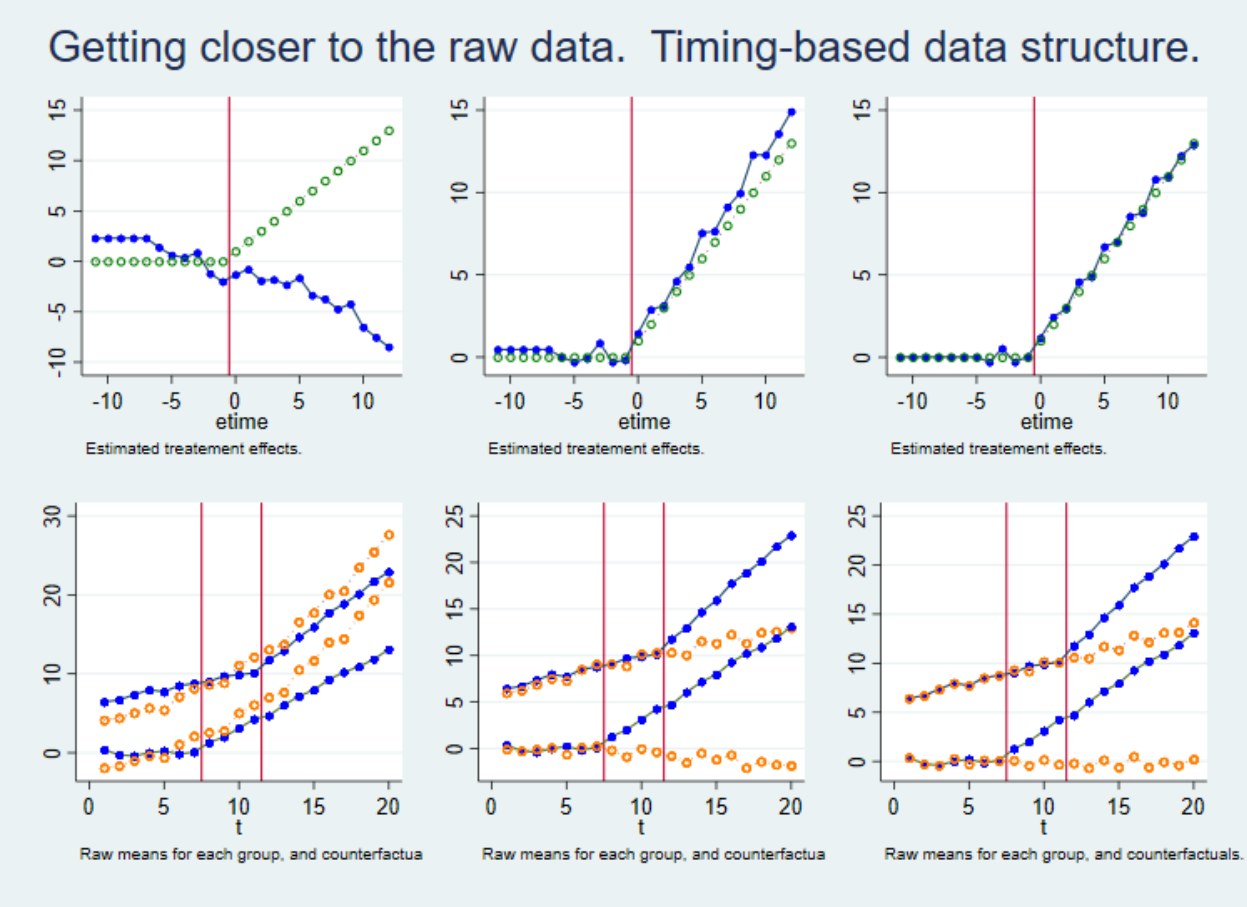
Next we consider the case of timing-based data structures and getting closer to the raw data. In the figure below, there are two unit types, one treated starting in period 8 and the other in period 12. The second unit type has a different underlying trend than the first. The figure shows three different 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.

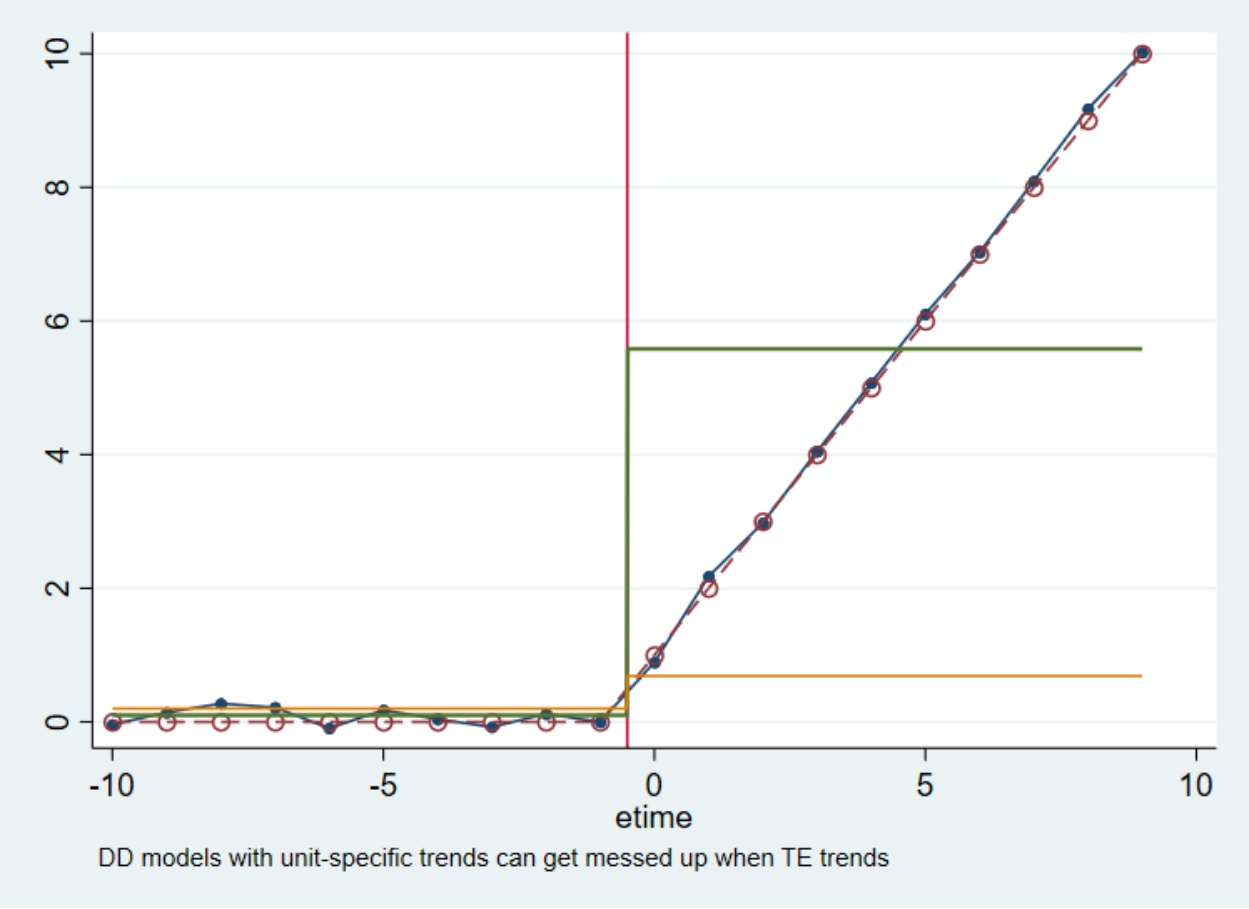
G.3 The Ben Olken Puzzle

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.