Measuring Heterogeneous Effects of Environmental Policies using Panel Data[*]

Douglas G. Steigerwald[†] Gonzalo Vazquez-Bare[‡] Jason Maier[§]

August 25, 2020

Abstract

To measure the effects of environmental policies, researchers often combine panel data with two-way fixed effects models. This approach implicitly assumes that the distribution of the policy effect is constant across units and over time. Yet many environmental policies have effects that differ depending on the unit exposed to the policy and the period in which the policy is applied. In this setting we detail why the model parameters generally do not capture a useful measure of the effects. We then show that in a multi-period setting, if the policy is applied in only one period, then the model parameters do capture a useful measure of the effects. In these settings, appropriate inference is based on cluster-robust standard errors. Because the resultant t-statistic may yield unreliable inference when clusters are heterogeneous, we present an appropriate measure of cluster heterogeneity and describe how the measure should be used to guide inference.

1 Introduction

In studying natural resources and the environment, the role of targeted programs and policies is often of paramount interest. In the absence of random assignment, many studies employ panel data and rely on two-way fixed effects to account for time-constant unobservable factors and common nonlinear effects over time. Two principle questions confront the researcher using this approach, namely how to interpret the parameters of the estimating equation in terms of the effect of the program and how to assess the significance of the findings. In this paper we address these questions in turn to provide a guide for estimation and inference in two-way fixed effects models with panel data.

To fix ideas, consider measuring the effect of a gasoline content regulation (as in Auffhammer and Kellogg 2011). The two-way fixed effects estimating equation is:

$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + \varepsilon_{it}, \quad t = 0, \ldots, T_i, \, i = 1, \ldots, G \tag{1}$$

where Y_{it} is a measure of air quality in state i at year t, α_i is a state-level fixed effect, δ_t is a year fixed effect and D_{it} is an indicator equal to one if state i in year t adopted a gasoline content regulation. To estimate the parameters of this equation it is typical to assume that the equation captures a feature of the conditional distribution of the outcome - most commonly the conditional mean. For this to be the case, it must be that $\mathbb{E}(\varepsilon_{it}|\alpha_i, \delta_t, D_{it}) = 0$, which identifies the coefficients in (1) as parameters of the conditional mean.

[*] We are grateful to Clément de Chaisemartin, Olivier Deschenes and Kyle Meng for valuable comments and discussions.

[†] Department of Economics, UC Santa Barbara, doug@ucsb.edu.

[‡] Department of Economics, UC Santa Barbara, gvazquez@econ.ucsb.edu.

[§] Bren School, UC Santa Barbara, jmaier@ucsb.edu.

When Equation (1) is seen as the true model generating the data, obtaining an unbiased and consistent estimate for the effect of the program, β, is a straightforward exercise in ordinary least squares algebra. But assuming Equation (1) is the correct specification implies imposing, among other things, homogeneity of the program effect over time and across units. Yet many environmental policies exhibit effects that are both heterogeneous and time-varying. In the case of estimating the effect of a gasoline content regulation discussed above, Auffhammer and Kellogg (2011) find that the flexibility allowed by regulators for refiners to choose which chemical components to remove from their gasoline resulted in substantial variation in the effectiveness of the policy.[1] As another example, firms with differing marginal abatement costs can respond differently to a pollution tax. Thus firm size, the energy intensity of the industry, and trade patterns can lead to different responses to the tax (e.g. Martin et al. 2014). Other examples include Auffhammer et al. (2009), Grainger (2012), Ferraro and Miranda (2013), Frondel and Vance (2013), Bento et al. (2015), and Sills and Jones (2018). While the two-way fixed effects estimator is one of the most popular tools in policy evaluation, most of the recent literature analyzing it under treatment effect heterogeneity draws rather pessimistic conclusions.

We review these recent advances in the methodological literature, couched within the framework in which the policy application is a treatment. A key component is to carefully describe the estimand of interest, which depends on the degree of heterogeneity in the *distribution* of treatment effects. We derive a representation of the parameter of Equation (1) that shows that, in general, it does not correspond to an estimand of interest, such as an average treatment effect. We detail why this is so and then describe the setting in which the model parameters do capture a useful measure of the treatment effects - namely, a multi-period setting in which the subset of units that receive treatment all do so in the same period. A special case of this is the classic difference-in-differences model of two periods, where all units are untreated in the first period and a subset of units are treated in the second period. In this setting a standard parallel trends assumption is sufficient for the parameters of (1) to provide useful information.

For the settings in which estimation of (1) provides an estimate of an average treatment effect, correct inference is based on a cluster-robust variance estimator. It is commonly believed that the resultant cluster-robust t statistic has a normal distribution if the number of clusters is large. This statement rests on an implicit assumption that the characteristics of the clusters are similar. If they are not, then even if the number of clusters is large, the t statistic can be non-normal. We show how to measure the dissimilarity across clusters to determine if the test statistic is approximately normal. If the dissimilarity is large, so that the normal approximation is poor, we show how to obtain critical values that provide a better approximation to the distribution of the t statistic.

In Section 2 we use a potential outcomes framework to carefully describe the possible assumptions about the heteogeneity in the distribution of treatment effects and the conditions needed to identify the resulting estimand of interest. In Section 3 we present a representation of β, and of the two-way fixed effects estimator, as a linear combination of the underlying average treatment effects. The linear combination, in general, does not correspond to an estimand of interest. Our results add to the existing literature in two ways. First, our result goes beyond previous results that focus only on properties of the estimator and make clear that the issue of weights is not only a property of the estimator but is intrinsic to the parameters of two-way fixed effects specifications. Second, we provide a convenient expression for the weights, to aid in understanding applied results. Section 4 presents the issues for inference that stem from cluster heterogeneity and describes a correct procedure for inference. Section 5 provides an empirical illustration of these findings and Section 6 provides a summary with practical recommendations.

[1] These authors have data that allow them to also construct regression discontinuity estimators, which can provide a useful alternative to two-way fixed effects estimators.

2 Specifying and Identifying Treatment Effects

2.1 Potential Outcomes and Estimands

Causal effects are naturally associated with potential outcomes (see Imbens and Rubin 2015, for a general treatment). To adapt potential outcomes to a multiperiod setting we must clearly define how the treatment is assigned over time and distinguish among the potential measures of causal effects.

To frame the discussion let the units under study be states, as in the motivating example of state regulation of gasoline content. The data come from a balanced panel, where for each state i in each time period $t = 0, 1, \ldots, T$, there exist two potential outcomes: the outcome that would hold if the state was subject to the regulation (treated), $Y_{it}(1)$, and the outcome that would hold if the state was not subject to the regulation (untreated), $Y_{it}(0)$. We assume that the potential outcomes for state i do not depend on the treatment status of any other state, that is, there are no treatment spillovers (see Deschenes and Meng 2018; Vazquez-Bare 2017, for recent work that allows for spillovers). Importantly, the potential outcomes, which could be measures of air quality or health, for example, are themselves random variables.

The observed outcomes for each state depend on the assignment of a treatment vector $\{D_{it}\}_{t=0}^{T}$, where each element is a binary treatment indicator. We assume that no state is treated in period 0, so that $D_{i0} = 0$ for all units. Once state i has been assigned to treatment it remains in treatment, so that $D_{it} \geq D_{it-1}$. In general, all units do not need to enter treatment in the same period (see Currie et al. 2015, for an example), which is often called a *staggered adoption* design. To capture this type of design, let the period in which the state is assigned to treatment, denoted t_i^*, be random. For states that are never assigned to treatment, $t_i^* = T + 1$. In detail,

$$t_i^* = \begin{cases} \min\{t : D_{it} = 1\} & \text{if } \sum_{t=0}^{T} D_{it} > 0 \\ T + 1 & \text{if } \sum_{t=0}^{T} D_{it} = 0 \end{cases}.$$

Because units never leave treatment, t_i^* alone conveys the same information as the entire vector of treatment indicators $\{D_{it}\}_{t=0}^{T}$, and the treatment indicator can be written as $D_{it} = \mathbb{1}(t \geq t_i^*)$.

The observed outcome, Y_{it}, is related to the potential outcomes in the following way:

$$Y_{it} = \begin{cases} Y_{i0}(0) & \text{if } t = 0 \\ Y_{it}(0) & \text{if } t \neq 0 \text{ and } D_{it} = 0 \\ Y_{it}(1) & \text{if } t \neq 0 \text{ and } D_{it} = 1 \end{cases},$$

which can be rewritten more compactly as:

$$Y_{it} = Y_{it}(0) + \tau_{it} D_{it}, \tag{2}$$

where $\tau_{it} = Y_{it}(1) - Y_{it}(0)$ is the treatment, or causal, effect for state i in period t. The treatment effect potentially differs over states and time periods, so this expression captures all possible heterogeneous treatment effects. Because the potential outcomes pair consists of the observed outcome and the counterfactual, unobserved outcome, the treatment effect is not observable.

In this context, it is possible to define several parameters of interest. The average treatment effect (ATE) in period t is the mean of the distribution of treatment effects in that period:

$$ATE_t = \mathbb{E}[\tau_{it}].$$

Identification of the ATE requires strong assumptions that rarely hold when treatment receipt is endogenous. For this reason, attention usually turns to the average treatment effect on the treated (ATT) in period t, which is the average effect for treated states in period t:

$$ATT_t = \mathbb{E}[\tau_{it} | D_{it} = 1].$$

To understand the complex dynamics of these parameters, consider the second period, $t = 2$, in detail:

$$ATT_2 = w_1 \mathbb{E}[\tau_{i2} \mid t_i^* = 1] + w_2 \mathbb{E}[\tau_{i2} \mid t_i^* = 2], \tag{3}$$

which is a weighted average of two components with weights $w_1 = \mathbb{P}[t_i^* = 1|D_{i2} = 1]$ and $w_2 = \mathbb{P}[t_i^* = 2|D_{i2} = 1]$. The first component is the average treatment effect in the second year of regulation for states that adopted the policy in period 1. The second component captures the effect for states that newly adopted the regulation in period 2. From this we can see that t_i^* partitions the population into disjoint groups according to when (and whether) they start receiving treatment. Because groups entering treatment in different periods can be heterogeneous and exhibit different average effects over time, this variable will play a key role in modeling treatment effect heterogeneity.

To isolate the components, and so aid in the identification of dynamic treatment effects and the analysis of two-way fixed effects estimators, we index the ATT parameters by the period in which a group enters treatment (Callaway and Sant'Anna 2019; Abraham and Sun 2020, also analyze this parameter). The resultant group-specific ATT is defined for each group entering treatment at time $s \leq t$ as:

$$ATT_t^s = \mathbb{E}[\tau_{it}|t_i^* = s].$$

The parameter ATT_2 in (3) can then be expressed as:

$$ATT_2 = w_1 ATT_2^1 + w_2 ATT_2^2.$$

We consider the set of group-specific treatment effects $\{\{ATT_t^s\}_{t=s}^T : s = 1, \ldots, T\}$ as the underlying parameters of interest. This is appropriate in the many contexts in which it is reasonable to assume that the treatment effects for all units that enter treatment at the same time are drawn from a common distribution, so that the ATT parameter is not indexed by i (see de Chaisemartin and D'Haultfœuille, forthcoming, for an alternative setup with non-identically distributed units). Our approach of allowing the ATT parameter to vary over both groups and time allows for a rich degree of heterogeneity. For example, this setting accommodates treatment effects that depend on how long the policy has been applied, in which case the treatment effects two years after the policy is implemented have a different mean than the treatment effects in the first year after the policy is implemented. The setting also allows for the treatment effects in the first year of adoption to differ with the period of adoption, as would be the case if states that are early adopters have a different distribution of effects than states that are later adopters.

2.2 Identification

We now focus on describing conditions for identification of group-specific average effects. While identification under a parallel trends assumption is well understood, we formally state a result to be clear about the assumptions needed to identify group-specific treatment effects.

Proposition 1 (Identification of group-specific effects) *Consider a pair of periods t and t' with $0 \leq t < t' \leq T$. Consider the set of units with $t_i^* = s$ where $t < s \leq t'$, who are untreated in period t and treated in period t', and the set of units with $t_i^* = u$ where $u > t'$, who remain untreated in both periods t and t'. If*

$$\mathbb{E}[Y_{it'}(0) - Y_{it}(0)|t_i^* = s] = \mathbb{E}[Y_{it'}(0) - Y_{it}(0)|t_i^* = u],$$

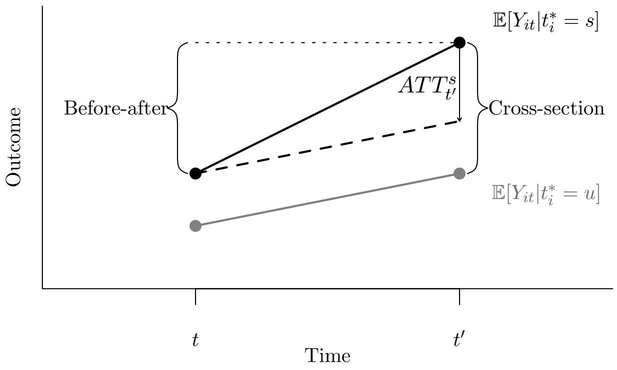
then $ATT_{t'}^s$ is identified as

$$\mathbb{E}[\tau_{it'}|t_i^* = s] = \mathbb{E}[Y_{it'} - Y_{it}|t_i^* = s] - \mathbb{E}[Y_{it'} - Y_{it}|t_i^* = u].$$

We provide a proof of this result in the appendix.

Proposition 1 shows that the group-specific ATT in period t' for units with $t_i^* = s$ can be expressed as a function of the observable data. More precisely, it shows that the average treatment effect in period t' for the group of units entering treatment at some previous (or the same) period s can be identified by comparing the evolution of their average outcome to the evolution of another group that remains untreated in both periods t and t'. The key assumption for this comparison to recover a causal parameter is the equality of the average trends of the

Figure 1: Difference-in-differences



$\mathbb{E}[Y_{it}|t_i^* = s]$
$ATT_{t'}^s$
Before-after
Cross-section
$\mathbb{E}[Y_{it}|t_i^* = u]$
Outcome
t
t'
Time

potential outcomes under no treatment between periods t and t', an assumption commonly known as the *parallel trends assumption.*

Figure 1 illustrates the parallel trends assumption. Graphically, the assumption holds when the observed average trend of the untreated group (the solid gray line) is parallel to the trend that would have been observed for the treated group, had the policy not been implemented (dashed black line).

2.3 Assessment of Identification

Identification assumptions are untestable by definition. The parallel trends assumption depends on counterfactual magnitudes that can never be observed. Hence the validity of this approach depends on the application: is the parallel trends assumption reasonable for the empirical question at hand?

While this question cannot be answered directly, related evidence is often marshalled to support identification. When multiple pre-treatment periods are available, researchers routinely assess the credibility of this assumption by checking whether trends were parallel before the treatment was implemented. Suppose that we have three periods, $t = 0, 1, 2$, where the policy is implemented in the last period $t = 2$ so that both $t = 0$ and $t = 1$ are pre-treatment periods, and a subgroup of units receives treatment in $t = 2$. Furthermore, suppose that both periods $t = 0$ and $t = 1$ are valid baselines, in the sense that the parallel trends assumption holds for both pairs:

$$\mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1] = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0]$$
$$\mathbb{E}[Y_{i2}(0) - Y_{i0}(0)|D_{i2} = 1] = \mathbb{E}[Y_{i2}(0) - Y_{i0}(0)|D_{i2} = 0].$$

The above immediately implies that the parallel trends assumption also holds between periods $t = 0$ and $t = 1$:

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_{i2} = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_{i2} = 0].$$

But since both $t = 0$ and $t = 1$ are pre-treatment periods, all units are untreated and hence the potential outcomes under no treatment are observable, $Y_{i0}(0) = Y_{i0}$ and $Y_{i1}(0) = Y_{i1}$. Therefore, the above equation reduces to:

$$\mathbb{E}[Y_{i1} - Y_{i0}|D_{i2} = 1] = \mathbb{E}[Y_{i1} - Y_{i0}|D_{i2} = 0] \tag{4}$$

and because all the magnitudes involved are observable, this condition can be tested in practice by simply comparing the observed outcome trends between the groups that will be treated and untreated in the last period.

Because the identification assumption is fundamentally untestable, the condition in (4) is neither necessary nor sufficient for the parallel-trends assumption in Proposition 1. The average trends between groups could be parallel between periods 0 and 1 and then diverge between periods 1 and 2, or vice versa. Nevertheless, condition (4) can provide some compelling evidence to support it. We discuss these issues in more detail in Section 6.

3 Two-Way Fixed Effects Estimation

Proposition 1 establishes identification of group-specific average effects under a parallel trends assumption without imposing parametric restrictions and without reference to any specific estimation strategy. In practice, these group-specific average parameters can be estimated nonparametrically by simply replacing:

$$\mathbb{E}[Y_{it'} - Y_{it} | t_i^* = s] - \mathbb{E}[Y_{it'} - Y_{it} | t_i^* = u]$$

by its sample counterpart based on observations drawn from the population of interest. It is important to note that the total number of parameters of interest can be large: one average effect for each group $t_i^* = s$ in each treated period $t \geq s$. This nonparametric estimation approach, to our knowledge, is not used in empirical research. Instead, researchers routinely rely on two-way fixed effects specifications as in Equation (1). The resultant two-way fixed effects estimator of β, based on pooling all observations, is the reported measure of the treatment effect. A natural question is therefore: how does this estimation method combine the heterogeneous treatment effects $\{ATT_t^s\}$ and does this summary measure have a clear causal interpretation?

To develop the links between the two-way fixed effects specification in (1) and the heterogeneous treatment effects, expand the decomposition of the observed outcomes (2) as

$$Y_{it} = Y_{i0}(0) + \sum_{j=1}^{t} \Delta_{ij}(0) + \tau_{it} D_{it},$$

where $\Delta_{ij}(0) = Y_{i,j}(0) - Y_{i,j-1}(0)$. This expression divides the observed outcome into a component that is time-invariant, $Y_{i0}(0)$, a term that changes over time but not with the treatment, $\sum_{j=1}^{t} \Delta_{ij}(0)$, and a treatment indicator with its corresponding (random) coefficient τ_{it}. Importantly, this expression allows for arbitrary heterogeneity of the coefficients across units and over time unlike the two-way fixed effects specification. Hence, this expression is similar to Equation (1), with the important difference that coefficients are random and can vary across units and over time.

To formally characterize the two-way fixed effects estimator rewrite (1) in deviations-from-mean form. We have

$$\ddot{Y}_{it} = \beta \ddot{D}_{it} + \ddot{\epsilon}_{it}, \tag{5}$$

where the double-demeaned variables are

$$\ddot{Y}_{it} = Y_{it} - \overline{Y}_i - \overline{Y}_t + \overline{Y}, \quad \ddot{D}_{it} = D_{it} - \overline{D}_i - \overline{D}_t + \overline{D}, \quad \ddot{\epsilon}_{it} = \epsilon_{it} - \overline{\epsilon}_i - \overline{\epsilon}_t + \overline{\epsilon},$$

and where

$$\overline{Y}_i = \frac{1}{T+1} \sum_{t=0}^{T} Y_{it}, \quad \overline{Y}_t = \frac{1}{G} \sum_{i=1}^{G} Y_{it}, \quad \overline{Y} = \frac{1}{G(T+1)} \sum_{t=0}^{T} \sum_{i=1}^{G} Y_{it},$$

with analogous definitions for the treatment variable and the error. The OLS estimator for β in (5) is the two-way fixed effects estimator:

$$\widehat{\beta}_{\text{FE}} = \frac{\sum_t \sum_i Y_{it} \ddot{D}_{it}}{\sum_t \sum_i D_{it} \ddot{D}_{it}}.$$

3.1 Staggered Adoption

We establish two properties under staggered adoption. First, we show that β, the treatment effect coefficient in the two-way fixed effects specification, is a linear combination of the underlying treatment effects given by $\{ATT_t^s\}$. While the weights in this linear combination sum to one, it is not necessarily a weighted-average because some of the weights may be negative. Even if all of the weights are non-negative, the resultant expression for β is not an intuitive weighted average of the underlying treatment effects and its causal interpretation is unclear. Second, we show that $\widehat{\beta}_{\mathrm{FE}}$ is both consistent and conditionally unbiased for β. Related results for $\widehat{\beta}_{\mathrm{FE}}$ have been shown in a variety of contexts (Laporte and Windmeijer 2005; Borusyak and Jaravel 2017; Athey and Imbens 2018; de Chaisemartin and D'Haultfœuille, forthcoming; Goodman-Bacon 2018; Imai and Kim, forthcoming).[2] Our results add to this literature by providing an alternative characterization of the weights that clearly shows the relationship between these weights and the treatment assignment (determined by t_i^*). According to this representation, the researcher can easily evaluate or estimate the weights solely based on the (observable) vector $\{t_i^*\}_{i=1}^G$.

In establishing the result, it will be helpful to define $D_i^* = \mathbb{1}(t_i^* \leq T)$, so that units with $D_i^* = 1$ are eventually treated and those with $D_i^* = 0$ are untreated in all the observed periods $t = 0, \ldots, T$.

Assumption 1 *The following conditions hold:*

1. ***Sampling:*** *the $(T+1)$-dimensional vectors $(Y_{it}(0), Y_{it}(1), t_i^*, D_{it})_{t=0}^T$ are iid across i.*
2. ***Treatment assignment:*** *each unit i enters treatment in period $t_i^* > 0$ and there is a subset of untreated units for which $t_i^* = T+1$.*
3. ***Parallel trends:*** *for each $t = 0, \ldots, T$, $\mathbb{E}[Y_{it}(0) - Y_{it-1}(0) | t_i^*] = \mathbb{E}[Y_{it}(0) - Y_{it-1}(0)]$.*
4. ***Regularity conditions:*** *all the required moments are bounded.*

Assumption 1.1 states that units are drawn independently from the same distribution, but allows for an arbitrary correlation structure within units over time (the correlation structure is further discussed in Section 4). Assumption 1.2 allows units to enter treatment in different periods as defined by t_i^*, subject to the existence of a pre-treatment period $t = 0$ in which no unit is treated, and the existence of a group that remains untreated in every period $t = 0, \ldots, T$. Assumption 1.3 is the parallel trends assumption, according to which the average trajectory of the potential outcomes under no treatment is the same regardless of when or whether units receive treatment. Finally, Assumption 1.4 imposes standard regularity assumptions to obtain probability limits.

To obtain a representation of β, let $\mathbb{E}[D_i^*]$ be the proportion of eventually treated units in the population and let $\mathbb{E}[t_i^* | D_i^* = 1]$ be the average entry period for these units. Also, for each period t let $\mathbb{E}[D_{it}] = \pi_t$ be the population proportion of units that are treated in period t.

Proposition 2 (Representation of β) *Under Assumption 1 the coefficient β from the two-way fixed effects specification (1) can be written as:*

$$\beta = \sum_{t=1}^{T} \sum_{s=1}^{t} \mathbb{E}[\tau_{it} | t_i^* = s]\, \omega_{st}$$

where

$$\omega_{st} = \frac{\left(\mathbb{E}[D_i^*] - \pi_t + \frac{1}{T+1}\left(s - \mathbb{E}[t_i^* | D_i^* = 1]\mathbb{E}[D_i^*]\right)\right) \mathbb{P}[t_i^* = s]}{\sum_{t=1}^{T} \sum_{s=1}^{t} \left(\mathbb{E}[D_i^*] - \pi_t + \frac{1}{T+1}\left(s - \mathbb{E}[t_i^* | D_i^* = 1]\mathbb{E}[D_i^*]\right)\right) \mathbb{P}[t_i^* = s]}.$$

[2] Related results for one-way fixed effects include Wooldridge (2005), Chernozhukov et al. (2013), Gibbons et al. (2018), and Słoczyński (2018).

We provide a proof of this result in the appendix.

To illustrate the fact that β is not necessarily a useful measure of treatment effects, consider the case with three time periods $t = 0, 1, 2$ and three groups. Group 1 enters treatment in the first period ($t_i^* = 1$), group 2 enters treatment in the second period ($t_i^* = 2$), and the remaining group is untreated. Suppose that 50 percent of the population is in group 1 and the remaining population is split evenly between the untreated group and group 2. The average treatment effects given by the proportionally-weighed average of the group-specific effects are

$$ATT_1 = ATT_1^1 \quad \text{and} \quad ATT_2 = \frac{2}{3}ATT_1^2 + \frac{1}{3}ATT_2^2.$$

If each period is equally weighted, then

$$ATT = \frac{1}{2}ATT_1 + \frac{1}{2}ATT_2 = \frac{1}{2}ATT_1^1 + \frac{1}{3}ATT_1^2 + \frac{1}{6}ATT_2^2. \tag{6}$$

Computation of ω_{st} reveals[3]

$$\beta = \frac{3}{5}ATT_1^1 + 0\, ATT_1^2 + \frac{2}{5}ATT_2^2,$$

which does not correspond to the ATT in (6). Moreover, β will not equal $ATT = \lambda ATT_1 + (1 - \lambda)ATT_2$ for any value of $\lambda \in [0, 1]$.

One of the most perplexing aspects of Proposition 2 is the possibility that some weights in the linear combination are negative, a point that has been emphasized in the recent literature. How does this come about? Mechanically, negative weights require $s < \mathbb{E}[t_i^*|D_i^* = 1]\, \mathbb{E}[D_i^*]$, which is more likely in later periods (so that $\mathbb{E}[D_i^*] - \pi_t$ is close to zero) for units who entered treatment in earlier periods (so that s is small). In the three period case, if 25 percent of the population is untreated, then when the proportion in group 1 falls below 50 percent of the population, $\omega_{12} < 0$. We display this in Figure 2, where the horizontal axis measures the proportion of the population in group 1. While ω_{11} and ω_{22} are always between 0 and 1, ω_{12} turns negative when the proportion in group 1 falls below 0.5.

Negative weights are likely to be particularly problematic when the effect of the treatment is lagged or accumulates over time. As an extreme case, suppose that the average effect of the treatment is 0 for all groups in the first period in which they are treated, so that $\mathbb{E}[\tau_{i1}|t_i^* = 1] = \mathbb{E}[\tau_{i2}|t_i^* = 2] = 0$, and equals some positive value $\theta > 0$ in the second year of treatment, $\mathbb{E}[\tau_{i2}|t_i^* = 1] = \theta$ (which is only observed for group 1). In this case, $\beta = \omega_{12}\theta$, so if $\omega_{12} < 0$ then $\beta < 0$ even when the average effect is non-negative for all units in all periods.

To derive the related properties for the *estimator* of β, let $\mathbf{D}_i = (D_{i1}, \ldots, D_{iT})$ and $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_G)$.

Proposition 3 (Properties of $\widehat{\beta}_{\text{FE}}$) *Under Assumption 1*

$$\mathbb{E}[\widehat{\beta}_{\text{FE}}|\mathbf{D}] = \sum_{t=1}^{T}\sum_{s=1}^{t} \mathbb{E}[\tau_{it}|t_i^* = s]\, \widehat{\omega}_{st}$$

with

$$\widehat{\omega}_{st} = \frac{\left(\overline{D}^* - \overline{D}_t + \frac{1}{T+1}\left(s - \overline{t}^*\overline{D}^*\right)\right)\widehat{p}_s}{\sum_{t=1}^{T}\sum_{s=1}^{t}\left(\overline{D}^* - \overline{D}_t + \frac{1}{T+1}\left(s - \overline{t}^*\overline{D}^*\right)\right)\widehat{p}_s}$$

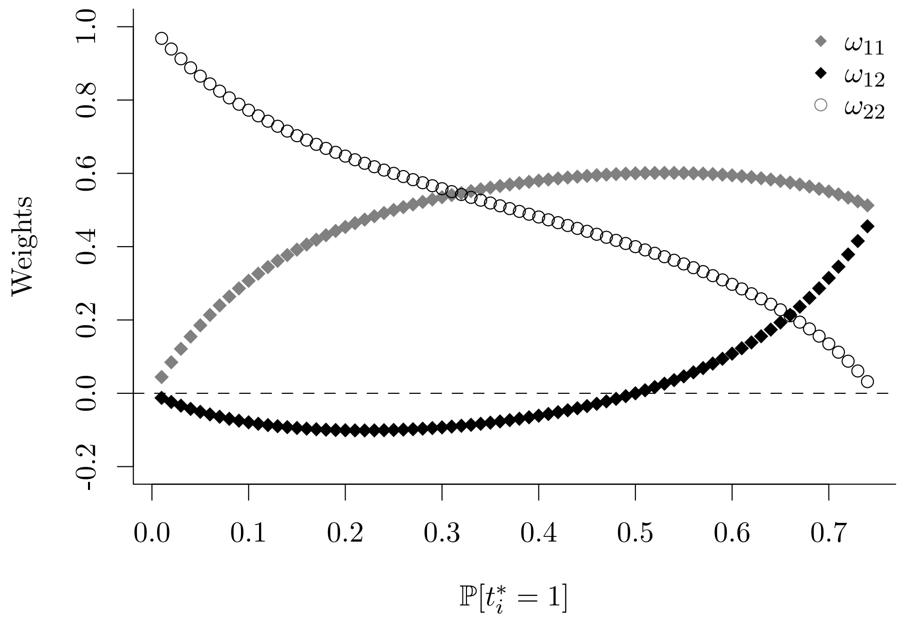
where $\overline{D}^* = \frac{1}{G}\sum_i D_i^*$ *is the sample proportion of eventually treated units,* $\overline{D}_t = \frac{1}{G}\sum_i D_{it}$ *is the sample proportion of treated units at time* t*,* $\overline{t}^* = \sum_i t_i^* D_i^* / \sum_i D_i^*$ *is the sample average entry period for treated units and* $\widehat{p}_s = \frac{1}{G}\sum_i \mathbb{1}(t_i^* = s)$ *is the sample proportion of units entering treatment at time* s*.*

Further, as $G \to \infty$ *with* T *fixed,*

$$\widehat{\beta}_{\text{FE}} \to_{\mathbb{P}} \beta.$$

[3] Here $\mathbb{E}(D_i^*) = 3/4$, $\mathbb{E}(t_i^* \mid D_i^* = 1) = 4/3$, $\pi_1 = 1/2$, $\pi_2 = 3/4$, $\mathbb{P}(t_i^* = 1) = 1/2$, and $\mathbb{P}(t_i^* = 2) = 1/4$.

Figure 2: Weights assigned by $\widehat{\beta}_{\text{FE}}$



We provide a proof of this result in the appendix.

Proposition 3 establishes that the expected value of the two-way fixed effects estimator is a linear combination of the underlying treatment effects where the weights are the sample counterparts of the weights in the representation for β. Unsurprisingly, as the number of units grows holding T fixed, $\widehat{\beta}_{\text{FE}}$ is consistent for β.

Mirroring the target parameter β, the two-way fixed effects estimator will not necessarily provide a useful measure of treatment effects, due to the weights in the linear combination. An alternative description of how negative weights arise is provided by Goodman-Bacon (2018) and Imai and Kim (2019) who show that $\widehat{\beta}_{\text{FE}}$ can be expressed as a linear combination of all possible individual difference-in-differences estimators. Each difference-in-differences estimator compares two groups over two time periods, such as the comparison

$$\mathbb{E}[Y_{i1} - Y_{i0} | t_i^* = 1] - \mathbb{E}[Y_{i1} - Y_{i0} | D_i^* = 0],$$

between post-treatment period $t = 1$ and the pre-treatment period $t = 0$ comparing the $t_i^* = 1$ group to the untreated group. Not all of the comparisons use only the untreated units to form the control group, and it is the use of a "pseudo-control" group that leads to negative weights. The pseudo-control group arises because the two-way fixed effects estimator implicitly compares units whose treatment status changes between periods to units whose treatment status remains unchanged between periods. Thus in period 2, the units that entered treatment in period 1 now form a pseudo-control group with a treatment effect that enters negatively in the two-way fixed effects estimator.

In all, under staggered adoption the parameter β and its corresponding two-way fixed effects estimator do not generally recover a meaningful measure of the effects of a policy. There is one case, however, in which they do. If the treatment effects $\{\tau_{it}\}$ are iid over both *i and t*, then the two-way fixed effects specification in (1) is correct and β is the average treatment effect on the treated.

Corollary 1 (Constant average treatment effects) *Suppose that, in addition to Assumption 1, there exists some τ such that $\mathbb{E}[\tau_{it}|t_i^* = s] = \tau$ for all $1 < s \leq T$ and $s \leq t \leq T$. Then*

$$\beta = \tau,$$

$$\mathbb{E}[\widehat{\beta}_{\text{FE}}|\mathbf{D}] = \beta,$$

and as $G \to \infty$ with T fixed,

$$\widehat{\beta}_{\text{FE}} \to_{\mathbb{P}} \beta.$$

We provide a proof of this result in the appendix.

This result shows that under constant average treatment effects, $\widehat{\beta}_{\text{FE}}$ is an unbiased and consistent estimator of the causal treatment effect. Note, this result does not require constant treatment effects, an unreasonably strong assumption that requires $\tau_{it} = \tau$. Rather, the requirement is that the distribution, or at least the expectation, of τ_{it} not vary across individuals and over time.

3.2 Simultaneous Adoption

The discussion so far suggests that the two-way fixed effects specification does not recover a causally interpretable estimand for average treatment effects. A key reason for this is the fact that previously treated groups (whose status remains unchanged) are classifed as untreated groups in later periods. Hence, we may suspect that under simultaneous adoption of the treatment, that is, when all treated units receive treatment in the same period, the problems with two-way fixed effects estimation disappear. The following corollary shows that this is indeed the case.

Corollary 2 (Simultaneous adoption) *Suppose that, in addition to Assumption 1, all treated units enter treatment in the same period $0 < t^* \leq T$, that is, $t_i^* = t^*$ for all i such that $D_i^* = 1$. Then,*

$$\beta = \frac{1}{T + 1 - t^*} \sum_{t=t^*}^{T} \mathbb{E}[\tau_{it}|D_i^* = 1],$$

$$\mathbb{E}[\widehat{\beta}_{\text{FE}}|\mathbf{D}] = \beta,$$

and as $G \to \infty$ with T fixed,

$$\widehat{\beta}_{\text{FE}} \to_{\mathbb{P}} \beta.$$

We provide a proof of this result in the appendix.

According to this result, when all treated units enter treatment at the same time, the two-way fixed effects estimator is unbiased and consistent for a target parameter (β) that is a simple average of the average treatment effects over the post-treatment periods. Notice that the sum runs over the periods $t^* \leq t \leq T$, a total of $T + 1 - t^*$ periods, giving equal weights $1/(T + 1 - t^*)$ to each period. Hence, with simultaneous adoption, two-way fixed effects estimation recovers an easily interpretable summary of the effect of a policy. Also note that the parameter β is implicitly a function of T, the period being modeled. An implication is that, by extending a data set a researcher explicitly changes T, which not only alters the estimator but also the underlying estimand.

3.3 Implications for Empirical Practice

As noted at the beginning of Section 3, under the parallel-trends assumption it is possible to estimate the group-specific treatment effects in each post-treatment period, which can give a more complete assessment of the treatment effect heterogeneity both across groups and over time. This is proposed by Callaway and Sant'Anna (2019), who also provide inference methods for both the group-specific effects and summary measures of these effects. Under a more general heterogeneity assumption with non-iid units, de Chaisemartin and D'Haultfœuille (forthcoming) propose an alternative estimand that focuses on the average effect for each group on the period in which they switch from untreated to treated.

With many groups and time periods, the total number of parameters can be quite large, and hence the researcher may be interested in summarizing these effects in some way. With staggered adoption, it is always possible to apply Corollary 2 by doing pairwise comparisons between each treated group and the untreated group. More precisely, one can estimate Equation (1) on the subset of observations with $t_i^* = s$ and the untreated group to obtain one estimator $\widehat{\beta}^s_{\mathsf{FE}}$ for each value of s, which gives an unbiased and consistent estimator of the average effect for group s. The set of coefficients $\{\widehat{\beta}^s_{\mathsf{FE}}\}_s$ can help analyze the heterogeneity in average effects across groups, or can be used to construct a summary measure:

$$\widehat{\beta}_\lambda = \sum_s \lambda_s \widehat{\beta}^s_{\mathsf{FE}}$$

where $\{\lambda\}_s$ is a set of weights chosen by the researcher (such as, e.g., the sample proportion of each group).

A key takeaway from this discussion is that the weighting problems pointed out in Proposition 3 are due exclusively to the choice of the 2WFE specification (1) and its corresponding estimator $\widehat{\beta}_{\mathsf{FE}}$, and are unrelated to identification. In other words, even when β and $\widehat{\beta}_{\mathsf{FE}}$ are unable to recover an accurate measure of the average treatment effects, these effects are identified as long as the parallel trends assumption holds. Thus, as serious as these problems can be, they can be easily avoided by considering more flexible estimators that do not pool all observations together.

3.4 Dynamic specifications

A common specification when estimating treatment effects with panel data is a dynamic specification of the form:

$$Y_{it} = \alpha_i + \delta_t + \sum_{l=1}^{L} \gamma_l D_{it+l} + \sum_{k=0}^{K} \beta_k D_{it-k} + \varepsilon_{it} \tag{7}$$

This specification has two sets of parameters of interest. First, β_k, corresponding to the current treatment indicator and its lags, aim at estimating the effects of the treatment over time, that is, how the treatment effect changes in each time period once a unit gets treated. Second, γ_l, corresponding to leads of the treatment variables, are usually interpreted as "placebo tests" that measure whether the trends in outcomes between treated and untreated groups were already different before the treatment. The γ_l coefficients are expected to be non-significant if the parallel trends assumption holds. Specification (7), sometimes called an *event study*, can therefore be seen as a generalization of model (1) that allows the researcher to estimate treatment effects more flexibly while providing evidence to support the parallel trends assumption.

The lead coefficients may capture anticipatory effects if units can react to the treatment before it is implemented (see e.g. Laporte and Windmeijer 2005). In such cases, the researcher may shift the start date of the treatment to the first period in which anticipatory effects may be a concern (for example, when the policy is first announced). Borusyak and Jaravel (2017) and Abraham and Sun (2020) also analyze this type of dynamic specification in staggered adoption designs with heterogeneous treatment effects. Similarly to what happens in two-way FE models, estimators for the β_k coefficients may average across different treatment effects with misleading weighting schemes including negative weights. Furthermore, Abraham and Sun (2020) show that the γ_l coefficients may pick up some of the treatment effects in future periods, yielding spurious significant estimates even when the parallel-trends assumption holds.

As suggested by the discussion in Section 3.2, these issues are not a concern when treatment adoption is simultaneous, and in such cases Equation (7) can be used to estimate time-varying treatment effects and test for pre-treatment parallel trends. On the other hand, Callaway and Sant'Anna (2019) and Abraham and Sun (2020) propose alternative estimators that are robust to effect heterogeneity based on separately estimating and averaging group-specific ATTs.

3.5 Controlling for Covariates

So far we have only discussed the canonical two-way fixed effects specification that includes a treatment indicator and unit and time fixed effects. In many cases, a researcher may suspect that the parallel trends assumption will hold after conditioning on a vector of exogenous covariates Z_{it}, or may want to include observable characteristics to reduce the variability of the estimates. A commonly used specification in this setting is one in which covariates enter linearly:

$$Y_{it} = \alpha_i + \delta_t + \beta D_{it} + Z_{it}^{\mathrm{T}}\gamma + \eta_{it}$$

While the inclusion of covariates may make the parallel trends assumption more credible in some contexts, covariate-adjusted estimators are still subject to the problems highlighted above related to the negative weights with staggered adoption (de Chaisemartin and D'Haultfœuille, forthcoming). See also Słoczyński (2018) for related work on one-way fixed effects models. Importantly, including covariates in this fashion implicitly imposes a linear relationship between the outcome and the covariates, which may add misspecification bias. Abadie (2005) proposes a semiparametric reweighting method to control for covariates in a difference-in-differences setting, and Callaway and Sant'Anna (2019) generalize this method to allow for variation in treatment timing and propose an estimator that is not subject to the weighting problems highlighted above.

4 Inference

For settings in which two-way fixed effects estimation does deliver interpretable measures of causal effects, the next question is how to perform correct inference.[4] As discussed in the sampling assumption above, it is common to treat all the observations for a given unit as belonging to a single cluster. The logic of this stems from the fact that environmental policies are often set at a regional level (e.g. a county) and there are many other factors that affect the outcome and are common within the region. Inference is based on statistics constructed from cluster-robust standard errors, which allow for arbitrary error correlation within each cluster but require that the errors be independent across clusters. These standard errors allow for flexibility in the correlations among the errors within a cluster but reduce the sample size on which the normality of the test statistic is based. We discuss how non-normality arises, what a researcher can do to determine if their data set likely results in non-normality of the test statistic, and how to conduct inference if the test statistic is non-normal.

Let Ω be the variance matrix for the errors $\{\varepsilon_{it}\}$. Because the errors are independent across clusters, Ω is a block diagonal matrix, where diagonal block i corresponds to cluster (unit) i. The structure is summarized as

$$\mathbb{E}[\varepsilon_{it}\varepsilon_{js}] \neq 0 \text{ if } i = j$$

otherwise $\mathbb{E}[\varepsilon_{it}\varepsilon_{js}] = 0$. Importantly, the correlation structure for one unit need not be the same as the correlation structure for any other unit.

Let X represent the matrix of all fixed effects and the treatment indicator and θ be the vector that captures all the coefficients in (1), the individual and time fixed effects as well as the treatment effect, β. Let $\widehat{\theta}$ be the OLS estimator of θ with variance matrix $V = Var(\widehat{\theta}|X)$. Because Ω is block diagonal, the common expression

$$V = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Omega X(X^{\mathrm{T}}X)^{-1},$$

simplifies considerably to

$$V = (X^{\mathrm{T}}X)^{-1}\sum_{i=1}^{G} X_i^{\mathrm{T}}\Omega_i X_i\,(X^{\mathrm{T}}X)^{-1},$$

[4] This discussion assumes that the sample is drawn from an arbitrarily large population. In settings in which the sample is a non-neglibible fraction of the population, an adjustment needs to be made. See Cochran 1946 and Abadie et al. 2020.

where X_i and Ω_i correspond to the explanatory variables and error variance matrix for unit i. This simplification greatly reduces the number of unknown terms in Ω, so it is possible to obtain a consistent estimator of V without further restrictions on Ω. The widely used cluster-robust variance estimator employs the OLS residuals for each unit, $\widehat{\varepsilon}_i$, as

$$\widehat{V} = (X^{\mathrm{T}}X)^{-1} \sum_{i=1}^{G} X_i^{\mathrm{T}} \widehat{\varepsilon}_i \widehat{\varepsilon}_i^{\mathrm{T}} X_i \, (X^{\mathrm{T}}X)^{-1}. \tag{8}$$

We study tests of a null hypothesis of the form $a^{\mathrm{T}}\theta$. This encompasses a test of any one coefficient as well as tests on linear combinations of the coefficients. The test statistic is the t-statistic

$$t^{stat} = \frac{a^{\mathrm{T}}(\widehat{\theta} - \theta)}{\sqrt{a^{\mathrm{T}}\widehat{V}a}}. \tag{9}$$

In general, when estimating (1) interest centers on the treatment effect through the specific null hypotheses $H_0 : \beta = 0$, for which the test statistic takes the familiar form

$$t^{stat} = \frac{\widehat{\beta}}{\widehat{s}_{\widehat{\beta}}},$$

with $\widehat{s}_{\widehat{\beta}}$ the cluster-robust estimator of the standard error for $\widehat{\beta}$.

4.1 Effective Sample Size and Approximate Normality of the t statistic

It is common practice to report the number of clusters when basing hypothesis tests on cluster-robust standard errors. It may not be clear exactly why this is done, given that the estimator of the coefficients is governed by the sample size n. The reason is that the estimator of the variance, $\widehat{V}$, depends only on the variation between clusters, and so is a function of the number of clusters rather than the sample size. To see this, consider an intercept-only model in place of (1), so that $\widehat{\varepsilon}_{it} = Y_{it} - \overline{Y}$. The cluster-robust estimator $\widehat{V}$ is a function of $X_i^{\mathrm{T}}\widehat{\varepsilon}_i$, which in this intercept only model equals

$$\sum_{t=1}^{T_i}(Y_{it} - \overline{Y}_i) + \sum_{t=1}^{T_i}(\overline{Y}_i - \overline{Y}),$$

where the first sum captures the within-cluster variation and the second sum captures the between-cluster variation. Because the first sum is identically zero by construction, the cluster-robust variance estimator is a function only of between cluster variation, which is governed by the number of clusters, rather than the total number of observations.

The cluster-robust variance estimator in (8) is sensitive to heterogeneity in the between-cluster variation. Carter et al. (2017) (CSS) establish that the mean-squared error of $\widehat{V}$ increases with this variation. In establishing the asymptotic normality of t^{stat}, CSS introduce an adjustment to the sample size, that is an adjustment to the number of clusters, to account for the heterogeneity. The adjusted number of clusters is called the effective number of clusters because this is the relevant quantity to determine the accuracy of the normal approximation.

To fix ideas, consider the intercept-only model for which the heterogeneity over clusters is given by the variation in $T_i(\overline{Y}_i - \overline{Y})$ over i, which depends on variation in both the cluster sizes, T_i, and the cluster covariance matrices, Ω_i. The adjustment is constructed as follows. First, for each cluster construct the measure

$$\gamma_i^* = (T_i)^{-1} \sum_{i,j} \Omega_{i,j} (T_i)^{-1}, \tag{10}$$

where $\sum_{i,j} \Omega_{i,j}$ is the sum of all of the elements in Ω_i. The adjustment term is

$$\Gamma^* = \frac{\frac{1}{G}\sum_{i=1}^{G}(\gamma_i^* - \overline{\gamma}^*)^2}{\overline{\gamma}^{*2}}, \tag{11}$$

where $\overline{\gamma}^* = \frac{1}{G}\sum_{i=1}^{G}\gamma_i^*$. If all the clusters have the same size, $T_i = T$ for all i, and the same error covariance matrices, $\Omega_i = \Omega$ for all i, then $\gamma_i^* = \overline{\gamma}_i^*$ and the adjustment term is zero. As the heterogeneity across clusters increases, as would occur with variation in cluster sizes, Γ^* increases.

The resulting effective number of clusters is

$$G^* = \frac{G}{1+\Gamma^*}. \tag{12}$$

Because the adjustment is multiplicative, the effective number of clusters can be substantially less than G, so that a large value of G is not sufficient to guarantee approximate normality of t^{stat}.

For a model with explanatory variables the cluster-level heterogeneity measure is

$$\gamma_i^* = a^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X_i^{\mathrm{T}}\Omega_i X_i(X^{\mathrm{T}}X)^{-1}a, \tag{13}$$

which is then used to form the adjustement term in (11). Two further points stand out. First, the adjustment differs with the coefficient under test. For example, if an explanatory variable takes very different values over clusters, then the adjustment will be larger for test of the corresponding coefficient. Second, the adjustment depends on the unknown cluster correlations through Ω_i. (Because Γ^* is scale invariant, the correlation matrix can be used to construct the adjustment.) In practice, the adjustment is often constructed with a unit correlation matrix, in which all correlations equal 1, which generally provides an upper bound on Γ^*. This would lead to conservative inference - if the adjustment constructed in this way leads to a large value of the effective number of clusters, then approximate normality of the t statistic is justified.

4.2 Correct Inference

If the effective number of clusters is large, then the distribution of t^{stat} is approximately normal. When the effective number of clusters is smaller, two factors lead to non-normality of the t-statistic. The first factor is the downward bias in the estimated standard error. Lower values of G^* lead to more substantial downward bias in the cluster-robust standard error. The second factor is the increasing variation in the standard error. Lower values of G^* lead to larger variation in the cluster-robust standard errors, which in turn leads to non-normality of the t-statistic. If G^* is extremely low, then the distribution of t^{stat} can be bimodal.

MacKinnon and Webb (2017) study how best to approximate the distribution of t^{stat} under cluster heterogeneity. In their study, cluster sizes vary to reflect the relative populations of the 50 US states. In this setting, t^{stat} is not well approximated by a normal distribution (the findings are based on a Student-t with $G-1$ degrees-of-freedom with $G = 50$ or 100, which is very close to a normal distribution).

To obtain more accurate inference, they study two alternative methods. The first is to use the critical values from a $Student - t(G^*)$. In their simulation setting, the true correlation matrix that enters G^* is a constant correlation matrix with all elements equal to 0.5. To more closely capture this reduced level of correlation, they replace the unit correlation matrix with a constant correlation matrix in which the correlation is estimated from the data. While the estimation introduces a pre-test bias, in this setting the use of the estimated correlation leads to an increase in the computed value of G^*, which in turn leads to rejection rates closer to the nominal rate of 5%.

The second is to use a form of the bootstrap to compute a P-value. To capture the within-cluster correlation structure the bootstrap resamples data by cluster, rather than by individual.

Randomness in the bootstrap is not obtained by sampling randomly with replacement from the set of clusters, but rather by introducing a scalar random variable for each cluster v_i, which is equally likely to be -1 or 1. The bootstrap sample is then constructed from $\{v_1\widehat{\varepsilon}_1, \ldots, v_G\widehat{\varepsilon}_G\}$ and is termed a wild cluster bootstrap. The P-value is then obtained as the proportion of bootstrap t-statistics that exceed, in magnitude, the t-statistic from estimation of the original data. In virtually all simulation settings, the rejection rate is remarkably close to the nominal rate of 5%.

Because the wild cluster bootstrap is computationally intensive, it would be helpful to know when the critical values from a normal distribution do provide rejection rates that are close to the nominal rate. CSS show that G^* is the key measure to answer this question and, that when G^* exceeds 50, the critical values from a normal distribution yield accurate inference.

The guide for testing in two-way fixed effects models, where the clusters have arbitrary correlation, can be conveniently summarized. First, compute the effective number of clusters for the coefficient under test. For example, Lee and Steigerwald (2018) provide a Stata command to do just this. Then, if the effective number of clusters exceeds 50, conduct standard inference in which the critical values from a normal distribution (typically ±1.96) are used. If the effective number of clusters is less than 50, use the wild cluster bootstrap to form the P-value.

One final note. If the data arise from a randomized controlled trial, so that (1) is replaced by

$$Y_{it} = \alpha + \beta D_{it} + \epsilon_{it},$$

then the cluster-robust variance estimator can be decomposed into two parts: the variance estimated from the treatment clusters and the variance estimated from the control clusters. Because there are no coefficients that are estimated across both the treatment and control groups, the computation of the effective number of clusters in (12) is done separately for the treatment and control groups and the effective number of clusters for estimation of β is then a weighted average of the effective number of clusters for the treatment G_1^* and control G_0^* groups.

5 Empirical Illustration

To illustrate how these results on estimation and inference can be incorporated in practice, we turn to the analysis by Davis and Wolfram (2012), who measure the effect of a law encouraging the sale of power generation plants by public utilities on the emission of greenhouse gases. The authors find that divestment - the sale of nuclear reactors by public utilities to independent (non-utility) power producers - led to a large gain in operating efficiency at these reactors (on the order of 10 percentage points) and a consequent reduction in the need for electricity from coal and natural gas power plants, leading to a large reduction in greenhouse gas emissions.

Within the analysis, observations from a nuclear reactor that is owned by independent power producers are in the treatment state, while those from reactors owned by public utilities are in the untreated state. Since switching from public to private ownership is likely to be correlated with both observable and unobservable firm characteristics, a simple comparison between privately- and publicly-owned firms would yield a biased and inconsistent estimator of the effect of interest. For this reason, the identification strategy is to compare the difference over time in operating efficiency for the reactors that changed ownership, with the corresponding difference for reactors that remained under the control of a public utility. This strategy fits perfectly within the context of Section 3, as the number of treated firms varies over time. The inclusion of firm-level and time fixed effects can mitigate the endogeneity of the treatment by controlling for (i) all observable and unobservable firm-level characteristics that are fixed over time, and (ii) secular trends and other time-varying factors that affect all firms equally.

To provide a measure of these treatment effects, Davis and Wolfram (2012) estimate Equation (1) with a forty-year monthly panel of all nuclear reactors in the United States. For illustration purposes, we reproduce this result after collapsing the data at the yearly level, and using data only after 1990 for reactors with complete reporting and no long outages. These data modifications generate a balanced panel of reactors that operate under normal load for the duration of the panel, consistent with the additional sensitivity tests performed by Davis and Wolfram. We estimate the same models as reported in Table 1 in Davis and Wolfram and display the results in Table 1. The two way fixed effects model is estimated in column (2), revealing that that a change in ownership structure is associated with an 8 percentage point improvement in reactor operating efficiency. This result is comparable to the estimated 10 percentage point improvement reported in Davis and Wolfram.

Table 1: Replication results

	(1)	(2)	(3)	(4)
D_{it}	5.59***	8.26***	7.88***	7.91***
	(1.34)	(2.64)	(2.31)	(2.24)
year-of-sample fixed effects (20 years)	Yes	Yes	Yes	Yes
Reactor fixed effects (89 reactors)	No	Yes	Yes	Yes
Reactor Age (cubic)	No	No	Yes	Yes
Other FE, Weights or Covariates	No	No	No	Yes
Number of cross sectional units	89	89	89	89
Number of clusters	54	54	54	54
Effective number of clusters	38.83	38.83	38.83	38.83
Observations	1,780	1,780	1,780	1,780
R^2	0.29	0.38	0.39	0.39

Notes: replication results of Davis and Wolfram Table 1, using modified data. The panel from Davis and Wolfram is collapsed at the yearly level, data prior to 1990 is dropped, and reactors with incomplete reporting or long outages are dropped. Results account for clustering at the plant level.

The authors demonstrate the importance of including firm-level and time fixed effects by estimating a restricted version of (1), in which the reactor fixed effects are omitted. These estimates are reproduced using our subsample in column (1) of Table 1. Omitting reactor fixed effects leaves out observed and unobserved factors associated with treatment adoption and substantially changes the estimate. They further assess the robustness of their results by including increasingly rich sets of explanatory variables (two of these sets are reproduced using our subsample in column (3) of Table 1 and finding that the estimated coefficient on the treatment indicator is largely unaffected.

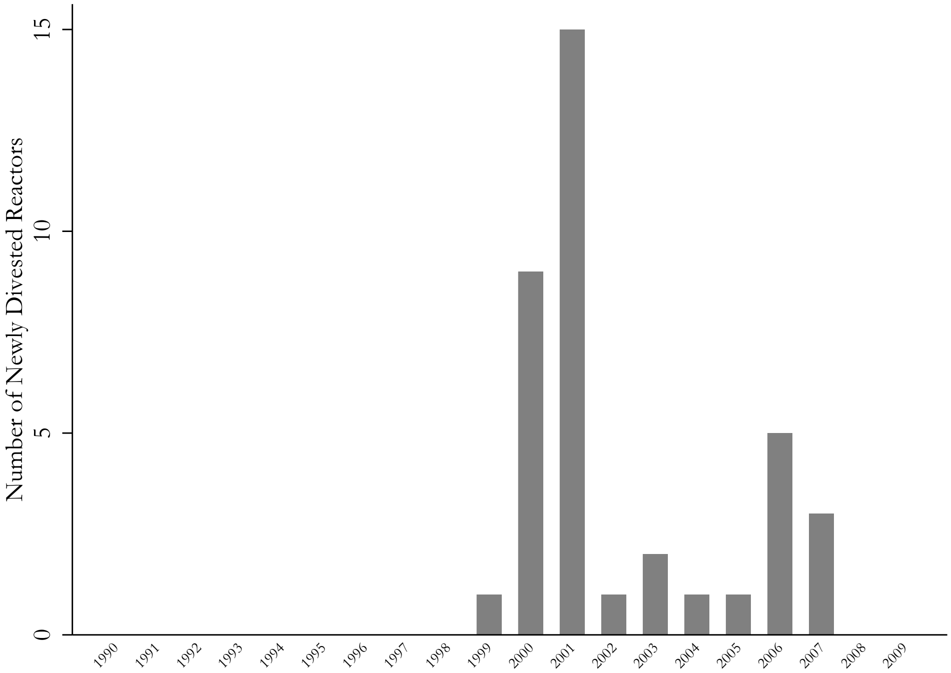
Within the context of the two-way fixed-effects estimating equation, there is a stable relationship between ownership change and operating efficiency. To assess the significance of the relation Davis and Wolfram construct cluster-robust standard errors. Because some nuclear plants contain more than one reactor, the standard errors are clustered at the plant level to account for unobserved factors that can affect multiple reactors at the same plant. In this setting, clusters are heterogeneous for several reasons: the number of reactors varies from plant to plant, the values of the additional explanatory variables (such as the age, type, or manufacturer of the reactor) vary over plants, and the correlations in the unobserved factors are likely unique to each plant.

We use the `clusteff` Stata command (Lee and Steigerwald 2018) to compute the effective number of clusters for test of significance of the treatment coefficient. We report the results for the baseline specification in column (2) of Table 1. There are 54 plants in the sample and the computed effective number of clusters is 38.83. The effective number of clusters may not be large enough to justify approximating the t-statistic by a normal distribution. As such, we check the robustness of the inference results by calculating wild-bootstrap p-values and 95% confidence intervals. The resultant p-value is ¡0.001 and the 95% confidence interval is [2.8,13.83] again allowing us to reject the null hypothesis of no effect. When cluster-robust standard errors are used the p-value is 0.003 and the 95% confidence interval is [2.97,13.56].

5.1 Average Treatment Effects

To interpret the estimated divestment coefficient as an average treatment effect, we first need to be clear about how the treatment effect differs over observations. Figure 3 contains the number of newly treated reactors in each year of the panel.

Figure 3: Newly treated reactors over time



As Figure 3 reveals, treatment assignment follows a staggered adoption design. With this staggered adoption of treatment, there are a number of treated groups (plants treated in the first year, plants treated in the second year, and so on), each with their own average treatment effect. Further, the average treatment effect in the year of entry may differ from the average treatment effect for the same group of reactors in the second year of treatment. In the two-way fixed-effects equation we have only one treatment coefficient to capture all of these effects, which is the source of the misspecification of the conditional mean.

With a staggered adoption design, the two-way fixed effects estimator is a linear combination of all the individual treatment effects. Proposition 3 shows that the weights can be computed simply from the fraction of treated units in each period and the timing of entry into treatment. We compute the weights test for the baseline specification from Table 1. The summary of the computed weights is contained in Table 2.

Table 2: Summary of weights

Number of weights	63
Mean	0.0159
Median	0.0069
Max	0.0579
Min	0.0017

Because the linear combination generated from these weights is not a weighted average of the period average treatment effects, it is not possible to interpret the two-way fixed effects estimator as an average treatment effect. Yet the presence of weights that are all positive does rule out one pathological case. If there were negative weights it would be possible to get a negative estimate of the coefficient even if all of the underlying treatment effects were positive. With all positive weights, this is not possible.

As shown in Corollary 2, the 2WFE estimator recovers an average of the ATTs over time when treatment adoption is simultaneous. Based on this fact, when treatment adoption is staggered, one can always estimate the treatment effects for each treated group separately, essentially recasting a staggered adoption design as multiple simultaneous adoption designs (with a common untreated group).

To illustrate how this idea can be used in practice, we restrict our sample to the groups entering treatment in 2000, the group entering treatment in 2001 and the untreated group. The results from the 2WFE estimation are shown in Table 3. While both estimates are positive, consistent with the results in Table 1, they suggest some heterogeneity across treated groups: the estimated average effects are 5.25 for the group treated in 2000 and 12.87 for the group treated in 2001. Two main factors can explain these differences. First, these estimates average over each group's treated period (see Corollary 2), and thus the average effect for the group treated in period 2000 includes one more treated period. On the other hand, differences in effects can be due to the fact that groups entering treatment in different periods can be heterogeneous and differ in both observed and unobserved characteristics. Additionally, the difference in coefficients could be explained by sampling variability. We further explore these differences in the upcoming sections.

Table 3: Estimation results - 2WFE

	Treated in 2000	Treated in 2001
D_{it}	5.25 (4.67)	12.87*** (3.72)
G_A^*	8.56	12.68
Bootstrap 95% CI	[-4.66,17.7]	[5.07, 21.3]
year-of-sample fixed effects	Yes	Yes
Reactor fixed effects	Yes	Yes
Number of cross-sectional units	60	66
Observations	1200	1320

Notes: estimation results from the two-way fixed effects model for the two group multiple period case, for the group treated in 2000 (column 1), and the group treated in 2001 (column 2). Results account for clustering at the plant level.

To assess significance, we again cluster at the plant level. The effective number of clusters for the groups treated in 2000 and 2001 are, respectively, 8.56 and 12.68, which strongly suggests that the normal approximation may not be accurate. Hence, we provide bootstrap confidence intervals for inference. Our inference procedure rejects the null of no average effect for the group treated in 2001, but fails to reject it for the group treated in 2000. Notice that this latter result may be due to low statistical power given the small number of units.

An alternative estimation method consists in running a 2WFE regression including an interaction term as follows:

$$Y_{it} = \alpha_i + \delta_t + \beta_{2001} D_{it} + \beta_{\mathsf{dif}} D_{it} \mathbb{1}(t_i^* = 2000) + \varepsilon_{it} \tag{14}$$

In Equation (14), β_{2001} estimates the average effect for the group treated in 2001, while β_{dif} estimates the difference between the effects for the groups treated in 2001 and 2000 (which implies that the effect for the group treated in 2000 is $\beta_{2001} + \beta_{\mathsf{dif}}$). This estimation method is a variation of the ones proposed by Gibbons et al. (2018) and Abraham and Sun (2020). The results, presented in Table 4, are very similar to the ones in Table 3, but also provide a straightforward way to evaluate the difference in treatment effects across groups by simply evaluating whether $\beta_{\mathsf{dif}} = 0$. The bootstrap confidence interval for this coefficient is $[-18.1, 3.75]$, which indicates that the effects for the two groups are not significantly different (although again this may be due to low statistical power given the small number of treated units).

Table 4: Estimation results - interacted model

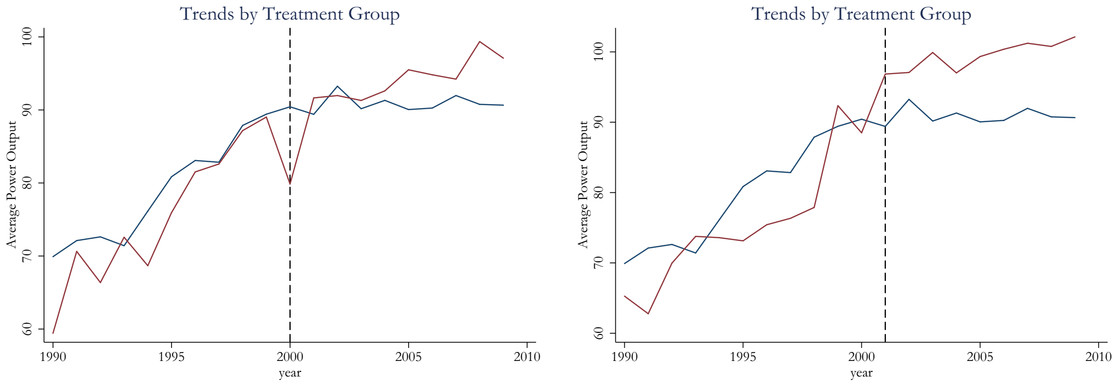
	(2)
D_{it}	12.46*** (3.62)
$D_{it}\mathbb{1}(t_i^* = 2000)$	-7.14 (4.77)
Bootstrapped 95% CI for β_{dif}	[-18.1, 3.76]
year-of-sample fixed effects (20 years)	Yes
Reactor fixed effects	Yes
Number of cross sectional units	75
Observations	1500

Notes: estimation results from the interacted model in Equation (14). Results account for clustering at the plant level.

5.2 Pre-Treatment Trends and Dynamic Effects

The possibility of interpreting the estimated results as (weighted averages of) causal effects depends crucially on the plausibility of the parallel trends assumption. While this assumption is not directly testable, some evidence can be provided to support its validity, as explained in Section 3.4. Figure 4 provides a visual inspection of the pre-treatment trends for treated and untreated firms for the group treated in 2000 (left panel) and the group treated in 2001 (right panel). Overall, the estimated trends seem to follow a similar pattern, but a visual inspection is not a formal test. We next describe a dynamic specification that does lead to a formal test of equality of pre-treatment trends.

Figure 4: Evolution of the outcome over time



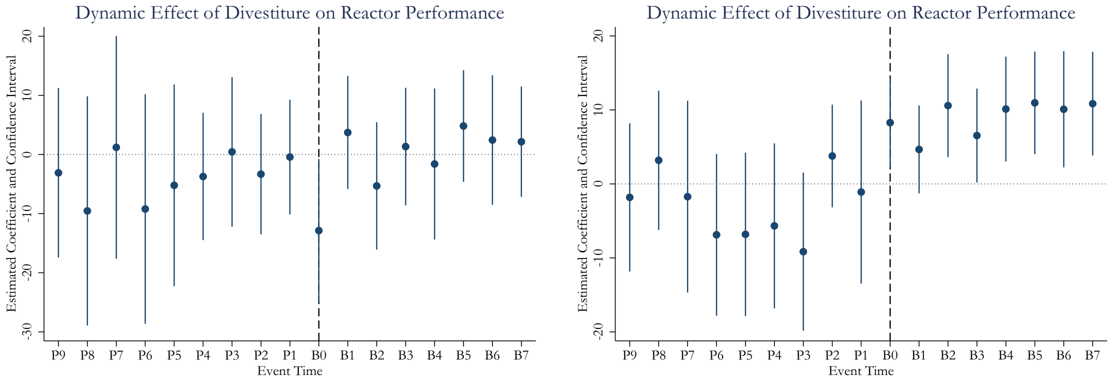
(a) Treated in 2000 versus untreated

(b) Treated in 2001 versus untreated

Notes: outcome trends for the treatment group (red) and the untreated group (blue). The left panel considers the treatment group to be the group treated in 2000, while the right panel considers the treatment group to be the group treated in 2001.

To provide a more formal analysis of the parallel trends assumption, as well as give direct estimates of the dynamic treatment effects, we now estimate Equation (7). The results are shown in Figure 5, with the left panel showing the results for the group treated in 2000 and the right panel showing the results for the group treated in 2001.

Figure 5: Dynamic effects



(a) Effects for units treated in 2000

(b) Effects for units treated in 2001

Notes: results of the dynamic treatment effect estimation for the two group multiple period case, with bootstrapped 95% confidence intervals. 'P' represents a placebo test, where the number corresponds to the number of years before treatment. 'B' represents the beta of treatment, where the number corresponds to the number of years after treatment. The left panel considers the treatment group to be the group treated in 2000, while the right panel considers the treatment group to be the group treated in 2001. Results account for clustering at the plant level.

In Figure 5 we are interested in both the pre-treatment effects (placebo tests) and the dynamic treatment effects. The figure illustrates that none of the placebo tests are statistically different from zero on their own for either the group treated in 2000 (panel a) or the group treated in 2001 (panel b) – providing supportive evidence of the common trends assumption for both groups. However, recent work raises issues about using the estimated coefficients to form valid placebo tests of the parallel trends assumption (see Abraham and Sun 2020). We discuss these issues in Section 6.

Figure 5 also shows that treatment effects over time are mostly positive and are statistically different from zero for the group treated in 2001 (panel b), and overall suggests that the effects are relatively stable over time, at least for this treated group during the period under study. However, for the group treated in 2000 (panel a), the treatment effects are not statistically different from zero.

6 Discussion and Some Practical Recommendations

6.1 Estimation

One takeaway from the discussion in this paper is that when all treated units enter treatment simultaneously, the two-way fixed effects estimator recovers a simple average of the average treatment effects on the treated over time, which has a clear causal interpretation as long as the parallel trends assumption holds. In addition, it is easy in this setting to separately estimate the ATTs at different time periods through a dynamic specification or event-study design to explore how the average effect of the policy varies over time.

On the other hand, when there is variation in treatment timing (staggered treatments), the two-way fixed effects estimator may not have a clear causal interpretation. In such cases, the researcher should favor more flexible specifications that do not pool all the data into a single regression. For instance, it is straightforward in this case to estimate average effects separately for different treatment groups, as we illustrate in our empirical application. With these estimates, the researcher can explore the heterogeneity of treatment effects across groups, or pool these effects into a single summary measure using a weighting scheme that is deemed reasonable, as suggested by Callaway and Sant'Anna (2019) and Abraham and Sun (2020). Another possibility is to focus on alternative estimands like the ones proposed by de Chaisemartin and D'Haultfœuille (forthcoming).

6.2 Inference

The key takeaway regarding inference is to carefully consider the sample size on which the estimator is based. The distribution of the cluster-robust test statistic is governed by the effective number of clusters, which can be no bigger than the total number of clusters in the sample. If the effective number of clusters is larger than 20, then a normal approximation is appropriate for the test statistic. If the effective number of clusters is smaller, then a wild-cluster bootstrap should be used to approximate the distribution of the test statistic. Hence, good practice is to report the effective number of clusters and to use it to justify either the normal or bootstrap approximations.

6.3 Assessing the Parallel Trends Assumption

The causal interpretation of the estimands and estimators discussed in this paper relies crucially on the parallel trends assumption, according to which treated and control units would have experienced the same average trend in outcomes had the policy not been implemented. The parallel trends assumption has become the focus of several recent an ongoing studies. While an in-depth discussion of this rapidly growing literature is beyond the scope of our paper, here we briefly discuss several important issues related to this assumption and how to assess it. We refer to reader to the original papers (and references therein) for further details.

6.3.1 Interpretation of the Pre-Treatment Trends Test

The parallel trends assumption is untestable because it involves counterfactual magnitudes, and researchers usually focus on testing whether trends were parallel before the policy is implemented. While finding parallel trends before the policy may, under some conditions, be interpreted as evidence to support the identification assumption, parallel pre-treatment trends are neither sufficient nor necessary for parallel trends after the policy. Therefore, failing to find statistically significant differences in pre-treatment trends does not guarantee that the identification assumption holds. Kahn-Lang and Lang (2019) argue that in such contexts, the researcher needs to justify why this pattern can be expected to continue after the policy, which also requires understanding why the groups differ in levels to begin with. These authors also stress that parallel trends tests may be sensitive to the choice of the time period, as analyzing long- and short-run trends often lead to different conclusions. On the other hand, Laporte and Windmeijer (2005) note that the pre-trends tests can fail when treatments have anticipatory effects.

Finally, as mentioned previously, Abraham and Sun (2020) find that in the dynamic specification (7), the lead coefficients commonly used to test for parallel trends may partially pick up future treatment effects yielding significant results even when the pre-treatment trends are indeed parallel. For these reasons, the results form the pre-treatment trends test should always be interpreted with caution.

6.3.2 Testing Issues

Several studies have noted that pre-trends tests may often have limited statistical power to detect plausible differences in trends before the treatment (Bilinski and Hatfield 2018; Freyaldenhoven et al. 2019; Kahn-Lang and Lang 2019; Roth 2019). Hence, failing to reject the hypothesis that pre-treatment trends are parallel is not the same as confirming that they are. Furthermore, Bilinski and Hatfield (2018) note that testing the null hypothesis that there is no difference in trends before the treatment inverts the role of type I and type II errors, as traditional testing methods guard against incorrectly concluding that an effect is non-zero (i.e. incorrectly rejecting the null of no effect). This implies that, if a parallel trends test has power of, say, 80 percent against a certain alternative, the probability of failing to detect this difference in trends when it actually exists is 20 percent. As a result, non-negligible differences in pre-treatment trends may remain undetected with high probability, which may lead to biased and inconsistent estimates.

On the other hand, conditioning the estimation of treatment effects on whether the pre-treatment trends test rejects or not can result in invalid inference due to pre-testing, a well known fact in statistics. Roth (2019) shows that the bias of estimators and the coverage rates of confidence intervals conditional on passing a pre-treatment trends test can be worse that the unconditional ones in relevant cases.

6.3.3 Robust Methods

Given the problems associated with pre-testing discussed above, empirical researchers may want to consider alternative methods that are robust to some violations of the parallel trends assumption. For example, Abadie (2005), Callaway and Sant'Anna (2019) and Sant'Anna and Zhao (2020) consider cases in which the parallel trends assumption holds after conditioning on a set of observed covariates and propose semiparametric inverse-probability-weighting estimators. Freyaldenhoven et al. (2019) exploit the availability of a covariate that is unaffected by the treatment but related to unobserved confounders. This covariate can be used as a measure of how much the estimated effect is due to the actual treatment effect, and how much is due to difference in outcome trends. Based on this covariate, they propose a 2SLS estimator that is robust to pre-event differential trends generated by this endogeneity in a linear setting with homogeneous effects.

A related literature considers partial identification of average effects when the trends are allowed to differ in specific ways. Manski and Pepper (2018) analyze the sensitivity of the identified set to different assumptions that allow for nonparallel trends, restricting the magnitude of the difference in trends without requiring it to be zero. Rambachan and Roth (2019) propose a method that allows the researcher to restrict the way in which the assumption can be violated incorporating information on pre-treatment trends. Based on this, they provide partial identification and uniformly valid inference results for average treatment effects.

A key message from this section is that while parallel pre-treatment trends tests may be useful in assessing the validity of the identification assumption, they should not be taken as an "all-or-nothing" rule that unequivocally determines the credibility of a study. The validity of the parallel trends assumption needs to be assessed and justified on a case by case basis. This identification assumption is fundamentally untestable and hence there is no statistical procedure that can ensure or rule out its validity. For this reason, whether the parallel-trends assumption holds cannot be determined solely on statistical grounds, and empirical evidence in its support needs to be complemented with institutional knowledge and economic theory. On the other hand, the recent methods discussed above that are robust to some violations of the parallel-trends assumption are a promising avenue to reinforce the credibility of studies based on panel data.

7 Conclusion

The two-way fixed effects specification with panel data is a common framework in which to assess the effectiveness of programs or policies with environmental targets. These programs can have effects that may be unique to each unit subject to the program. If the effects are heterogeneous, then the estimated coefficient from the two-way fixed effects equation may not accurately reflect these effects and could show a negative effect of the program when each of the individual effects are positive. If the treatment occurs in only a single period, then this problem does not arise and the estimated coefficient captures an average of the individual treatment effects.

Inference on this parameter is most often done with cluster-robust standard errors, which allow for arbitrary correlations over time for each unit in the panel. It is common to base the approximate normality of the test statistic on the number of units but this requires that the units have similar error covariance matrices. As the point of cluster-robust inference is to allow for dissimilarity in these matrices, it is highly likely that the units are dissimilar. We show that when the units are dissimilar, the number of units must be adjusted downward to reflect this dissimilarity. If this adjusted number is large (in practice, above 50), then standard inference with critical values from a normal distribution is appropriate. If the adjusted number is smaller, then inference should be done with a wild-cluster bootstrap.

References

Abadie, Alberto. 2005. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72 (1): 1–19.

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2020. Sampling-Based versus design-based uncertainty in regression analysis. *Econometrica* 88 (1): 265–296.

Abraham, Sarah, and Liyang Sun. 2020. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Unpublished manuscript, Department of Economics, MIT.

Athey, Susan, and Guido Imbens. 2018. Design-based analysis in difference-in-differences settings with staggered adoption. NBER Working Paper 24963, National Bureau of Economic Research, Cambridge MA.

Auffhammer, Maximilian, Antonio M. Bento, and Scott E. Lowe. 2009. Measuring the effects of the Clean Air Act Amendments on ambient PM10 concentrations: The critical importance of a spatially disaggregated analysis. *Journal of Environmental Economics and Management* 58 (1): 15 –26.

Auffhammer, Maximilian, and Ryan Kellogg. 2011. Clearing the air? The effects of gasoline content regulation on air quality. *American Economic Review* 101 (6): 2687–2722.

Bento, Antonio, Matthew Freedman, and Corey Lang. 2015. Who benefits from environmental regulation? Evidence from the Clean Air Act Amendments. *The Review of Economics and Statistics* 97 (3): 610–622.

Bilinski, Alyssa, and Laura A. Hatfield. 2018. Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions. *arXiv:1805.03273.*

Borusyak, Kirill, and Xavier Jaravel. 2017. Revisiting event study designs, with an application to the estimation of the marginal propensity to consume. *https://ssrn.com/abstract=2826228.*

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2019. Difference-in-differences with multiple time periods. *https://ssrn.com/abstract=3148250.*

Carter, Andrew V, Kevin T Schnepel, and Douglas G Steigerwald. 2017. Asymptotic behavior of at-test robust to cluster heterogeneity. *Review of Economics and Statistics* 99 (4): 698–709.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. Average and quantile effects in nonseparable panel models. *Econometrica* 81 (2): 535–580.

Cochran, William G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics* 17 (2): 164–177.

Currie, Janet, Lucas Davis, Michael Greenstone, and Reed Walker. 2015. Environmental health risks and housing values: Evidence from 1,600 toxic plant openings and closings. *American Economic Review* 105 (2): 678–709.

Davis, Lucas W, and Catherine Wolfram. 2012. Deregulation, consolidation, and efficiency: Evidence from US nuclear power. *American Economic Journal: Applied Economics* 4 (4): 194–225.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. Forthcoming. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review.*

Deschenes, Olivier, and Kyle C. Meng. 2018. Quasi-experimental methods in environmental economics: Opportunities and challenges. In *Handbook of Environmental Economics,* edited by Partha Dasgupta, Subhrendu K. Pattanayak, and V. Kerry Smith. Amsterdam: Elsevier.

Ferraro, Paul J., and Juan José Miranda. 2013. Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment. *Resource and Energy Economics* 35 (3): 356 –379.

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.

Frondel, Manuel, and Colin Vance. 2013. Heterogeneity in the effect of home energy audits: Theory and evidence. *Environmental and Resource Economics* 55, no. 3 (July): 407–418.

Gibbons, Charles E., Juan Carlos Suárez-Serrato, and Michael B. Urbancic. 2018. Broken or fixed effects? *Journal of Econometric Methods* 8 (1): 20170002.

Goodman-Bacon, Andrew. 2018. Difference-in-differences with variation in treatment timing. NBER Working Paper 25018, National Bureau of Economic Research, Cambridge MA.

Grainger, Corbett A. 2012. The distributional effects of pollution regulations: Do renters fully pay for cleaner air? *Journal of Public Economics* 96 (9): 840 –852.

Imai, Kosuke, and In Song Kim. Forthcoming. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis.*

———. 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science* 63 (2): 467–490.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Kahn-Lang, Ariella, and Kevin Lang. 2019. The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics* 38 (3): 613–320.

Laporte, Audrey, and Frank Windmeijer. 2005. Estimation of panel data models with binary indicators when treatment effects are not constant over time. *Economics Letters* 88 (3): 389 –396.

Lee, Chang Hyung, and Douglas G Steigerwald. 2018. Inference for clustered data. *The Stata Journal* 18 (2): 447–460.

MacKinnon, James G., and Matthew D. Webb. 2017. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics* 32 (2): 233–254.

Manski, Charles F., and John V. Pepper. 2018. How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics* 100 (2): 232–244.

Martin, Ralf, Laure B. de Preux, and Ulrich J. Wagner. 2014. The impact of a carbon tax on manufacturing: Evidence from microdata. *Journal of Public Economics* 117 (2014): 1 –14.

Rambachan, Ashesh, and Jonathan Roth. 2019. An honest approach to parallel trends. Unpublished manuscript, Department of Economics, Harvard University.

Roth, Jonathan. 2019. Pre-test with caution: Event-study estimates after testing for parallel trends. Unpublished manuscript, Department of Economics, Harvard University.

Sant'Anna, Pedro H.C., and Jun B. Zhao. 2020. Doubly robust difference-in-differences estimators. *arXiv:1812.01723.*

Sills, Erin O., and Kelly Jones. 2018. Causal inference in environmental conservation: The role of institutions. In *Handbook of Environmental Economics,* edited by Partha Dasgupta, Subhrendu K. Pattanayak, and V. Kerry Smith. Amsterdam: Elsevier.

Słoczyński, Tymon. 2018. A general weighted average representation of the ordinary and two-stage least squares estimands. IZA Working Paper 11866, IZA Institute of Labor Economics, Bonn, Germany.

Vazquez-Bare, Gonzalo. 2017. Identification and estimation of spillover effects in randomized experiments. *arXiv:1711.02745.*

Wooldridge, Jeffrey M. 2005. Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *The Review of Economics and Statistics* 87 (2): 385–390.

Appendix

A Proof of Proposition 1

Because units with $t_i^* = s$ are untreated in period t and treated in period t', whereas units with $t_i^* = u$ remain untreated in both periods,

$$\begin{aligned}
\mathbb{E}[Y_{it'} - Y_{it}|t_i^* = s] - \mathbb{E}[Y_{it'} - Y_{it}|t_i^* = u] &= \mathbb{E}[Y_{it'}(1) - Y_{it}(0)|t_i^* = s] - \mathbb{E}[Y_{it'}(0) - Y_{it}(0)|t_i^* = u] \\
&= \mathbb{E}[Y_{it'}(1) - Y_{it'}(0)|t_i^* = s] \\
&\quad + \mathbb{E}[Y_{it'}(0) - Y_{it}(0)|t_i^* = s] - \mathbb{E}[Y_{it'}(0) - Y_{it}(0)|t_i^* = u] \\
&= \mathbb{E}[Y_{it'}(1) - Y_{it'}(0)|t_i^* = s] \\
&= \mathbb{E}[\tau_{it'}|t_i^* = s],
\end{aligned}$$

where the third equality follows from the parallel-trends assumption. ■

B Proof of Proposition 2

To analyze the population parameter β, define the population analogs of the double-demeaned variables $\ddot{D}_{it}$ and $\ddot{Y}_{it}$ as:

$$\begin{aligned}
\widetilde{D}_{it} &= D_{it} - \overline{D}_i - \pi_t + \mathbb{E}[\overline{D}_i] \\
\widetilde{Y}_{it} &= Y_{it} - \overline{Y}_i - \mathbb{E}[Y_{it}] + \mathbb{E}[\overline{Y}_i]
\end{aligned}$$

which replace the sample averages over i by their corresponding population expectations (recall that T is fixed in our setting). By construction, $\mathbb{E}[\widetilde{D}_{it}] = \sum_{t=0}^{T} \widetilde{D}_{it} = 0$, which can be verified by direct calculation. Based on these variables, it is possible to eliminate the individual and time fixed effects in Equation (1) by rewriting it as:

$$\widetilde{Y}_{it} = \beta \widetilde{D}_{it} + \widetilde{\varepsilon}_{it}, \quad t = 0, \ldots, T.$$

which is the population version of the double-demeaned model. By definition, the linear projection coefficient β in this equation is given by:

$$\beta = \frac{\sum_{t=0}^{T} \mathbb{E}[\widetilde{Y}_{it}\widetilde{D}_{it}]}{\sum_{t=0}^{T} \mathbb{E}[\widetilde{D}_{it}\widetilde{D}_{it}]} = \frac{\sum_{t=0}^{T} \mathbb{E}[Y_{it}\widetilde{D}_{it}]}{\sum_{t=0}^{T} \mathbb{E}[D_{it}\widetilde{D}_{it}]}$$

where the second equality uses that $\mathbb{E}[\widetilde{D}_{it}] = \sum_{t=0}^{T} \widetilde{D}_{it} = 0$. Plugging in the potential outcomes,

$$\beta = \frac{\sum_{t=0}^{T} \mathbb{E}[Y_{it}\widetilde{D}_{it}]}{\sum_{t=0}^{T} \mathbb{E}[D_{it}\widetilde{D}_{it}]} = \frac{\sum_{t=0}^{T} \mathbb{E}[(Y_{it}(0) - \overline{Y}_i(0))\widetilde{D}_{it}]}{\sum_{t=0}^{T} \mathbb{E}[D_{it}\widetilde{D}_{it}]} + \frac{\sum_{t=1}^{T} \mathbb{E}[\tau_{it}D_{it}\widetilde{D}_{it}]}{\sum_{t=1}^{T} \mathbb{E}[D_{it}\widetilde{D}_{it}]},$$

where we used the facts that $\sum_{t=0}^{T} \mathbb{E}[\overline{Y}_i(0)\widetilde{D}_{it}] = \mathbb{E}[\overline{Y}_i(0)\sum_{t=0}^{T} \widetilde{D}_{it}] = 0$ for the first term and that $D_{i0} = 0$ for the second term. For the first term, use the fact that

$$Y_{it}(0) - \overline{Y}_i(0) = \sum_{s=1}^{t} \frac{s}{T+1}\Delta_{is}(0) - \sum_{s=t+1}^{T} \left(1 - \frac{s}{T+1}\right)\Delta_{is}(0)$$

where $\Delta_{is}(0) = Y_{is}(0) - Y_{is-1}(0)$, and thus the parallel-trends assumption implies that

$$\mathbb{E}[Y_{it}(0) - \overline{Y}_i(0)|\mathbf{D}_i] = \mathbb{E}[Y_{it}(0) - \overline{Y}_i(0)]$$

so the first term equals zero using again that $\mathbb{E}[\widetilde{D}_{it}] = 0$. Next, consider the term $\mathbb{E}[\tau_{it} D_{it} \widetilde{D}_{it}]$. Recall that $D_{it} = \mathbb{1}(t_i^* \leq t)$. We have that

$$\begin{aligned}
\mathbb{E}[\tau_{it} D_{it} \widetilde{D}_{it}] &= \mathbb{E}\{\mathbb{E}[\tau_{it}|t_i^*] D_{it} \widetilde{D}_{it}\} = \mathbb{E}\left\{\sum_{s=1}^{T+1} \mathbb{E}[\tau_{it}|t_i^* = s] \mathbb{1}(t_i^* = s) D_{it} \widetilde{D}_{it}\right\} \\
&= \mathbb{E}\left\{\sum_{s=1}^{T+1} \mathbb{E}[\tau_{it}|t_i^* = s] \mathbb{1}(t_i^* = s) \mathbb{1}(t_i^* \leq t) \widetilde{D}_{it}\right\} = \mathbb{E}\left\{\sum_{s=1}^{t} \mathbb{E}[\tau_{it}|t_i^* = s] \mathbb{1}(t_i^* = s) \mathbb{1}(t_i^* \leq t) \widetilde{D}_{it}\right\} \\
&= \sum_{s=1}^{t} \mathbb{E}[\tau_{it}|t_i^* = s] \mathbb{E}\left\{\mathbb{1}(t_i^* = s) D_{it} \widetilde{D}_{it}\right\}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\overline{D}_i &= \frac{1}{T+1} \sum_{t=0}^{T} D_{it} = \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{1}(t_i^* \leq t) = 1 - \frac{t_i^*}{T+1} \\
\mathbb{E}[t_i^*] &= \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*] + (T+1)(1 - \mathbb{E}[D_i^*]) \\
\mathbb{E}[\overline{D}_i] &= \mathbb{E}[D_i^*] - \frac{\mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1} \\
D_{it} \widetilde{D}_{it} &= D_{it} \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{t_i^* - \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1}\right) \\
\mathbb{E}[D_{it} \widetilde{D}_{it}] &= \mathbb{E}\left[\mathbb{1}(t_i^* \leq t) \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{t_i^* - \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1}\right)\right] \\
&= \sum_{s=1}^{t} \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{s - \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1}\right) \mathbb{P}[t_i^* = s] \\
\mathbb{1}(t_i^* = s) D_{it} \widetilde{D}_{it} &= \mathbb{1}(t_i^* = s) \mathbb{1}(s \leq t) \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{s - \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1}\right) \\
\mathbb{E}[\mathbb{1}(t_i^* = s) D_{it} \widetilde{D}_{it}] &= \mathbb{P}[t_i^* = s] \mathbb{1}(s \leq t) \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{s - \mathbb{E}[t_i^*|D_i^* = 1]\mathbb{E}[D_i^*]}{T+1}\right).
\end{aligned}$$

Collecting all the results,

$$\beta = \frac{\sum_{t=1}^{T} \sum_{s=1}^{t} \mathbb{E}[\tau_{it}|t_i^* = s] \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{s - \mathbb{E}[t_i^*|D_i^*=1]\mathbb{E}[D_i^*]}{T+1}\right) \mathbb{P}[t_i^* = s]}{\sum_{t=1}^{T} \sum_{s=1}^{t} \left(\mathbb{E}[D_i^*] - \mathbb{E}[D_{it}] + \frac{s - \mathbb{E}[t_i^*|D_i^*=1]\mathbb{E}[D_i^*]}{T+1}\right) \mathbb{P}[t_i^* = s]}$$

as required. ■

C Proof of Proposition 3

$$\widehat{\beta}_{\mathsf{FE}} = \frac{\sum_t \sum_i Y_{it}(0) \ddot{D}_{it}}{\sum_t \sum_i D_{it} \ddot{D}_{it}} + \frac{\sum_t \sum_i \tau_{it} D_{it} \ddot{D}_{it}}{\sum_t \sum_i D_{it} \ddot{D}_{it}} = \frac{\sum_t \sum_i (Y_{it}(0) - \overline{Y}_i(0)) \ddot{D}_{it}}{\sum_t \sum_i D_{it} \ddot{D}_{it}} + \frac{\sum_t \sum_i \tau_{it} D_{it} \ddot{D}_{it}}{\sum_t \sum_i D_{it} \ddot{D}_{it}}$$

where the second equality follows from the fact that $\sum_t \ddot{D}_{it} = 0$. The first term equals zero in expectation by a previous argument. Next note that for $t = 0, \ldots, T$,

$$\begin{aligned}
\overline{D}_i &= \frac{1}{T+1} \sum_{t=0}^{T} D_{it} = \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{1}(t_i^* \leq t) = 1 - \frac{t_i^*}{T+1} \\
\overline{D} &= \frac{1}{N} \sum_{i=1}^{N} \overline{D}_i = 1 - \frac{\overline{t}^* \overline{D}^* + (T+1)(1 - \overline{D}^*)}{T+1} = \overline{D}^* - \frac{\overline{t}^* \overline{D}^*}{T+1}
\end{aligned}$$

where $\overline{D}^* = \frac{1}{G}\sum_i D_i^*$, $\overline{D}_t = \frac{1}{G}\sum_i D_{it}$ and $\overline{t}^* = \sum_i t_i^* D_i^* / \sum_i D_i^*$. Therefore,

$$\ddot{D}_{it} = D_{it} - 1 - \overline{D}_t + \overline{D}^* + \frac{t_i^* - \overline{t}^*\overline{D}^*}{T+1}$$

and

$$\begin{aligned} D_{it}\ddot{D}_{it} &= D_{it}\left(\overline{D}^* - \overline{D}_t + \frac{t_i^* - \overline{t}_1^*\overline{D}^*}{T+1}\right) = \mathbb{1}(t_i^* \leq t)\left(\overline{D}^* - \overline{D}_t + \frac{t_i^* - \overline{t}^*\overline{D}^*}{T+1}\right) \\ &= \sum_{s=1}^{t} \mathbb{1}(t_i^* = s)\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right) \end{aligned}$$

This implies that

$$\begin{aligned} \sum_i \tau_{it} D_{it}\ddot{D}_{it} &= \sum_i \sum_{s=1}^{t} \tau_{it}\mathbb{1}(t_i^* = s)\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right) \\ &= \sum_{s=1}^{t}\left\{\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\sum_i \tau_{it}\mathbb{1}(t_i^* = s)\right\} \\ \sum_i D_{it}\ddot{D}_{it} &= \sum_i \sum_{s=1}^{t} \mathbb{1}(t_i^* = s)\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right) \\ &= \sum_{s=1}^{t}\left\{\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\sum_i \mathbb{1}(t_i^* = s)\right\} \\ &= \sum_{s=1}^{t}\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)N_s^* \end{aligned}$$

where $N_s^* = \sum_i \mathbb{1}(t_i^* = s)$. Taking conditional expectations of the first term,

$$\begin{aligned} \mathbb{E}\left[\sum_i \tau_{it} D_{it}\ddot{D}_{it}\middle|\mathbf{D}\right] &= \sum_{s=1}^{t}\left\{\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\sum_i \mathbb{E}[\tau_{it}|\mathbf{D}_i]\mathbb{1}(t_i^* = s)\right\} \\ &= \sum_{s=1}^{t}\left\{\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\sum_i \mathbb{E}[\tau_{it}|t_i^* = s]\mathbb{1}(t_i^* = s)\right\} \\ &= \sum_{s=1}^{t}\left\{\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\mathbb{E}[\tau_{it}|t_i^* = s]\sum_i \mathbb{1}(t_i^* = s)\right\} \\ &= \sum_{s=1}^{t}\mathbb{E}[\tau_{it}|t_i^* = s]\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)N_s^* \end{aligned}$$

where the second equality uses independence across i and the fact that $\mathbf{D}_i$ is a deterministic function of t_i^* and the third equality uses identical distributions across i. Collecting all the results,

$$\begin{aligned} \mathbb{E}[\widehat{\beta}_{\mathsf{FE}}|\mathbf{D}] &= \frac{\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{G}\tau_{it}D_{it}\ddot{D}_{it}\middle|\mathbf{D}\right]}{\sum_{t=1}^{T}\sum_{i=1}^{G}D_{it}\ddot{D}_{it}} \\ &= \frac{\sum_{t=1}^{T}\sum_{s=1}^{t}\mathbb{E}[\tau_{it}|t_i^* = s]\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\hat{p}_s}{\sum_{t=1}^{T}\sum_{s=1}^{t}\left(\overline{D}^* - \overline{D}_t + \frac{s - \overline{t}^*\overline{D}^*}{T+1}\right)\hat{p}_s} \end{aligned}$$

where $\hat{p}_s = \frac{1}{G}N_s^*$, which gives the required result.

Finally, for the second part, for any variable Z_{it},

$$\frac{1}{G}\sum_i Z_{it}(D_{it} - \overline{D}_i - \overline{D}_t + \overline{D}) = \frac{1}{G}\sum_i Z_{it}(D_{it} - \overline{D}_i) - (\overline{D}_t - \overline{D})\frac{1}{G}\sum_i Z_{it}$$

and

$$\overline{D}_t = \frac{1}{G}\sum_i D_{it} \to_{\mathbb{P}} \mathbb{E}[D_{it}], \quad \overline{D} = \frac{1}{T+1}\sum_t \overline{D}_t \to_{\mathbb{P}} \mathbb{E}[\overline{D}_i].$$

Then, by the law of large numbers, as long as the required moments are bounded, as $G \to \infty$,

$$\frac{1}{G}\sum_i Z_{it}\ddot{D}_{it} = \frac{1}{G}\sum_i Z_{it}\widetilde{D}_{it} + o_{\mathbb{P}}(1) \to_{\mathbb{P}} \mathbb{E}[Z_{it}\widetilde{D}_{it}].$$

It follows that

$$\widehat{\beta}_{\mathsf{FE}} = \frac{\sum_t Y_{it}\widetilde{D}_{it}}{\sum_t D_{it}\widetilde{D}_{it}} + o_{\mathbb{P}}(1) \to_{\mathbb{P}} \frac{\sum_t \mathbb{E}[Y_{it}\widetilde{D}_{it}]}{\sum_t \mathbb{E}[D_{it}\widetilde{D}_{it}]} = \beta$$

which completes the proof. ■

D Proof of Corollary 1

From the proof of Proposition 3,

$$\beta = \frac{\sum_{t=0}^T \mathbb{E}[Y_{it}\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]}.$$

Therefore,

$$\begin{aligned}\beta = \frac{\sum_{t=0}^T \mathbb{E}[Y_{it}\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]} &= \frac{\sum_{t=0}^T \mathbb{E}[Y_{it}(0)\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]} + \frac{\sum_{t=0}^T \mathbb{E}[\tau_{it}D_{it}\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]} \\ = \frac{\sum_{t=0}^T \mathbb{E}[\mathbb{E}[\tau_{it}|D_{it}=1]D_{it}\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]} &= \tau\frac{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]}{\sum_{t=0}^T \mathbb{E}[D_{it}\widetilde{D}_{it}]} = \tau\end{aligned}$$

using that for any t, $\mathbb{E}[\tau_{it}|D_{it} = 1] = \sum_{s=1}^T \mathbb{E}[\tau_{it}|t_i^* = s]\mathbb{P}[t_i^* = s|D_{it} = 1] = \tau$. The remaining results follow by applying Proposition 3 to the case with constant treatment effect. ■

E Proof of Corollary 2

In this case we have that $\overline{D}_t = \overline{D}^*\mathbb{1}(t^* \leq t)$, $\overline{t}_1^* = t^*$ and $\hat{p}_s = \overline{D}^*\mathbb{1}(s = t^*)$ and thus

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_{\mathsf{FE}}|\mathbf{D}] &= \frac{\sum_{t=1}^T\sum_{s=1}^t \mathbb{E}[\tau_{it}|t_i^* = s]\left(\overline{D}^* - \overline{D}_t + \frac{s-\overline{t}^*\overline{D}^*}{T+1}\right)\hat{p}_s}{\sum_{t=1}^T\sum_{s=1}^t \left(\overline{D}^* - \overline{D}_t + \frac{s-\overline{t}^*\overline{D}^*}{T+1}\right)\hat{p}_s} \\ &= \frac{\sum_{t=t^*}^T \mathbb{E}[\tau_{it}|D_i^* = 1]t^*\overline{D}^*(1-\overline{D}^*)/(T+1)}{\sum_{t=t^*}^T t^*\overline{D}^*(1-\overline{D}^*)/(T+1)} \\ &= \frac{\sum_{t=t^*}^T \mathbb{E}[\tau_{it}|D_i^* = 1]}{T+1-t^*}\end{aligned}$$

where we used that $t_i^* = t^* \Leftrightarrow D_i^* = 1$. Similarly, $\mathbb{E}_t[D_{it}] = \mathbb{E}[D_i^*]\mathbb{1}(t^* \leq t)$, $\mathbb{E}[t_i^*|D_i^* = 1] = t^*$ and $\mathbb{P}[t_i^* = s] = \mathbb{E}[D_i^*]\mathbb{1}(s = t^*)$ and the result for the probability limit follows analogously. ■