Measuring the Effect of Environmental Policies using Panel Data

Douglas G. Steigerwald† Gonzalo Vazquez-Bare‡

December 3, 2019

Abstract

Two-way fixed effects models are a popular tool for measuring the effect of environmental policies using panel data. Recent methodological literature has emphasized two potential pitfalls of this technique. First, in the presence of treatment effect heterogeneity, two-way fixed effects estimators may recover a weighted average of treatment effects in which some effects may receive negative weights. Second, the commonly employed cluster-robust standard error estimators may yield unreliable inference when clusters are heterogeneous. We review these recent advances and provide a guide for estimation and inference in two-way fixed effects models.

1 Introduction

In studying natural resources and the environment, the role of targeted programs and policies is often of paramount interest (see Deschenes and Meng, 2018, for a recent survey). 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 secular trends that can be correlated with policy adoption. 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:

$$Y_{it} = \alpha_i + \mu_t + \beta D_{it} + \varepsilon_{it}, \quad t = 0, \ldots, T, \quad i = 1, \ldots, G$$ (1)

where $Y_{it}$ is a measure of air quality in state $i$ at year $t$, $\alpha_i$ is a state-level fixed effect, $\mu_t$ 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 $\mathbb{E}(\varepsilon_{it}|\alpha_i, \mu_t, D_{it}) = 0$, which identifies the coefficients in (1) as parameters of the conditional mean.

When Equation (1) is seen as the true model generating the data, obtaining an unbiased and consistent estimate $\hat{\beta}$ for the effect of the program $\beta$ 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 restriction 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. As another example, firms with differing marginal abatement costs can respond differently to a pollution tax. Thus firm size, industry energy intensity, and trade patterns can lead to different responses to the tax (see e.g. Martin, de Preux, and Wagner, 2014). See also Sills and Jones (2018); Bento, Freedman, and Lang (2015); Ferraro and Miranda (2013); Frondel and Vance (2013); Grainger (2012); Auffhammer, Bento, and Lowe (2009). Importantly, recent methodological literature shows that a specification like (1) can yield misleading estimates in the presence of such treatment effect heterogeneity.

Based on these recent advances, we provide conditions under which the two-way fixed effects estimator does provide useful information on the treatment effects. As we will see, there are two straightforward sets of conditions. First, if the setting has only two periods (and no unit is treated in the first period), then the standard parallel trends assumption is sufficient for the parameters of (1) to provide useful information. Second, in a setting with multiple periods, if all units adopt the policy at the same time, then again the parameters of (1) are guaranteed to provide useful information. On the other hand, variation in the timing of the policy adoption can yield weighted averages that assign negative weights to the effects for certain units in certain periods (de Chaisemartin and D’Haultfœuille, 2019).

In this setting researchers also need to control for error correlations amongst observations for a unit over time through the use of cluster-robust standard errors. It is commonly believed that the cluster-robust $t$ statistic has a normal distribution if the number of units is large. This statement rests on an implicit assumption that the characteristics of the units are similar. If they are not, then even if the number of units is large, the $t$ statistic can be non-normal. We show how to measure the dissimilarity across units 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.

## 2 Identifying and Estimating Treatment Effects

### 2.1 Specifying Heterogeneous Treatment Effects

To specify individual treatment effects, we rely on the potential outcomes framework (see Imbens and Rubin, 2015; Deschenes and Meng, 2018, for a general treatment and an application to environmental economics, respectively). To describe the framework of heterogeneous treatment effects, we return to the effect of a gasoline content regulation. For each state $i$ in each time period $t$, there exist two potential outcomes: the pollution emission that would hold if the state was subject to the regulation, $Y_{it}(1)$, and the pollution emission that would hold if the state was not subject to the regulation, $Y_{it}(0)$. Note that these potential outcomes are themselves random variables. Throughout we assume that the potential outcomes for state $i$ do not depend on the treatment status of any other state, which rules out the possibility of treatment spillovers between units (see Deschenes and Meng, 2018; Vazquez-Bare, 2019, for recent work in the analysis of spillovers). As is well known, because only one of these outcomes can be observed, the pair always consists of the actual outcome and the unobserved, counterfactual outcome. Paired with the potential outcomes is a binary treatment indicator, $D_{it}$, so the observed outcome is

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

where $\tau_{it} = Y_{it}(1) - Y_{it}(0)$ is the treatment, or causal, effect for state $i$ in period $t$. Because the treatment effect potentially differs over states and time periods, this expression captures all possible heterogeneous treatment effects.

Since both elements of the potential outcome vector can never be simultaneously observed, $\tau_{it}$ is unobservable. The conditions required to identify $\tau_{it}$ are very strong and unlikely to hold in most contexts, so attention usually turns to features of the distribution of the treatment
effects, the most common of which is the average treatment effect among treated units (ATT):

$$E[\tau_{it}|D_{it} = 1] = E[Y_{it}(1) - Y_{it}(0)|D_{it} = 1].$$

(3)

The goal of the upcoming sections is to analyze the relationship between specification (1) and features of $\tau_{it}$ and, more precisely, to provide conditions under which Equation (1) gives useful information about the average treatment effect under effect heterogeneity.

### 2.2 A General Identification Result

A necessary condition to be able to estimate and conduct inference on parameters of interest is identification. Parameter identification is concerned with the question of when we can expect to learn something about our target parameter using the available data. More precisely, identification seeks to provide restrictions on the data generating process to ensure that the (unobservable) parameter of interest can be expressed as a function of observable data.

Under what conditions can the ATT parameter be identified? To state the conditions, it is helpful to first consider how the pollution emission of a state evolves over time in the absence of the gasoline content regulation. Consider two time periods $t$ and $t'$ with $t < t'$. In the absence of the regulation, the average baseline trend in pollution emission is

$$E[Y_{it}(0) - Y_{it}(0)].$$

If we introduce the regulation for some states in period $t'$, then the key assumption for identification of the ATT parameter is that the baseline trend is identical for states that are regulated (for whom $D_{it'} = 1$) and for states that are not regulated (for whom $D_{it'} = 0$). The assumption of identical baseline trends is the parallel trends assumption.

We now formalize the identification result.

**Proposition 1 (Identification)** Consider a pair of periods $0 \leq t < t' \leq T$ such that:

1. $P[D_{it} = 0] = 1$ and $0 < P[D_{it'} = 1] < 1$,
2. $E[Y_{it'}(0) - Y_{it}(0)|D_{it'} = 1] = E[Y_{it'}(0) - Y_{it}(0)|D_{it'} = 0]$.

Then,

$$E[\tau_{it'}|D_{it'} = 1] = E[Y_{it'} - Y_{it}|D_{it'} = 1] - E[Y_{it'} - Y_{it}|D_{it'} = 0].$$

Proposition 1 shows that the average treatment effect on the treated can be expressed as a function of the observable data - the common difference-in-differences expression. The two conditions are: (1) the availability of a “pre-treatment” period $t$ in which the probability of treatment is zero and a “post-treatment” period $t'$ in which the probability of treatment is strictly between 0 and 1; and (2) the equality of baseline trends between the treatment and untreated groups. Intuitively, the second condition allows treated and untreated units to differ in their baseline outcome levels, which is often plausible in observational settings where the treatment is not randomly assigned.

Because identification is a concept related to a population (or data generating process) and not a specific sample, this result does not make any reference to sampling issues such as sample size, independence or identical distributions across observations. In practice, these population quantities are unknown and need to be estimated using a panel sample of independent copies of $(Y_{it}(0), Y_{it}(1), D_{it})$, which we will denote $(Y_{it}(0), Y_{it}(1), D_{it})_{i \geq 1}^{G}$. Next we study the performance of different estimation strategies that use a sample of panel data.

### 3 Properties of Treatment Effects Estimators

Our main focus in this section will be a two-way fixed effects regression like Equation (1) or variations thereof. Our goal is to understand the relationship between different estimators for $\beta$ and the average treatment effects defined in previous sections. This equation is not believed
to be the correct representation of how the outcome arises from the treatment, rather it is an approximation where the presence of the fixed effects are designed to control for unobserved heterogeneity. The key is to relate the parameters of (1) to the heterogeneous treatment effects and then to determine conditions on these effects under which the parameters provide useful information about the effects.

3.1 Two-period case

Let us start by considering the simple case of a balanced panel in which all units are observed in two periods, \( t = 0 \) and \( t = 1 \), and where a binary treatment is administered to a subset of units only in the second period, so that \( P[D_{0i} = 0] = 1 \) and \( 0 < P[D_{1i} = 1] < 1 \) for all \( i \). As described in Section 2.1, each unit \( i \) at time \( t \) has a vector consisting of two potential outcomes, \((Y_{it}(0), Y_{it}(1))\) which indicates the outcome that would be observed without and with the treatment, respectively. Since the treatment is administered in period \( t = 1 \) only,

\[
Y_{i0} = Y_{i0}(0) \\
Y_{i1} = Y_{i1}(0) + \tau_{i1}D_{i1}
\]

from which it follows that we can write the observed outcome as:

\[
Y_{it} = Y_{i0}(0) + (Y_{i1}(0) - Y_{i0}(0))t + \tau_{i1}D_{it}.
\]

This is simply a compact version of the following information:

\[
Y_{it} = \begin{cases} 
Y_{i0}(0) & \text{if } t = 0 \\
Y_{i1}(0) & \text{if } t = 1 \text{ and } D_{i1} = 0 \\
Y_{i1}(1) & \text{if } t = 1 \text{ and } D_{i1} = 1
\end{cases}
\]

(5)

There are two important features of this equation that are worth emphasizing. First, Equation (4) is fully nonparametric. Since Equations (4) and (5) are equivalent, we see that Equation (4) enumerates all the possible cases and therefore does not impose any parametric or functional form assumptions on the observed or potential outcomes.

Second, as in Equation (1), Equation (4) 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, \((Y_{i1}(0) - Y_{i0}(0))t\), and a treatment indicator with its corresponding (random) coefficient \( \tau_{it} \). The main difference between Equations (1) and (4), however, is that Equation (4) allows for arbitrary heterogeneity of the coefficients across units and over time.

We can now use Equation (4) to understand how a regression model like (1) performs in a setting with unrestricted heterogeneity in treatment effects. A simple strategy to estimate (1) is the first-difference approach, consisting of taking the difference between periods \( t = 1 \) and \( t = 0 \). This yields:

\[
Y_{i1} - Y_{i0} = Y_{i1}(0) - Y_{i0}(0) + \tau_{i1}D_{i1}
\]

(6)

The above equation can be mapped into a simple linear regression including a constant and one covariate, using observations from period \( t = 1 \) only:

\[
\Delta Y_i = \alpha_{FD} + \beta_{FD}D_{i1} + u_i
\]

(7)

where \( \Delta Y_i = Y_{i1} - Y_{i0} \) and noting that \( D_{i1} - D_{i0} = D_{i1} \). The least squares estimator for \( \beta_{FD} \) is:

\[
\hat{\beta}_{FD} = \frac{\sum_{i=1}^{G} \Delta Y_i (D_{i1} - \bar{D}_1)}{\sum_{i=1}^{G} D_{i1} (D_{i1} - \bar{D}_1)}, \quad \bar{D}_1 = \frac{1}{G} \sum_{i=1}^{G} D_{i1}
\]

After simple algebra, this estimator can be rewritten as:

\[
\hat{\beta}_{FD} = \frac{\sum_i \Delta Y_i D_{i1}}{\sum_i D_{i1}} - \frac{\sum_i \Delta Y_i (1 - D_{i1})}{\sum_i (1 - D_{i1})}
\]
which is a difference in the average outcome changes between treated and untreated units. For this reason, this estimator is commonly known as the difference-in-differences (DiD) estimator. Using Equation (4) and taking conditional expectations,

\[ E[\hat{\beta}_{FD}|D] = \frac{\sum_i E[\tau_{i1}|D_{i1} = 1]D_{i1}}{\sum_i D_{i1}} + \frac{\sum_i E[\Delta Y_i(0)|D_{i1} = 1]D_{i1}}{\sum_i D_{i1}} - \frac{\sum_i E[\Delta Y_i(0)|D_{i1} = 0](1 - D_{i1})}{\sum_i (1 - D_{i1})} \]

where \( D = (D_{it})_{i,t} \) is the vector of all treatment indicators. Thus, the expectation of \( \hat{\beta}_{FD} \) equals the average ATT in period 1 across all treated units, plus a bias term that equals the difference in average trends of the average potential outcomes between treated and untreated units. Intuitively, by comparing the change in the outcomes between treated and untreated units, the first-difference estimator is mixing two factors: the actual effect of the treatment on the treated units, and the difference in outcome trends that would have been observed in the absence of the treatment. Without further assumptions, the treated group could have experienced a different trend in the absence of treatment compared to the untreated group. If this is so, the first-difference estimator is biased. Under the parallel-trends assumption, the bias disappears and the conditional expectation of \( \hat{\beta}_{FD} \) equals the average ATT across treated units. This idea is illustrated in Figure 1, and formally summarized in Proposition 2.

**Proposition 2 (Difference-in-differences)** Suppose the following conditions hold:

1. **Sampling:** for \( t = 0, 1 \), \( (Y_{it}(0), Y_{it}(1), D_{it}) \) are independent across \( i \).
2. **Treatment assignment:** \( D_{i0} = 0 \) for all \( i \) and \( D_{i1} = 1 \) for some (but not all) \( i \).
3. **Parallel trends:** there is a constant \( \kappa_1 \) such that \( E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}] = \kappa_1 \) for all \( i \).

Then,

\[ E[\hat{\beta}_{FD}|D] = \frac{\sum_{i=1}^{G} E[\tau_{i1}|D_{i1} = 1]D_{i1}}{\sum_{i=1}^{G} D_{i1}}. \]

Note that Proposition 2 allows for arbitrary heterogeneity in the average treatment effects. That is, not only \( \tau_{it} \), but also its distribution is allowed to vary across units and over time. In particular, this result allows the ATT to be different for different units. For example, if units represent US states, the same policy can have different effects in, say, states \( i \) and \( j \), so
where \( \mathbb{E}[\tau_{it}|D_{it} = 1] \neq \mathbb{E}[\tau_{it}|D_{it} = 1] \). In other cases, it may be reasonable to assume that observations are drawn from the same population, and hence the average effect is the same for all units. In that case, the above result simplifies to \( \mathbb{E}[\beta_{FD}|D] = \mathbb{E}[\tau_{1i}|D_{1i} = 1] \).

An alternative strategy commonly used to implement Equation (1) in practice is by either directly estimating the fixed effects including \( G \) individual dummies and \( T \) time dummies (least squares dummy variables) or, equivalently, estimating a linear regression on the double-demeaned variables:

\[
\hat{Y}_{it} = \beta_{FE}\bar{D}_{it} + \nu_{it}
\]

where

\[
\hat{Y}_{it} = Y_{it} - \bar{Y} - \tilde{D}_{it} + \bar{D}, \quad \bar{D}_{it} = D_{it} - \bar{D}_{it} + \bar{D}
\]

and where

\[
\bar{Y}_t = \frac{1}{T+1} \sum_{t=0}^{T} Y_{it}, \quad \bar{D}_t = \frac{G}{G(T+1)} \sum_{t=0}^{T} \sum_{i=1}^{G} Y_{it}
\]

with analogous definitions for the treatment variable. It can be shown that, with only two time periods, \( \hat{\beta}_{FD} = \hat{\beta}_{FE} \). This equivalence, however, does not hold with more than two periods, and this fact has important implications for estimation, as we discuss next.

### 4 Multiple periods

Now suppose that we have access to a data set with multiple time periods \( t = 0, \ldots, T \), where a policy is implemented at time \( t^* \) with \( 0 < t^* < T \). For simplicity, we will assume we have a balanced panel in which all observations are observed in all time periods. Note that in this setting, the treatment indicator can be written as:

\[
D_{it} = D^*_i \mathbb{I}(t \geq t^*)
\]

where \( D^*_i = D_{it^*} \) indicates whether unit \( i \) is eventually treated. We can generalize Equation (4) as:

\[
Y_{it} = Y_{i0}(0) + \sum_{s=1}^{T} (Y_{is}(0) - Y_{i0}(0)) \mathbb{I}(s = t) + \sum_{s=1}^{T} \tau_{is}D_{is}\mathbb{I}(s = t)
\]

Proposition 1 ensures that if the parallel-trends assumption holds between any two periods \( t_0 < t^* \) and \( t_1 \geq t^* \), \( \mathbb{E}[\tau_{it}|D^*_i = 1] \) is identified as long as there is sufficient variation in treatment assignment, and the methods discussed in the previous section can be immediately applied to compare any two periods \( t_0 \) and \( t_1 \).

The availability of multiple pre-treatment periods has an important empirical implication that can be exploited to assess the validity of the identification strategy. Suppose that we have three periods, \( t = 0, 1, 2 \), where the policy is implemented in the last period period \( t^* = T = 2 \) so that both \( t = 0 \) and \( t = 1 \) are pre-treatment periods. 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:

\[
\mathbb{E}[Y_{i1} - Y_{i0}|D_{i2} = 1] = \mathbb{E}[Y_{i1} - Y_{i0}|D_{i2} = 0] \quad \text{(8)}
\]
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.

It is important to note that condition (8) is neither necessary nor sufficient to achieve identification of the ATT. The average trends between groups could be parallel between periods 0 and 1 and then diverge between periods 1 and 2, or vice versa. Since the counterfactual trend for the treated group between periods 1 and 2 is unobservable, the identification assumption is fundamentally untestable. Nevertheless, condition (8), while not a formal test of the identification assumption, can provide some compelling evidence to support it.

4.1 Two-way Fixed Effects Estimators

With multiple time periods, researchers generally pool information from all time periods into a single regression including unit and time fixed effects as in Equation (1). The parameter of interest is estimated after double-demeaning the variables of interest:

\[ \ddot{Y}_{it} = \beta_{FE} \ddot{D}_{it} + \nu_{it} \]

where

\[ \ddot{Y}_{it} = Y_{it} - \bar{Y}_{i} - \bar{Y}_{t} + \bar{Y}, \quad \ddot{D}_{it} = D_{it} - \bar{D}_{i} - \bar{D}_{t} + \bar{D} \]

and where

\[ \bar{Y}_{i} = \frac{1}{T+1} \sum_{t=0}^{T} Y_{it}, \quad \bar{Y}_{t} = \frac{1}{G} \sum_{i=1}^{G} Y_{it}, \quad \bar{Y} = \frac{1}{G(T+1)} \sum_{t=0}^{T} \sum_{i=1}^{G} Y_{it} \]

with analogous definitions for the treatment variable. The OLS estimator for \( \beta_{FE} \) is:

\[ \hat{\beta}_{FE} = \frac{\sum_{t} \sum_{i} Y_{it} \ddot{D}_{it}}{\sum_{t} \sum_{i} \ddot{D}_{it} \ddot{D}_{it}}. \]

The following result characterizes the expected value of \( \hat{\beta}_{FE} \).

**Proposition 3 (Two-way fixed effects)** Suppose the following conditions hold:

1. **Sampling:** for each \( t = 0, \ldots, T \), \( (Y_{it}(0), Y_{it}(1), D_{it}) \) are independent across \( i \).
2. **Treatment assignment:** there is a time period \( 0 < t^* \leq T \) such that \( D_{it} = 0 \) for all \( i \) and \( t < t^* \) and \( D_{it} = 1 \) for some (but not all) \( i \) and for \( t \geq t^* \). Moreover, letting \( D_{it^*} = D^*_{i} \), we have that \( D_{it} = D^*_{i} \) for all \( t \geq t^* \).
3. **Parallel trends:** for each \( t = 1, \ldots, T \) there exists a constant \( \kappa_{t} \) such that \( E[Y_{it}(0) - Y_{it-1}(0)|D^*_i] = \kappa_{t} \) for all \( i \).

Then,

\[ E[\hat{\beta}_{FE}|D] = \frac{1}{T + 1 - t^*} \sum_{t \geq t^*} \left\{ \frac{\sum_{i} E[\tau_{it}|D^*_i = 1] D^*_i}{\sum_{i} D^*_i} \right\}. \]

Proposition 3 shows that under conditions 1-3, the conditional expectation of the two-way fixed-effects is a simple average of the average treatment effects on the treated over all the post-treatment periods. Condition 2 states that the program of interest is implemented in some period \( t^* \) before which nobody is treated. After period \( t^* \), a subset of units is treated and remain treated until the last period. Note that this requirement rules out staggered adoption designs in which different units enter the treatment in different periods (we will relax this assumption in the next section). On the other hand, this condition does not necessarily mean that the treatment is implemented repeatedly in all post-treatment periods. Suppose for example that we observe a panel of firms from years 2000 to 2010, and that a state introduces a gasoline content regulation in 2006. Even if this policy change is only implemented in 2005 and then removed in the following years, its effect could be persistent over time. Thus, putting this
treated state back into the untreated pool in the following years could render the estimator biased. Setting $D_{it} = D_i^*$ for all $t \geq t^*$ avoids this issue.

As in Proposition 2, the distribution of treatment effects is allowed to vary arbitrarily across units and time periods. In the particular case in which units are drawn from the same distribution, the inner sum in the above result reduces to $E[\tau_{it}|D_i^* = 1]$, so $\hat{\beta}_{FE}$ recovers a simple average of the ATT across post-treatment periods.

4.2 Variation in Treatment Timing

The previous section discussed conditions under which, with multiple periods, the two-way FE estimator is unbiased for the simple average of ATTs over treated units and treated periods. One of the conditions to obtain this result is that all eventually treated units enter the treatment simultaneously at some period $0 < t^* \leq T$. In general, however, different units may enter the treatment at different times. This occurs, for example, when different states adopt a certain policy in different years. These settings are sometimes known as staggered adoption designs. For example, suppose we analyze the effect of industrial plant openings on air quality and health outcomes in US states. Since plants openings can vary widely in their timing, different geographic sites become treated in different periods. See e.g. Currie, Davis, Greenstone, and Walker (2015).

Several studies have shown that, with variation in the timing of the treatment, fixed effects estimators may not recover meaningful causal parameters. In earlier work, Wooldridge (2005) provided sufficient conditions for consistently estimating the average slope in one-way fixed effects models with random coefficients, whereas Chernozhukov, Fernández-Val, Hahn, and Newey (2013) showed that, under random sampling, the one-way fixed effects estimator is inconsistent for the ATT. See also Gibbons, Suárez-Serrato, and Urbancic (2018) and Imai and Kim (2019b) for more recent work on one-way fixed effects. Borusyak and Jaravel (2017), Athey and Imbens (2018) and de Chaisemartin and D’Haultfoeuille (2019) showed, under different conditions, that when treatment effects are heterogeneous across units, the two-way fixed effects estimator recovers a weighted average of treatment effects across units and treated periods, where the weights can be negative.

We now introduce some additional notation to elaborate on these ideas. Let $t^*_i > 0$ indicate the time at which each unit $i$ enters the treatment. For units that are not treated in the observed period $0, \ldots, T$, we set $t^*_i = T + 1$. More formally,

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

For $t = 0, \ldots, T$ the treatment indicator is $D_{it} = 1(t^*_i \leq t)$. Finally, let $D_i^* = 1(t^*_i \leq T)$ be an indicator that unit $i$ is eventually treated, and $t^* = \frac{1}{G} \sum_{i} t^*_i$ be the average treatment adoption period. The following result is a simplified version of Theorem 1 in de Chaisemartin and D’Haultfoeuille (2019).

**Proposition 4 (Fixed effects with variation in treatment timing)** Suppose the following conditions hold:

1. **Sampling:** for each $t = 0, \ldots, T$, $(Y_{it}(0), Y_{it}(1), D_{it})$ are independent across $i$.
2. **Treatment assignment:** each unit $i$ enters treatment in period $t^*_i$, and $D_{it} = 1(t^*_i \leq t)$.
3. **Parallel trends:** for each $t = 1, \ldots, T$ there exists a constant $\kappa_t$ such that $E[Y_{it}(0) - Y_{it-1}(0)|D_{it}] = \kappa_t$ for all $i$.

Then,

$$E[\hat{\beta}_{FE}|D] = \sum_{i} \sum_{t^*_i \geq t^*_i} E[\tau_{it}|D_{it} = 1] \omega_{it}$$
Figure 2: Weights assigned by $\hat{\beta}_{FE}$ to a unit with $t_i^* = 1$

where

$$\omega_{it} = \frac{1 - D_t + \frac{t_i^* - t^*}{T+1}}{\sum_{i: D_i^* = 1} \sum_{t \geq t_i^*} \left(1 - D_t + \frac{t_i^* - t^*}{T+1}\right)}.$$ 

See de Chaisemartin and D’Haultfoeuille (2019) for a more general result that allows for variable group sizes, among other things. Proposition 4 shows that the conditional expectation of the two-way fixed effects estimator can be written as weighted average of average treatment effects on the treated across units and over treated periods. The weights assigned to each $E[\tau_{it}|D_{it} = 1]$ depends on the proportion of treated units in each time period, $D_t$, the time at which unit $i$ enters treatment, $t_i^*$, and the average time at which units in the sample enter treatment, $\bar{t}$. Importantly, these weights can be negative. More precisely, weights can be negative for units with $t_i^* < \bar{t}^*$ and for periods in which $1 - D_t$ is small enough. This shows that the two-way fixed effects estimator can give negative weights in later periods to units that get the treatment in earlier periods.

We illustrate this issue in Figure 2. We simulate a sample with three time periods $t = 0, 1, 2$ and $G = 500$ units, out of which 25 percent are never treated ($G_0 = 125$), a proportion $P[t_i^* = 1]$ is treated in period $t = 1$ ($G_1 = \lfloor G \times P[t_i^* = 1] \rfloor$, where $\lfloor x \rfloor$ indicates the integer value of $x$), and the remaining $G - G_0 - G_1$ are treated in period $t = 2$. We vary $P[t_i^* = 1]$ from 0.01 to 0.74. Figure 2 shows that the weights for a unit with $t_i^* = 1$ in $t = 1$ are always positive regardless of the value of $P[t_i^* = 1]$, but that the weights for that same unit in $t = 2$ can be negative when the total proportion of units getting treatment at $t = 1$ is small. In other words, when only a few units enter treatment in period $t = 1$, these units receive negative weights in $t = 2$. In this example, all weights become positive when $P[t_i^* = 1]$ crosses the 50 percent threshold.

Goodman-Bacon (2019) provides some insight into this phenomenon by showing that the two-way-fixed effects estimator can be written as a variance-weighted average of all possible two-group/two-period DiD estimators that can be formed based on treatment timing and time
This specification has two sets of parameters of interest. First, $\beta$ similarly to what happens in two-way FE models, estimators for the specification in a setting with heterogeneous treatment effects and variation in treatment timing.

Although the parallel-trends assumption holds. Specification (9), sometimes called an event study parallel-trends assumption could 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.

For instance, de Chaisemartin and D’Haultfœuille (2019) propose an alternative causal estimand that can be estimated unbiasedly under certain conditions. Imai and Kim (2019a) propose an alternative causal estimand and estimators that may provide more accurate information on causal parameters. For instance, de Chaisemartin and D’Haultfœuille (2019) propose an alternative causal estimand that can be estimated unbiasedly under certain conditions. Imai and Kim (2019a) propose a multi-period DiD estimator that eliminates the bias of the two-way FE estimator. Callaway and Sant’Anna (2019) derive an inverse-probability-weighted estimator for group-time average treatment effects that exploits the availability of covariates. We refer the reader to the original papers for further details and software implementation of these alternative strategies.

4.3 Estimating Time-Varying Effects

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

$$Y_{it} = \alpha + \mu_t + \sum_{l=1}^{L} \beta_l D_{lt+l} + \sum_{s=0}^{S} \beta_s D_{it-s} + \epsilon_{it}$$

(9)

This specification has two sets of parameters of interest. First, $\beta_s$, 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, $\beta_l$, 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 $\beta_l$ coefficients are expected to be non-significant if the parallel-trends assumption holds. Specification (9), 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.

Borusyak and Jaravel (2017) and Abraham and Sun (2019) analyze this type of dynamic specification in a setting with heterogeneous treatment effects and variation in treatment timing. Similarly to what happens in two-way FE models, estimators for the $\beta_s$ coefficients may average
across different treatment effects with misleading weighting schemes including negative weights. Furthermore, Abraham and Sun (2019) show that the $\beta_t$ coefficients may pick up some of the treatment effects in future periods, yielding spurious significant estimates even when the parallel-trends assumption holds.

4.4 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:

$$Y_{it} = \alpha_i + \mu_t + \beta D_{it} + Z_{it}' \gamma + \eta_{it}$$

Importantly, including covariates in this fashion implicitly imposes a linear relationship between the outcome and the covariates, which may result in 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. 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 (de Chaisemartin and D’Haultfoeuille, 2019).

5 Inference

For panel data estimation with a two-way fixed effects model as in (1) 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 $\Omega$ be the variance matrix for the errors $\{\varepsilon_{it}\}$. Because the errors are independent across clusters, $\Omega$ is a block diagonal matrix, where diagonal block $i$ corresponds to cluster (unit) $i$. The structure is summarized as

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

otherwise $E[\varepsilon_{it}\varepsilon_{js}] = 0$. 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 $\theta$ be the vector that captures all the coefficients in (1), the individual and time fixed effects as well as the treatment effect, $\beta$. Let $\hat{\theta}$ be the OLS estimator of $\theta$ with variance matrix $V = Var(\hat{\theta}|X)$. Because $\Omega$ is block diagonal, the common expression

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

simplifies considerably to

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

11
where $X_i$ and $\Omega_i$ correspond to the explanatory variables and error variance matrix for unit $i$. This simplification greatly reduces the number of unknown terms in $\Omega$, so it is possible to obtain a consistent estimator of $V$ without further restrictions on $\Omega$. The widely used cluster-robust variance estimator employs the OLS residuals for each unit, $\hat{\varepsilon}_i$, as

$$
\hat{V} = (X^TX)^{-1} \sum_{i=1}^{G} X_i^T \hat{\varepsilon}_i \hat{\varepsilon}_i^T X_i (X^TX)^{-1}.
$$

(10)

We study tests of a null hypothesis of the form $a^T \theta$. 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

$$
t_{\text{stat}} = \frac{a^T(\hat{\theta} - \theta)}{a^T \hat{V} a}.
$$

(11)

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

$$
t_{\text{stat}} = \frac{\hat{\beta}}{\hat{s}_\beta},
$$

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

### 5.1 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, $\hat{V}$, 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 $\hat{\varepsilon}_{it} = Y_{it} - \bar{Y}$. The cluster-robust estimator $\hat{V}$ is a function of $X_i^T \hat{\varepsilon}_i$, which in this intercept only model equals

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

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.

If the units are homogeneous, in the sense that $E[XT_i \Omega_i X_i]$ is identical for all $i$, then standard methods of proof establish that $t_{\text{stat}} \rightarrow N(0, 1)$ as $G \rightarrow \infty$. From this emerges the conventional wisdom that if the number of units, $G$, is sufficiently large then the distribution of $t_{\text{stat}}$ is approximately normal. Yet how likely is it that the units are homogeneous? One of the main reasons to employ cluster-robust standard errors is to allow $\Omega_i$ to differ over units. Moreover, in almost all data sets the units do not have the exact same values for the matrix of explanatory variables, so $X_i$ also differs over units. If that is the case, it is very likely that $E[XT_i \Omega_i X_i]$ also differs over units. Thus it is highly unlikely that any data set satisfies the requirement of homogeneous units.

Carter, Schneipel, and Steigerwald (2017) (CSS) were the first to establish conditions for asymptotic normality of the test statistic $t_{\text{stat}}$ while allowing clusters to be heterogeneous. Their method of proof introduces an adjustment to the number of units to account for heterogeneity across units.\(^1\) The adjustment term increases with degree of heterogeneity and is used to produce

\(^1\)Djogbenou, MacKinnon, and Nielsen (2019) provide a more traditional proof of asymptotic normality, but their conditions do not yield an adjustment term.
an effective sample size. Because the sample size is based on the number of clusters, CSS term the adjusted quantity the effective number of clusters and show that it is the effective number of clusters that must grow to obtain asymptotic normality of $t_{\text{stat}}$.

The adjustment is constructed as follows. First, for each unit construct the measure

$$
\gamma_i = a^T X^T X^{-1} X_i^T \Omega_i X_i (X^T X)^{-1} a
$$

and $\bar{\gamma} = \frac{1}{G} \sum_{i=1}^{G} \gamma_i$. With these in hand, construct the adjustment

$$
\Gamma_a = \frac{1}{\bar{\gamma}^2} \sum_{i=1}^{G} (\gamma_i - \bar{\gamma})^2.
$$

The resulting effective number of clusters is

$$
G_{a}^* = \frac{G}{1 + \Gamma_a}.
$$

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 $t_{\text{stat}}$.

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 $\Omega_i$. (Because $\Gamma_a$ 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 $\Gamma_a$. 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.

### 5.2 Correct Inference

If the effective number of clusters is large, then the distribution of $t_{\text{stat}}$ 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 $G_{a}^*$ 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 $G_{a}^*$ lead to larger variation in the cluster-robust standard errors, which in turn leads to non-normality of the $t$-statistic. If $G_{a}^*$ is extremely low, then the distribution of $t_{\text{stat}}$ can be bimodal.

MacKinnon and Webb (2017) study how best to approximate the distribution of $t_{\text{stat}}$ under cluster heterogeneity. In their study, cluster sizes vary to reflect the relative populations of the 50 US states. In this setting, $t_{\text{stat}}$ 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 $\text{Student} - t(G_{a}^*)$. In their simulation setting, the true correlation matrix that enters $G_{a}^*$ 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 $G_{a}^*$, 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 \( \{v_1\hat{\varepsilon}_1, \ldots, v_G\hat{\varepsilon}_G\} \) 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 \( \pm 1.96 \)) are used. If the effective number of clusters is less than 50, use the wild cluster bootstrap to form the \( P \)-value.

6 Conclusion

The two-way fixed effects equation 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. We show that 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

ic Studies, 72(1), 1–19.


Appendix

A Proof of Proposition 1

For units that are treated in period \( t' \),

\[
E[Y_{it'} - Y_{it} | D_{it'} = 1] = E[Y_{it'}(1) - Y_{it}(0) | D_{it'} = 1]
\]

\[
= E[Y_{it'}(1) - Y_{it}(0) | D_{it'} = 1] + E[Y_{it}(0) - Y_{it'}(0) | D_{it'} = 1]
\]

\[
= E[\tau_{it'} | D_{it'} = 1] + E[Y_{it'}(0) - Y_{it}(0) | D_{it'} = 0]
\]

\[
= E[\tau_{it'} | D_{it'} = 1] + E[Y_{it'} - Y_{it} | D_{it'} = 0]
\]

where the second line adds and subtracts \( E[Y_{it'}(0) | D_{it'} = 1] \), the third line uses the definition of treatment effect, the fourth line uses the parallel trends assumption and the last line uses that for units that are never treated, the observed outcomes equal the potential outcomes under no treatment. \( \square \)

B Proof of Proposition 2

The first difference estimator is:

\[
\hat{\beta}_{FD} = \sum_i \frac{\Delta Y_i (D_{it} - \bar{D}_i)}{\sum_i D_{it} (D_{it} - \bar{D}_i)}
\]

\[
= \frac{(1 - \bar{D}_1) \sum_i \Delta Y_i D_{it} - \bar{D}_1 \sum_i \Delta Y_i (1 - D_{it})}{(1 - D_{it}) \sum_i D_{it}}
\]

\[
= \frac{\sum_i \Delta Y_i D_{it}}{\sum_i D_{it}} - \frac{\sum_i \Delta Y_i (1 - D_{it})}{\sum_i (1 - D_{it})}.
\]

Taking conditional expectations, using the fact that \( \Delta Y_i = \Delta Y_i(0) + \tau_{it} D_{it} \) and the parallel trends assumption,

\[
E[\hat{\beta}_{FD} | D] = \sum_i \frac{E[\tau_{it} | D_{it} = 1] D_{it}}{\sum_i D_{it}} + \sum_i \frac{E[\Delta Y_i(0) | D_{it} = 1] D_{it}}{\sum_i D_{it}} - \sum_i \frac{E[\Delta Y_i(0) | D_{it} = 0] (1 - D_{it})}{\sum_i (1 - D_{it})}
\]

\[
= \sum_i \frac{E[\tau_{it} | D_{it} = 1] D_{it}}{\sum_i D_{it}} + \kappa_1 \sum_i \frac{D_{it}}{\sum_i D_{it}} - \kappa_1 \sum_i (1 - D_{it})
\]

\[
= \sum_i \frac{E[\tau_{it} | D_{it} = 1] D_{it}}{\sum_i D_{it}}. \square
\]

C Proof of Proposition 3

Recall that:

\[
\hat{\beta}_{FE} = \frac{\sum_i \sum_t Y_{it} \bar{D}_{it}}{\sum_i \sum_t D_{it} \bar{D}_{it}}, \quad \bar{D}_{it} = D_{it} - \bar{D}_i - \bar{D}_t + \bar{D}
\]
and thus

$$E[\hat{\beta}_{FE}] = \frac{\sum_t \sum_i E[Y_{it}|D] \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}}$$

$$= \frac{\sum_t \sum_i E[Y_{it}(0)|D] \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}} + \frac{\sum_t \sum_i E[\tau_{it}|D] D_{it} \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}}$$

$$= \frac{\sum_t \sum_i E[Y_{it}(0)|D] \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}} + \frac{\sum_t \sum_i E[\tau_{it}|D] D_{it} \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}}$$

$$= \frac{\sum_t \sum_i E[Y_{it}(0) - \bar{Y}_{i}(0)|D] \hat{D}_{it} + \sum_t \sum_i E[\tau_{it}|D] D_{it} \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}}$$

where the third equality follows from $D_{it} = D_i^* 1(t \geq t^*)$ and the fourth equality uses the fact that $\sum_i \hat{D}_{it} = 0$. For the first term in the last equation, use that:

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

where $\Delta_{is}(0) = Y_{is}(0) - Y_{i(s-1)}(0)$. Under the assumption that $E[\Delta_{is}(0)|D_i^*]$ does not vary over $i$, the first term becomes zero using that $\sum_i \hat{D}_{it} = 0$. We also have the following facts that can be verified by direct calculation:

$$\hat{D}_{it} = \left(1(t \geq t^*) + \frac{t^*}{T+1}\right) (D_i^* - \bar{D}^*)$$

$$D_{it} \hat{D}_{it} = 1(t \geq t^*) \frac{t^*}{T+1} D_i^*(1 - \bar{D}^*)$$

$$\sum_i D_{it} \hat{D}_{it} = 1(t \geq t^*) \frac{t^*}{T+1} N \bar{D}^*(1 - \bar{D}^*)$$

$$\sum_t \sum_i D_{it} \hat{D}_{it} = (T + 1 - t^*) \frac{t^*}{T+1} \bar{D}^*(1 - \bar{D}^*)$$

It follows from these equalities that:

$$\frac{\sum_t \sum_i E[\tau_{it}|D_i^* = 1] D_{it} \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}} = \frac{\sum_{t \geq t^*} \sum_t \sum_i E[\tau_{it}|D_i^* = 1] D_i^*(1 - \bar{D}^*)}{\sum_{t \geq t^*} (T + 1 - t^*) N \bar{D}^*(1 - \bar{D}^*)}$$

$$= \frac{1}{T + 1 - t^*} \sum_{t \geq t^*} \left\{ \frac{\sum_i E[\tau_{it}|D_i^* = 1] D_i^*}{\sum_i D_i^*} \right\} \Box$$

## D Proof of Proposition 4

This result is a particular case of Theorem 1 in de Chaisemartin and D'Haultfœuille (2019), but we provide a direct proof here for completeness. We have that

$$\hat{\beta}_{FE} = \frac{\sum_t \sum_i Y_{it} \hat{D}_{it}}{\sum_t \sum_i D_{it} \hat{D}_{it}}, \quad \hat{D}_{it} = D_{it} - \bar{D}_{it} - \hat{D}_{i} + \bar{D}$$
and thus

\[ \hat{\beta}_{FE} = \frac{\sum_t \sum_i Y_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} = \frac{\sum_t \sum_i Y_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} + \frac{\sum_t \sum_i \tau_{it} D_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} = \frac{\sum_t \sum_i (Y_{it} - \bar{Y}_i(0)) \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} + \frac{\sum_t \sum_i \tau_{it} D_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} \]

where the third equality uses the fact that \( \sum_t \bar{D}_{it} = 0 \). Taking expectations,

\[ E[\hat{\beta}_{FE}|D] = \frac{\sum_t \sum_i E[Y_{it}(0) - \bar{Y}_i(0)|D] \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} + \frac{\sum_t \sum_i E[\tau_{it}|D] D_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} \]

Note that for \( t = 0, \ldots, T \),

\[ D_{it} = \mathbb{1}(t_i^* \leq t) \]

\[ \bar{D}_t = \frac{1}{T+1} \sum_{i=0}^T D_{it} = \frac{1}{T+1} \sum_{i=0}^T \mathbb{1}(t_i^* \leq t) = 1 - \frac{t_i^*}{T+1} \]

\[ \bar{D} = \frac{1}{N} \sum_{i=1}^N \bar{D}_i = 1 - \frac{\bar{t}^*}{T+1} \]

and therefore

\[ D_{it} \bar{D}_{it} = D_{it}(D_{it} - \bar{D}_i - \bar{D}_t + \bar{D}) = D_{it} \left(1 - \bar{D}_t + \frac{t_i^* - \bar{t}^*}{T+1}\right). \]

Thus under the parallel-trends assumption,

\[ E[\hat{\beta}_{FE}|D] = \frac{\sum_t \sum_i E[\tau_{it}|D_{it} = 1] D_{it} \bar{D}_{it}}{\sum_t \sum_i D_{it} \bar{D}_{it}} = \sum_{i:D_{it}^*=1} \sum_{t \geq t_i^*} E[\tau_{it}|D_{it} = 1] \omega_{it} \]

where

\[ \omega_{it} = \frac{1 - \bar{D}_t + \frac{t_i^* - \bar{t}^*}{T+1}}{\sum_{i:D_{it}^*=1} \sum_{t \geq t_i^*} \left(1 - \bar{D}_t + \frac{t_i^* - \bar{t}^*}{T+1}\right)}. \]