Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators

Preprint · August 2021

CITATIONS
0

READS
9,371

1 author:

Jeffrey M. Wooldridge
Michigan State University
246 PUBLICATIONS **69,621** CITATIONS

SEE PROFILE

Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators

Jeffrey M. Wooldridge
Department of Economics
Michigan State University

This Version: August 16, 2021

Abstract: I establish the equivalence between the two-way fixed effects (TWFE) estimator and an estimator obtained from a pooled ordinary least squares regression that includes unit-specific time averages and time-period specific cross-sectional averages, which I call the two-way Mundlak (TWM) regression. This equivalence furthers our understanding of the anatomy of TWFE, and has several applications. The equivalence between TWFE and TWM implies that various estimators used for intervention analysis – with a common entry time into treatment or staggered entry, with or without covariates – can be computed using TWFE or pooled OLS regressions that control for time-constant treatment intensities, covariates, and interactions between them. The approach allows considerable heterogeneity in treatment effects across treatment intensity, calendar time, and covariates. The equivalence implies that standard strategies for heterogeneous trends are available to relax the common trends assumption. Further, the two-way Mundlak regression is easily adapted to nonlinear models such as exponential models and logit and probit models.

Acknowledgments: Trang Hoang, Jeff Zabel, and participants at the 15th New York Camp Econometrics provided helpful comments on earlier drafts.

1. Introduction

Panel data structures are used routinely across many fields in attempts to determine causality and estimate the effects of policy interventions. At the micro level, panels are often characterized by a small number of time periods (T) and a large cross section sample size (N). At more aggregated levels, the number of time periods may be substantial, possibly even larger than the cross-sectional dimension.

Regardless of the sizes of T and N, a very common approach to estimating a linear model is to include both unit fixed effects and time fixed effects in ordinary least squares estimation. The resulting estimator is often called the "two-way fixed effects" (TWFE) estimator. As is well known, including unit fixed effects in a linear regression is identical to removing unit-specific time averages and applying pooled ordinary least squares (OLS) to the transformed data. (For this reason, the estimator obtained from including unit-specific dummy variables is often called the *within estimator*.) Including time fixed effects then removes secular changes in the economic environment that have the same effect on all units.

Another important algebraic equivalence involving the FE estimator, usually invoked in microeconometric settings, is the equivalence between the FE estimator that removes unit-specific effects – the one-way FE estimator (OWFE) – and the Mundlak (1978) device, which includes unit-specific time averages of time-varying variables and estimates the resulting equation by random effects (RE). Wooldridge (2019) provides a recent analysis, showing that an entire class of regressions – including pooled OLS – reproduce the FE estimator, even in the unbalanced case, provided one is careful about using only the complete cases in defining the unit-specific time averages. This equivalence has many important applications. For one, it leads to a robust, variable addition version of the Hausman (1978) test

for choosing between FE estimation and random effects (RE) estimation, and, even in the balanced case, it suggests simple hybrid approaches that represent a compromise between FE and traditional RE. A related point is that it makes clear that the pre-testing problem inherent in choosing between RE and FE is virtually the same as pre-testing on a set of regressors. In addition, the equivalence between the one-way FE estimator and the Mundlak regression in the linear case suggests natural ways to allow correlation between explanatory variables and unobserved heterogeneity in nonlinear models when the number of time periods is small. Wooldridge (2019) shows how this can be accomplished even in the context of unbalanced panels.

In exploiting the equivalence between the one-way FE estimator and the Mundlak regression in the small T case, time dummies are usually included among the time-varying covariates because their coefficients can be precisely estimated with a large N. This is the approach taken in Wooldridge (2019). In the balanced case, the time averages of the time dummies are redundant because they all average to $1/T$; in the unbalanced case, the time average generally differs by unit. Nevertheless, the equivalence result is purely algebraic, and so the dimensions of N and T are irrelevant provided they are large enough to actually produce the FE estimates.

In this paper, I explicitly consider the two-way FE estimator and show that a simple extension of the Mundlak device reproduces the TWFE estimates. In particular, adding both the unit-specific time series averages and the period-specific cross-sectional averages in a POLS regression reproduce the two-way FE estimates. I call the regression with the two sets of time averages the *two-way Mundlak* (TWM) regression, and the corresponding estimator the TWM estimator.

The algebraic equivalence between the TWFE and TWM estimators is not too surprising, but it appears to not have been explicitly stated in the literature. The equivalence has several applications. On a basic level, it is valuable to understand the mechanics of commonly used estimation methods. The equivalence of TWFE and TWM emphasizes that accounting for lots of unit and time heterogeneity – by including a full set of two-way "fixed effects" in regression – can be accomplished by using pooled OLS (or random effects) and including covariates of much lower dimension. As a specific application, one can see simple ways to test the basic two-way FE estimator against alternatives that include substantial heterogeneity. I briefly consider this possibility in Section 4.

Another application of the general result is to common timing difference-in-differences (DiD) designs – without or with covariates – and also staggered interventions. In Section 5 I show that, with a common intervention date, a pooled OLS regression that includes an indicator for eventually being "treated," a post-treatment time period dummy, and the treatment indicator – three regressors in addition to an overall constant – is numerically the same as the full TWFE estimator. (The POLS estimator is, in turn, equal to a commonly used difference-in-differences estimator.) A simple extension applies when time-constant covariates are added in a flexible way, showing that several different approaches to estimation – TWFE, pooled OLS, random effects, and standard difference-in-differences – lead to the same place.

There are good reasons for knowing that TWFE can be relied on in intervention analysis with staggered intervention times. For one, we know that TWFE is somewhat resilient to certain kinds of missing data problems – see, for example, Wooldridge (2010, 2019). The equivalence between TWFE and pooled regressions for intervention analysis breaks down in the unbalanced case. One can obtain equivalent pooled OLS regressions, as in Wooldridge

(2019), but it is easier to use TWFE once the proper interaction terms have been constructed, and those come from first studying the balanced case.

Under weak dependence conditions in the time series dimension (with random sampling in the cross-sectional dimension), TWFE has an asymptotic bias on the order of $1/T$ when the covariates violate strict exogeneity; see Wooldridge (2010, Chapter 10). Further, using N, $T \to \infty$ asymptotics with independent sampling in the cross-sectional dimension and weak dependence in the time-series dimension, Hansen (2007) shows that the TWFE estimator is consistent and asymptotically normal without strict exogeneity. In addition, the TWFE estimator is easily extended to allow for heterogeneous trends, and so we can estimate the same set of treatment effects while explicitly allowing for some violations of the common trend assumption. Finally, it is easy to modify the TWFE estimator to allow for spillover effects. For example, a policy change in county i may have an effect in an adjacent county, h, and this is easily handled in a TWFE framework by expanding the treatment variables to allow spillovers.

For staggered interventions, the basic TWFE estimator has come under considerable scrutiny lately – see, for example, de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Borusyak, Jaravel, and Spiess (2021) [BJS (2021)]. In Section 6, I obtain equivalences between an extended TWFE (ETWFE) estimator and pooled OLS regressions in staggered designs with lots of heterogeneity in treatment effects. In doing so, I show how the Sun and Abraham (2021) TWFE approach can be extended to allow covariates to enter flexibly, and I provide the corresponding pooled OLS estimator that controls for different treatment cohorts, calendar time effects, and covariates. Plus, I provide fairly simple arguments to show that, under a

conditional common trends assumption, the ETWFE approach identifies the average treatment effects for different cohort/time period treatment effects. I cover both the case with a never treated group and where all units are treated in the last time period. Moreover, because the POLS/ETWFE estimator is also equivalent to random effects with known (rather than estimated) transformation parameter, under the standard no serial correlation and homoskedasticity assumptions, the POLS estimator is best linear unbiased. Thus, this simpler estimator has the same efficiency as the comparable imputation estimator in BJS (2021).

In Section 7 I show how the algebraic equivalence results lead to simple and easily interpretable tests of the so-called common (or parallel) trends assumption in difference-in-differences settings. In particular, the tests are exclusion restriction tests in a full, unrestricted model, and so insights from pretesting a set of regressors apply directly. I also propose simple extensions to the basic equation that allows violation of parallel trends.

Section 8 contains brief discussions of the unbalanced case, how to allow for heterogeneous trends, and how the DiD-type estimators in the linear case can be extended to nonlinear models, with discussions of binary, fractional, and nonnegative responses. Section 9 contains some concluding remarks. An appendix provides Stata commands for created data sets that illustrate the relative simple mechanics of computing the estimators and proper standard errors.

2. Basics of the Two-Way Fixed Effects Estimator

The typical motivation for the TWFE estimator is an equation of the form

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + f_t + u_{it}, \ t = 1,\ldots,T; \ i = 1,\ldots,N, \tag{2.1}$$

where $\mathbf{x}_{it}$ is $1 \times K$ and $\boldsymbol{\beta}$ is $K \times 1$. The c_i are unit-specific effects (heterogeneity) and f_t are the

time-specific effects. [The panel is assumed to be balanced. I briefly discuss the unbalanced case in Section 8.] We need not take a stand on whether c_i or f_t are properly considered parameters to estimate or as outcomes of random variables: the results in the section are purely algebraic. In fact, there is no need to write down an underlying model; equation (2.1) is purely for motivational purposes.

To describe the TWFE estimator, for each i define a row vector of unit dummy variables as $\mathbf{c}_i = (c1_i, c2_i, \ldots, cN_i)$, where $ch_i = 1$ if $h = i$, $ch_i = 0$ if $h \neq i$. Therefore, for each row indicated by (i,t) pairs, exactly one element of $\mathbf{c}_i$ is equal to unity. Also, the time dummies for period t are $\{fs_t : t = 2, \ldots, T\}$ with $fs_t = 1$ if $s = t$, $fs_t = 0$ if $s \neq t$. We drop the first time period dummy because it is redundant.

The so-called two-way fixed effects estimator, $\hat{\boldsymbol{\beta}}_{FE}$, is obtained as the vector of coefficients on $\mathbf{x}_{it}$ in the pooled OLS regression

$$y_{it} \text{ on } \mathbf{x}_{it}, c1_i, c2_i, \ldots, cN_i, f2_t, \ldots, fT_t, t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{2.2}$$

Along with $\hat{\boldsymbol{\beta}}_{FE}$ we obtain estimates of the so-called unit fixed effects, the coefficients on $c1_i$, $c2_i$, ..., cN_i, and the time fixed effects, the coefficients on $f2_t, \ldots, fT_t$. We are not interested in these coefficients for the purposes of this paper and we do not discuss them further.

A more common way to characterize $\hat{\boldsymbol{\beta}}_{FE}$ is to drop, say, the first unit dummy variable and include an overall intercept:

$$y_{it} \text{ on } \mathbf{x}_{it}, 1, c2_i, \ldots, cN_i, f2_t, \ldots, fT_t, \; t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{2.3}$$

This distinction between (2.2) and (2.3) is unimportant for this paper as they lead to the same $\hat{\boldsymbol{\beta}}_{FE}$.

In the small-T, large-N literature, the time effects are often absorbed into $\mathbf{x}_{it}$, in which case

can study the one-way FE estimator. In the current setup, $\mathbf{x}_{it}$ only includes variables that have some variation across both i and t.

In the large-T panel literature, where one is interested in obtaining valid inference on $\boldsymbol{\beta}$ as $T \to \infty$ (usually along with $N \to \infty$), a "double-demeaning" characterization is used for $\hat{\boldsymbol{\beta}}_{FE}$. See Baltagi (2001). To describe the procedure, define the unit-specific averages over time as

$$\bar{\mathbf{x}}_{i\cdot} = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it} \tag{2.4}$$

and let

$$\bar{\mathbf{x}}_{\cdot t} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_{it} \tag{2.5}$$

be the cross-sectional average for each t. The overall average is

$$\bar{\mathbf{x}} = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it} = N^{-1} \sum_{i=1}^{N} \bar{\mathbf{x}}_{i\cdot} = T^{-1} \sum_{t=1}^{T} \bar{\mathbf{x}}_{\cdot t} \tag{2.6}$$

be the total average. Define

$$\ddot{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot}) - N^{-1} \sum_{i=1}^{N} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot}) = \mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot t} + \bar{\mathbf{x}}, \tag{2.7}$$

As shown in Baltagi (2001), $\hat{\boldsymbol{\beta}}_{FE}$ is the pooled OLS estimator from

$$y_{it} \text{ on } \ddot{\mathbf{x}}_{it}, \; t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{2.8}$$

Alternatively, $\hat{\boldsymbol{\beta}}_{FE}$ is obtained from the POLS regression

$$\ddot{y}_{it} \text{ on } \ddot{\mathbf{x}}_{it}, \; t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{2.9}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$, and this regression also produces the same residuals as the dummy variable regression in (2.2). For most purposes, the R-squared from (2.9) gives a more realistic

measure of goodness-of-fit than (2.2) because it nets out the explanatory power of the unit and time period fixed effects.

This paper is does not formally consider asymptotic analysis under different scenarios, but it is worth noting that the problem of computing valid standard errors can be studied by using one of the two equivalent expressions:

$$\hat{\boldsymbol{\beta}}_{FE} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}' \ddot{y}_{it} \right) = \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} \ddot{\mathbf{x}}_{it}' \ddot{y}_{it} \right). \tag{2.10}$$

After suitable normalization, the first expression is useful under fixed-T, $N \to \infty$ asymptotics with random sampling across i – as in White (1984) and Arellano (1987) – and also when $T \to \infty$ slowly enough along with $N \to \infty$ with random sampling across i and weak dependence across t – as in Hansen (2007). The second expression is useful in $T \to \infty$ setting where cross-sectional dependence is allowed, as in Driscoll and Kraay (1998) and Vogelsang (2012) – again, after suitable normalization by the sample sizes.

3. The Two-Way Mundlak Regression

Mundlak (1978) showed that, with $\mathbf{x}_{it}$ including any explanatory variables that vary only across t, including any time period dummies, the one-way FE estimator can be obtained as a particular GLS estimator by adding the time averages, $\bar{\mathbf{x}}_{i\cdot}$, as additional explanatory variables along with a constant and $\mathbf{x}_{it}$. Wooldridge (2019) showed that an entire class of estimators based on GLS-like transformations are equivalent to FE. The one we focus on in this paper is the pooled OLS estimator:

$$y_{it} \text{ on } 1, \mathbf{x}_{it}, \bar{\mathbf{x}}_{i\cdot}, \; t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{3.1}$$

However, here we want to explicitly separate the time period dummies from the elements of

$\mathbf{x}_{it}$. As in Section 2, henceforth $\mathbf{x}_{it}$ includes only variables that have some variation across both i and t. The key algebraic result in this paper is that the two-way FE estimate, $\hat{\boldsymbol{\beta}}_{FE}$, can be obtained from from an extension of the usual Mundlak regression:

$$y_{it} \text{ on } 1, \mathbf{x}_{it}, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t}, \; t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{3.2}$$

I call this the *two-way Mundlak* (TWM) regression and the coefficients on $\mathbf{x}_{it}$, say, $\hat{\boldsymbol{\beta}}_{M}$, the TWM estimates.

THEOREM 3.1: For a panel data set of dimensions T and N, assume that

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \tag{3.3}$$

is nonsingular, where $\ddot{\mathbf{x}}_{it}$ is defined as in (2.7). Let $\hat{\boldsymbol{\beta}}_{FE}$ be the two-way FE estimate obtained from (2.2) and let $\hat{\boldsymbol{\beta}}_{M}$ be the coefficient on $\mathbf{x}_{it}$ in (3.2). Then

$$\hat{\boldsymbol{\beta}}_{M} = \hat{\boldsymbol{\beta}}_{FE} \tag{3.4}$$

Proof: By the Frisch-Waugh (F-W) partialling out theorem, it suffices to show that the $\ddot{\mathbf{x}}_{it}$ are the (vector) residuals from the pooled regression

$$\mathbf{x}_{it} \text{ on } 1, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t}, \; t = 1, \ldots, T; \; i = 1, \ldots, N \tag{3.5}$$

By another simple application of F-W, the residuals from (3.5) regression can be obtained by removing the means from all variables. The common mean is $\bar{\mathbf{x}}$, and so the desired residuals are obtained from

$$\mathbf{x}_{it} - \bar{\mathbf{x}} \text{ on } \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}} \tag{3.6}$$

It is easily seen that the two sets of regressors in (3.6) are orthogonal in sample:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{\cdot t}-\bar{\mathbf{x}}) = \left[\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\right]'\sum_{t=1}^{T}(\bar{\mathbf{x}}_{\cdot t}-\bar{\mathbf{x}}) = \mathbf{0} \tag{3.7}$$

because both sums are for vectors deviated from the overall mean. It follows that in finding the $K \times K$ matrix of OLS coefficients on each $1 \times K$ vector in (3.6), we can focus on each term separately. We now show that the matrix of OLS coefficients on $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$ is $\mathbf{I}_K$. Let

$$\begin{aligned}
\hat{\mathbf{\Pi}} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\mathbf{x}_{it}-\bar{\mathbf{x}}) \\
&= \left[T\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\right]^{-1}\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\left[\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}})\right] \\
&= \left[\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\right]^{-1}\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\left[T^{-1}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}})\right] \\
&= \left[\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\right]^{-1}\sum_{i=1}^{N}(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})'(\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}}) = \mathbf{I}_K
\end{aligned}$$

because $T^{-1}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}) = (\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})$.

A symmetric argument shows that the $K \times K$ matrix of OLS coefficients on $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ in regression (3.6) is also $\mathbf{I}_K$. Therefore, the residuals from the regression (3.6) are

$$\begin{aligned}
\hat{\mathbf{r}}_{it} &\equiv (\mathbf{x}_{it}-\bar{\mathbf{x}}) - (\bar{\mathbf{x}}_{i\cdot}-\bar{\mathbf{x}})\mathbf{I}_K - (\bar{\mathbf{x}}_{\cdot t}-\bar{\mathbf{x}})\mathbf{I}_K \\
&= \mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot t} + \bar{\mathbf{x}} = \ddot{\mathbf{x}}_{it}. \quad \square
\end{aligned}$$

Also Theorem 3.1 is purely algebraic, it shows that $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$ effectively act as sufficient statistics in accounting for any unit-specific heterogeneity and time-specific heterogeneity that is correlated with $\mathbf{x}_{it}$. Rather than having to include $(N-1)+(T-1)$ control variables, it suffices to include $2K$ control variables, $(\bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t})$.

We can say something even stronger. Once $(\bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t})$ have been included in the regression, adding variables that change only across i, say $\mathbf{z}_i$, or only across t, say $\mathbf{m}_t$, does not affect the

coefficients on $\mathbf{x}_{it}$. Logically, this is satisfying because we know all such variables are eliminated by the TWFE transformation. The following result extends Wooldridge (2019, Proposition 2.1) to the two-way FE setting.

THEOREM 3.2: In the two-way Mundlak regression, include time-constant variables $\mathbf{z}_i$ and time-varying variables $\mathbf{m}_t$:

$$y_{it} \text{ on } 1, \mathbf{x}_{it}, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t}, \mathbf{z}_i, \mathbf{m}_t, t = 1,\ldots,T; i = 1,\ldots,N. \tag{3.8}$$

Let $\tilde{\boldsymbol{\beta}}_M$ be the $K \times 1$ vector of coefficients on $\mathbf{x}_{it}$. Then $\tilde{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}_M$.

Proof: We apply Frisch-Waugh multiple times and show that the residuals from

$$\mathbf{x}_{it} - \bar{\mathbf{x}} \text{ on } \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}, \mathbf{z}_i - \bar{\mathbf{z}}, \mathbf{m}_t - \bar{\mathbf{m}} \tag{3.9}$$

are still $\ddot{\mathbf{x}}_{it}$. The first step is to partial out $(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}})$ from $\mathbf{x}_{it} - \bar{\mathbf{x}}$; from the proof of Theorem 3.1 we know the residuals are $\ddot{\mathbf{x}}_{it}$. Next, we partial $(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}})$ out from both $\mathbf{z}_i - \bar{\mathbf{z}}$ and $\mathbf{m}_t - \bar{\mathbf{m}}$. As in Theorem 3.1, because $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ are orthogonal in sample, we can regress $\mathbf{z}_i - \bar{\mathbf{z}}$ separately on $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ to obtain their (matrix) coefficients. Using the same argument from Theorem 3.1, the coefficients on $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ are zero. Further, the residuals from $\mathbf{z}_i - \bar{\mathbf{z}}$ on $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$, say $\ddot{\mathbf{e}}_i$, depend only on i and $\sum_{i=1}^{N} \ddot{\mathbf{e}}_i = \mathbf{0}$ (because the both the dependent variables and independent variables have been centered about their means). Flipping around the subscripts, the residuals from regressing $\mathbf{m}_t - \bar{\mathbf{m}}$ on $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ gives zero coefficients on the first term and so the residuals depend only on t, say $\ddot{\mathbf{a}}_t$, with $\sum_{t=1}^{T} \ddot{\mathbf{a}}_t = \mathbf{0}$. By Frisch-Waugh, the residuals from (3.9) are the same as those from

$$\ddot{\mathbf{x}}_{it} \text{ on } \ddot{\mathbf{e}}_i, \ddot{\mathbf{a}}_t, t = 1,\ldots,T; i = 1,\ldots,N \tag{3.10}$$

But it is easily seen that $\ddot{\mathbf{x}}_{it}$ is orthogonal, in sample, to each of the regressors:

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{e}}_i'\ddot{\mathbf{x}}_{it} = \sum_{i=1}^{N}\ddot{\mathbf{e}}_i'\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it} = \mathbf{0}$$
$$\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{a}}_t'\ddot{\mathbf{x}}_{it} = \sum_{t=1}^{T}\ddot{\mathbf{a}}_t'\sum_{i=1}^{N}\ddot{\mathbf{x}}_{it} = \mathbf{0}$$

In fact, this shows that the matrix coefficients on $\mathbf{z}_i - \bar{\mathbf{z}}$ and $\mathbf{m}_t - \bar{\mathbf{m}}$ in (3.10) are identically zero, and so the residuals from (3.10) are simply the $\ddot{\mathbf{x}}_{it}$. □

The pooled OLS regression in (3.8) implicitly includes the time averages of all variables that have some time variation because the time averages of $\bar{\mathbf{x}}_{\cdot t}$ and $\mathbf{m}_t$ are simply vectors of constants, and an overall intercept is included in the regression. Therefore, Proposition 2.1 in Wooldridge (2019) applies immediately.

COROLLARY 3.3: The pooled OLS estimator $\tilde{\boldsymbol{\beta}}_M$ from (3.8), and therefore the TWFE estimator, are identical to one-way random effects GLS estimator (with a cross-sectional "random effect") using the same regressors as in (3.8). □

Wooldridge's proposition shows that the equivalence holds if one treats the variances in the typical random effects specification as known, and so POLS is actually best linear unbiased under standard random effects assumptions. Of course it is also asymptotically efficient under those same assumptions; see Wooldridge (2010, Section 10.4).

Theorems 3.1 and 3.2 have some simple but useful implications. Suppose, for example, that an element of $\mathbf{x}_{it}$ can be expressed as an interaction between a time-constant variable and time-varying variable:

$$x_{itj} = z_{ij} \cdot m_{tj} \tag{3.11}$$

Then

$$\bar{x}_{i\cdot j} = z_{ij} \cdot \bar{m}_j, \; \bar{x}_{\cdot tj} = \bar{z}_j m_{tj}, \tag{3.12}$$

where $\bar{m}_j = T^{-1}\sum_{s=1}^{T} m_{sj}$ and $\bar{z}_j = N^{-1}\sum_{h=1}^{N} z_{hj}$. Therefore, the two-way Mundlak regression will include z_{ij} and m_{tj} as separate regressors (because the averages are constant multiples of z_{ij} and m_{tj}). We will use this simple observation in the sections on intervention analysis.

4. Implications for Panel Factor Models

For the purposes of motivation, again start with the model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + f_t + u_{it},\ t = 1,\ldots,T;\ i = 1,\ldots,N. \tag{4.1}$$

The factor model literature replaces the additive terms $c_i + f_t$ with a factor structure, say, $\mathbf{f}_t\mathbf{c}_i$ for $\mathbf{f}_t$ a $1 \times P$ vector of unobserved factors and $\mathbf{c}_i$ a $P \times 1$ vector of unobserved loadings (heterogeneity). We know from Theorem 3.1 that estimating (4.1) by TWFE is identical to estimate the following equation by pooled OLS:

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_{i\cdot}\boldsymbol{\lambda} + \bar{\mathbf{x}}_{\cdot t}\boldsymbol{\xi} + e_{it}, \tag{4.2}$$

with $\hat{\boldsymbol{\beta}}_{FE}$ is the vector on $\mathbf{x}_{it}$. Therefore, one might think of proxying $\mathbf{f}_t\mathbf{c}_i$ by including interactions between $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$, as in

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_{i\cdot}\boldsymbol{\lambda} + \bar{\mathbf{x}}_{\cdot t}\boldsymbol{\xi} + (\bar{\mathbf{x}}_{i\cdot} \otimes \bar{\mathbf{x}}_{\cdot t})\boldsymbol{\pi} + e_{it} \tag{4.3}$$

As a simple test of whether having additive unit and time effects is sufficient, one can test $H_0 : \boldsymbol{\pi} = \mathbf{0}$ after POLS estimation of (4.3). Or, drop $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$ from (4.3) and estimate in by TWFE. If H_0 is rejected, one might even use (4.3) as a simply way of accounting for a factor structure.

Pesaran (2006) suggests alternatives to TWFE in the large T case, with a popular estimator being the pooled common correlated effects (CCEP) estimator. The CCEP estimator involves projecting the $\mathbf{x}_{it}$ onto the cross-sectional averages, $(\bar{\mathbf{x}}_{\cdot t}, \bar{y}_{\cdot t})$, unit-by-unit. The CCEP approach is untenable with small T and moderate K as the unit-specific regressions will provide perfect

fits. By contrast, the regression in (4.3) can be used for small T and large K, and knowing that (4.2) reproduces the TWFE estimator suggests estimating (4.3) can be informative.

One case easily allow heterogeneous slopes that are correlated with $\mathbf{x}_{it}$ by expanding (4.3) even further. A simple approach is to add terms $\mathbf{x}_{it} \otimes (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})$ and $\mathbf{x}_{it} \otimes (\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}})$ (where the centering about the overall mean is done to give $\boldsymbol{\beta}$ an average partial effect interpretation). Such approaches allow for substantial heterogeneity and are not limited by either small T or large K. How one does asymptotic inference requires some care, as there are differences between the small-T, large-N scenarios and situations where $T \to \infty$ asymptotics makes more sense. But under the null that none of the interactions are needed, the estimator collapses to TWFE. Even if the interaction terms are statistically significant, changes in the estimates of $\boldsymbol{\beta}$ could be relatively minor.

5. Application to Interventions with Common Treatment Timing

We now consider a intervention analysis setting with panel data, where, in time periods $q - 1$ and earlier, no units are subject to the treatment or intervention. At $t = q$, some units are subject to the intervention, and the intervention stays in place for the remaining time periods. This particular setup means we can define a treatment indicator as

$$w_{it} = d_i \cdot p_t, \tag{5.1}$$

where d_i is a dummy variable that equals one if unit i was eventually subjected to the intervention and p_t is a dummy variable indicating the post-treatment time periods: $p_t = 0$, $t = 1, \ldots, q - 1$, and $p_t = 1$, $t = q$, ..., T. Note that we can write

$$p_t = fq_t + \cdots + fT_t$$

where fs_t is a dummy variable equal to one of $s = t$ and zero otherwise. In addition to common treatment timing, this setup means there is no reversibility: once a unit is subjected to the intervention it stays in place through time period T. The outcome variable is denoted y_{it}, and we may observe pre-intervention covariates, $\mathbf{x}_i$, which do not change across t.

5.1. Homogeneous Time Effects

For estimating a single policy effect, A simple model underlying causal estimation of the intervention is

$$y_{it} = \beta w_{it} + c_i + g_t + u_{it}, t = 1, \ldots, T; i = 1, 2, \ldots, N, \tag{5.2}$$

where, initially, we estimate a constant effect. To allow w_{it} to be correlated with the unit heterogeneity c_i and to account for the possibility that the timing of the intervention might correspond with secular changes in the aggregate, it is common to use the two-way FE estimator, $\hat{\beta}_{FE}$. The dimensions of T and N, for the purposes of estimation, are essentially unrestricted. The case $T = 2$ is allowed but not necessary.

Viewed from the dummy variable regression perspective, it appears that the two-way FE approach, compared with the simple difference-in-differences approach of comparing the differences in the average changes across time, allows much more flexibility concerning the nature of endogeneity of the intervention. However, we can use Theorem 3.1 to show this is not the case. In fact, including time-constant controls or variables changing only across t do not affect estimation of β. We know that such variables would drop out of TWFE estimation if added. Theorem 3.1 implies that including them in the pooled OLS regression has no effect.

COROLLARY 5.1: Let $\hat{\beta}_{FE}$ be the two-way FE estimator and let $\hat{\beta}_{DD}$ be the coefficient on $w_{it} = d_i \cdot p_t$ from the regression

$$y_{it} \text{ on } 1, w_{it}, d_i, p_t, t = 1,\ldots,T;\ i = 1,\ldots,N \tag{5.3}$$

Provided $\sum_i^N \sum_{t=1}^N \ddot{w}_{it}^2 > 0$, $\hat{\beta}_{DD} = \hat{\beta}_{FE}$. Further, adding time constant variables or cross-sectional constant variables does not affect $\hat{\beta}_{DD}$.

Proof: From Theorem 3.1, it suffices to show that regression (5.3) is the Mundlak regression, in the case of a single covariate, $w_{it} = d_i \cdot p_t$, that has variation across i and t. But notice that the time average of w_{it} is simply

$$\bar{w}_i = d_i \cdot \bar{p}, \tag{5.4}$$

where $\bar{p} = (T - q + 1)/T$ is the fraction of treated periods. Further, for each t the cross sectional average is

$$\bar{w}_t = \bar{d} \cdot p_t, \tag{5.5}$$

where $\bar{d}$ is the fraction of units in the sample that are (eventually) treated. Therefore, the two-way Mundlak regression is

$$y_{it} \text{ on } 1, w_{it}, d_i \cdot \bar{p}, \bar{d} \cdot p_t, t = 1,\ldots,T;\ i = 1,\ldots,N.$$

Compared with the regression in (5.3), the two control variables, d_i and p_t, have been multiplied by constants. This leaves the coefficient on w_{it} unchanged, and so $\hat{\beta}_{DD} = \hat{\beta}_M$. The second part of the claim follows from Theorem 3.2. □

An immediate implication of Corollary 5.1 is that the estimates are unchanged if we replace p_t with a full set of time period dummies, $f2_t, f3_t, \ldots, fT_t$. Moreover, adding time-constant controls, $\mathbf{x}_i$, and also interacting them with the time-constant treatment indicator, d_i, does not change the estimate of β. In other words, the long regression

$$y_{it} \text{ on } 1, w_{it}, d_i, f2_t, \ldots, fT_t, \mathbf{x}_i, d_i \cdot \mathbf{x}_i, t = 1,\ldots,T; i = 1,\ldots,N$$

produces the same coefficient on w_{it} as the short regression in (5.3). This somewhat surprising result appears to have been overlooked in the literature.

The estimated effect does change if we interact $\mathbf{x}_i$ with p_t to allow the effects of the covariates in the untreated state to change over time or interact w_{it} with $\mathbf{x}_i$ to allow the treatment effect to change with $\mathbf{x}_i$. In other words, consider the expanded equation

$$y_{it} = \alpha + \beta w_{it} + [w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma} + \mathbf{x}_i\boldsymbol{\xi} + \zeta d_i + (d_i \cdot \mathbf{x}_i)\boldsymbol{\lambda} + \theta p_t + (p_t \cdot \mathbf{x}_i)\boldsymbol{\delta} + e_{it}, \tag{5.6}$$

where $\boldsymbol{\mu}_1 \equiv E(\mathbf{x}_i|d_i = 1)$ is the average of the covariates over the treated subpopulation. As discussed below, centering $\mathbf{x}_i$ about $\boldsymbol{\mu}_1$ gives β an interesting meaning under standard assumptions: it is the average treatment effect on the treated, ATT. With the addition of $p_t \cdot \mathbf{x}_i$ and $w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ to the TWFE estimation, we now need to add both $\mathbf{x}_i$ and $d_i \cdot \mathbf{x}_i$ to the POLS estimation of (5.6) to ensure POLS is the same as TWFE estimation. Again, this follows from Theorem 3.1, where the time average of $p_t \cdot \mathbf{x}_i$ is proportional to $\mathbf{x}_i$, the the time average of $w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1) = d_i \cdot p_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ is proportional to $d_i \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$, and the cross-sectional averages of $p_t \cdot \mathbf{x}_i$ and $d_i \cdot p_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ consists of terms that are multiples of p_t. For the terms not involving w_{it}, there is no need to center $\mathbf{x}_i$ because doing so does not change estimation of β or $\boldsymbol{\gamma}$.

In obtaining a feasible estimate, the population mean is replaced with the average over the treated units, $\bar{\mathbf{x}}_1 \equiv N_1^{-1} \sum_{i=1}^{N} d_i \cdot \mathbf{x}_i$. Therefore, the POLS version of the estimation is

$$y_{it} \text{ on } 1, w_{it}, w_{it} \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1), \mathbf{x}_i, d_i, d_i \cdot \mathbf{x}_i, p_t, p_t \cdot \mathbf{x}_i, t = 1, \ldots, T; \; i = 1, \ldots, N. \tag{5.7}$$

Comparing (5.7) with (5.3) shows how easy it is to allow substantial heterogeneity in common-timing DiD designs while staying within a simple estimation framework. Whether one implements the method using POLS or extended TWFE is irrelevant because they are

identical.

As a technical point, whether one uses POLS with the controls $\mathbf{x}_i$, d_i, $d_i \cdot \mathbf{x}_i$, and p_t or drops all of these terms and applies TWFE, one should consider adjusting the standard errors to account for the sampling error in $\bar{\mathbf{x}}_1$. If one does not adjust the standard errors, it is the same as conditioning inference on the observed $\mathbf{x}_i$. In practice, the adjustment seems to make only a small difference in the standard errors. Nevertheless, statistical packages such as Stata have a built-in command that computes average partial effects and accounts for sampling error in the sample averages that appear in the APE. In using such commands, it is important to define the treatment indicator, w_{it}, and compute the APE with respect to it and average over the $d_i = 1$ subgroup. The remaining terms involving d_i and p_t effectively act as controls along with the covariates. Incidentally, it is incorrect to replace w_{it} in (5.6) with $d_i \cdot p_t$ and compute the APE with respect to d_i.

When $T = 2$, we can make a connection with the Heckman, Ichimura, and Todd (1997) [HIT (1997)] regression adjustment approach when applied to panel data. It is easily seen when $T = 2$ that the pooled regression in (5.7), which is saturated across all for groups, is identical to estimating four separate linear regression functions: for the control group in $t = 1$ and $t = 2$ and for the (eventually) treated group in $t = 1$ and 2. The coefficient on w_{it} is identical to average the difference-in-differences of the estimated regression functions across the $d_i = 1$ subsample. This is a linear regression version of the estimator proposed by HIT (1997).

We can connect the TWFE and POLS estimation to standard difference-in-differences estimation by noting further algebraic equivalences. First, without covariates, one can show that the coefficient on w_{it} in regression (5.3) can be written as

$$\hat{\beta} = N_1^{-1} \sum_{i=1}^{N} d_i \cdot \Delta\bar{y}_i - N_0^{-1} \sum_{i=1}^{N} (1 - d_i) \cdot \Delta\bar{y}_i,$$

where

$$\Delta\bar{y}_i \equiv (T - q + 1)^{-1} \sum_{t=q}^{T} y_{it} - (q - 1)^{-1} \sum_{t=1}^{q-1} y_{it} = \bar{y}_{i,post} - \bar{y}_{i,pre} \tag{5.8}$$

is the change in means for unit i from the pre-treatment period to the post-treatment period. In other words, we can collapse the data to a simple cross section consisting of $\{(\Delta\bar{y}_i, d_i) : i = 1, \ldots, N\}$ and then $\hat{\beta}$ is the simple difference-in-means estimator using $\Delta\bar{y}_i$ as the outcome variable. In the $T = 2$ case, we have the usual difference $\Delta\bar{y}_i = \Delta y_i = y_{i2} - y_{i1}$. This characterization of $\hat{\beta}$ is useful in situations where N is small but we are willing to assume that the equation $\Delta\bar{y}_i = \alpha + \beta d_i + u_i$ roughly satisfies the classical linear model assumptions (CLM)– see Wooldridge (2019, Chapter 4) – so that inference about β can be based on the $\mathcal{T}_{N-2}$ distribution. Under the CLM assumptions, we could even have a single treated unit along with as few as two control units. If T is reasonably large and weak dependence holds across t, the central limit theorem helps ensure $\Delta\bar{y}_i$ has an approximate normal distribution.

When covariates are added, as in (5.6) or the TWFE equivalent, the estimate of $\hat{\beta}$ is obtained by applying standard regression adjustment (RA) to the cross sectional data $\{(\Delta\bar{y}_i, d_i, \mathbf{x}_i) : i = 1, \ldots, N\}$. In particular, suppose we estimate separate linear regression functions (for $d_i = 0$ and $d_i = 1$) with dependent variable $\Delta\bar{y}_i$ and covariates $\mathbf{x}_i$ and call the resulting "treatment effect" estimator $\hat{\tau}$. Then $\hat{\tau} = \hat{\beta}$ from (5.7) (and, therefore, its TWFE equivalent). Equivalently, $\hat{\beta}$ obtained from the regression

$$\Delta\bar{y}_i \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1), i = 1, \ldots, N. \tag{5.9}$$

When $T = 2$ the regression in (5.9) becomes

$$\Delta y_i \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1), i = 1, \ldots, N, \tag{5.10}$$

which is numerically identical to TWFE. As discussed above, this regression produces a parametric version of the HIT (1997) estimator for panel data. Below we will discuss implications for applying other treatment effects esimators, including doubly robust estimators that combine regression adjustment and propensity score weighting.

5.2. Heterogeneous Time Effects

We can easily find equivalent estimators in settings where the treatment effect is allowed to vary by time and even by both time and covariates. Without covariates, consider

$$y_{it} = \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + c_i + g_t + u_{it}, \tag{5.11}$$

where, again, fr_t is a dummy variable equal to unity of $r = t$ and zero otherwise. This specification allows the policy effect to be different in each of the treated periods. As usual, the two-way FE estimator removes c_i and g_t. In showing the equivalent pooled OLS regression based on Theorem 3.1, note that, for $r = q, \ldots, T$,

$$w_{it} \cdot fr_t = d_i \cdot p_t \cdot fr_t = d_i(fq_t + \cdots + fT_t)fr_t = d_i fr_t \tag{5.12}$$

because $fr_t fs_t = 0$, $r \neq t$, and $fr_t fr_t = fr_t$. Therefore, equation (5.11) is is equivalent to

$$y_{it} = \beta_q(d_i \cdot fq_t) + \cdots + \beta_T(d_i \cdot fT_t) + c_i + g_t + u_{it}$$

The time averages of $d_i \cdot fr_t$ are simply d_i/T for all r and the cross-sectional averages are $\bar{d} \cdot fr_t$, where, again, $\bar{d}$ is the fraction of treated units in the sample. Therefore, the two-way Mundlak equation involves adding d_i and the period dummies fq_t, ..., fT_t. Note that it is no longer sufficient to add just the post-period dummy, p_t, to reproduce TWFE. Therefore, estimating the following equation by POLS is, for the purposes of obtaining $\hat{\beta}_q$, $\hat{\beta}_{q+1}$, ..., $\hat{\beta}_T$,

equivalent to TWM:

$$\begin{aligned} y_{it} &= \alpha + \beta_q(d_i \cdot fq_t) + \cdots + \beta_T(d_i \cdot fT_t) + \zeta d_i + \theta_q fq_t + \cdots + \theta_T fT_t + e_{it} \\ &= \alpha + \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + \zeta d_i + \theta_q fq_t + \cdots + \theta_T fT_t + e_{it} \end{aligned} \tag{5.13}$$

Moreover, estimating this equation by random effects (in the cross-sectional dimension) gives exactly the same estimates.

In estimating (5.13), one can test whether the β_r are constant across r. Equivalently, w_{it} can be included by itself so its coefficient is the base estimate for the first period, q, and then the coefficients on the interactions $w_{it} \cdot fr_t = d_i \cdot fr_t$, $r = q + 1, ..., T$, become the differences with period q. It follows from Theorem 3.2 that including the time period dummies before period q does not change the $\hat{\beta}_r$. In fact, Theorem 3.2 implies that including additional variables that change only across i or only across t will not change the estimates.

It also can be shown that the $\hat{\beta}_r$ are obtained by separate DiD analyses by collapsing the $q - 1$ pre-periods into a single control period and then using periods q, $q + 1$, ..., T in turn as treatment periods. This is also the same as defining, for each i and $r \in \{q, \ldots, T\}$, the difference between y_{ir} and the pre-intervention average,

$$\dot{y}_{ir} \equiv y_{ir} - (q-1)^{-1} \sum_{t=1}^{q-1} y_{it} = y_{ir} - \bar{y}_{i,pre}, \tag{5.14}$$

and then computing

$$\hat{\beta}_r = N_1^{-1} \sum_{i=1}^{N} d_i \dot{y}_{ir} - N_0^{-1} \sum_{i=1}^{N} (1 - d_i) \dot{y}_{ir}$$

We can include time constant controls, $\mathbf{x}_i$, although, as putting them in additive form with constant coefficient has no effect – again a consequence of Theorem 3.2. A flexible specification also allows the policy effects to change with $\mathbf{x}_i$, as in the equation

$$y_{it} = \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + [w_{it} \cdot fq_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_q + \cdots + [w_{it} \cdot fT_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_T \quad (5.15)$$
$$+ (fq_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_q + \cdots + (fT_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_T + c_i + g_t + u_{it}$$

In addition to allowing separate treatment effects in each time period, and TEs that can also change with $\mathbf{x}_i$, this equation allows the controls $\mathbf{x}_i$ to have different effects on the control units in the different time periods. As before, $\boldsymbol{\mu}_1$ would be replaced with $\bar{\mathbf{x}}_1 = N_1^{-1}\sum_{i=1}^{N} d_i \cdot \mathbf{x}_i$.

Two-way fixed effects estimation of (5.15) is straightforward. Along with the $\hat{\beta}_r$ one can see how the effect policy changes with $\mathbf{x}_i$ and across time, or both. A slightly more convenient formulation for testing is to replace $w_{it} \cdot fq_t$ with w_{it} and $w_{it} \cdot fq_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ with $w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ and then the remaining coefficients are compared relative to the first intervention period. Compared with the usual TWFE estimator, which replaces all terms in (5.15) with the single variable w_{it}, the extended version allows much more flexibility. This highlights the important point that there is nothing inherently wrong with TWFE, which is an estimation method. The problem with how TWFE is implemented in DiD settings is that it is applied to a restrictive model.

We can, of course, impose restrictions on (5.15), which may be desirable if N is not large. To impose a constant effect across time, but allow the effect to depend on $\mathbf{x}_i$, include only w_{it} and $w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ and not interactions with the time dummies. Hybrids are possible, too, such as including $w_{it} \cdot fq_t$, ..., $w_{it} \cdot fT_t$ but only including interactions between the treatment and covariates, $w_{it} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$, and imposing the same coefficients in all time periods.

What is the Mundlak regression corresponding to (5.15)? As before, d_i must be included, as must be fq_t, ..., fT_t. The time average of $fr_t \cdot \mathbf{x}_i$ is simply $\mathbf{x}_i/T$, and so we must include $\mathbf{x}_i$ on its own. The cross-sectional average is $fr_t \cdot \bar{\mathbf{x}}$, which means we just need to include the period dummies, fq_t, ..., fT_t – which we already knew. The time averages of $d_i \cdot fr_t(\mathbf{x}_i - \boldsymbol{\mu}_1)$ is

$d_i \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)/T$, and so $d_i \cdot \mathbf{x}_i$ must appear. Again, the cross-sectional averages are just constant multiples of fr_t. The two-way Mundlak equation therefore looks like

$$y_{it} = \alpha + \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_T(w_{it} \cdot fT_t) + [w_{it} \cdot fq_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_q + \cdots + [w_{it} \cdot fT_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)]\boldsymbol{\gamma}_T \quad (5.16$$
$$+ (fq_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_q + \cdots + (fT_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_T + \zeta d_i + \mathbf{x}_i\boldsymbol{\xi} + (d_i \cdot \mathbf{x}_i)\boldsymbol{\lambda} + \theta_q fq_t + \cdots + \theta_T fT_t + e_{it}$$

and we estimate this by pooled OLS. Again, as discussed at the end of Section 3, RE estimation of this same equation produces identical estimates. With large N and small T, we can try to improve efficiency beyond the POLS/RE estimator by using a feasible GLS procedure on (5.16) that allows any pattern of changing variances or serial correlation. Because there still could be heteroskedasticity as a function of $(d_i, \mathbf{x}_i)$, we should use fully robust inference because any such procedure should be viewed as quasi-GLS.

As in the case without covariates, estimating (5.16) by POLS is the same as analyzing each post-treatment period separately with the first $q - 1$ periods collapsed into a single control period. One way to obtain the estimates is to use the panel data in periods 1, 2, ..., $q - 1$ and a given $r \in \{q, q + 1, \ldots, T\}$ and estimate the equation

$$y_{it} = \alpha + \beta_r w_{it} + w_{it} \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1)\boldsymbol{\gamma}_r + (fr_t \cdot \mathbf{x}_i)\boldsymbol{\delta}_r + \zeta d_i + d_i \cdot \mathbf{x}_i\boldsymbol{\lambda} + \theta_r fr_t + error_{it} \quad (5.17)$$

by POLS. Equivalently, drop $\mathbf{x}_i$, d_i, and $d_i \cdot \mathbf{x}_i$ and use two-way FE.

In order to use something other than regression adjustment, it is useful to know that $\hat{\beta}_r$ is obtained by defining $\dot{y}_{ir}$ as in (5.14) and then using regression adjustment (RA) on the resulting cross section, where coefficients on $\mathbf{x}_i$ are allowed to vary across $d_i = 0$ and $d_i = 1$. The OLS regression would be

$$\dot{y}_{ir} \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_1), i = 1, \ldots, N \quad (5.18)$$

and then $\hat{\tau}_r$ is the coefficient on d_i. As usual, built-in features of packages such as Stata make

it easy to adjust the standard errors for sampling error in $\bar{\mathbf{x}}_1$.

5.3. What is Being Estimated?

Estimating a flexible equation such as (5.15) by TWFE or, equivalently, (5.16) by POLS or RE, to obtain the effects of policy intervention intuitively seems like a good idea, especially given that these include much more heterogeneity in the treatment effects compared with the constant-coefficient TWFE estimator. From a traditional perspective it is easy to figure out what the coefficients on the treatment indicators are consistently estimating. For example, take the equation without covariates but allowing time-varying treatment effects:

$$E(y_t|d) = \alpha + \beta_q(d \cdot fq_t) + \cdots + \beta_T(d \cdot fT_t) + \zeta d + \theta_q fq_t + \cdots + \theta_T fT_t$$

Setting $d = 0$ gives

$$\begin{aligned} E(y_t|d = 0) &= \alpha, t = 1, \ldots, q-1 \\ &= \alpha + \theta_t, t = q, \ldots, T \end{aligned}$$

and setting $d = 1$ gives

$$\begin{aligned} E(y_t|d = 1) &= \alpha + \zeta, t = 1, \ldots, q-1 \\ &= \alpha + \zeta + \theta_t + \beta_t, t = q, \ldots, T \end{aligned}$$

Therefore, for given $r \geq q$ and any $t < q$,

$$\begin{aligned} \beta_r &= [E(y_r|d = 1) - E(y_r|d = 0)] - [E(y_t|d = 1) - E(y_t|d = 0)] \\ &= [E(y_r|d = 1) - E(y_t|d = 1)] - [E(y_r|d = 0) - E(y_t|d = 0)] \end{aligned} \tag{5.19}$$

which is a population difference-in-differences expression.

In modern discussions of intervention analysis, there is more interest in determining when parameters such as the β_r can be interepreted as average treatment effects in a setting with explicitly potential outcomes. Here I characterize what is identified when we use an extension of the TWFE estimator that allows for time-varying treatment effects and for flexibly including

covariates, $\mathbf{x}_i$. The underlying sampling scenario is random sampling across i, so we can state assumptions in terms of the population.

We take the T time periods, $t = 1,\ldots,T$, as given. Randomness comes from assignment to the control ($d = 0$) or treatment ($d = 1$) groups. The initial period, at the least, is a control period. The intervention occurs at time q for all treated units and any treated unit remains treated through period T. Because of the simple intervention pattern, we can denote the potential outcomes are $\{[y_t(0), y_t(1)] : t = 1,\ldots,T\}$, where $y_t(0)$ is the outcome in the control state and $y_t(1)$ is the outcome in the treated state, where the treatment is defined by d. We observe either $y_t(0)$ or $y_t(1)$, but not both.

For each t, the treatment effect (due to the intervention occuring at $t = q$) is

$$te_t = y_t(1) - y_t(0) \tag{5.20}$$

Because the intervention occurs at $t = q$, we have no way of identifying average treatment effects prior to the intervention. In fact, below we will assume such effects are zero. The focus is naturally on average effects for $t \geq q$. In particular, we are interested in the average treatment effect on the treated in each of the treated time periods:

$$\tau_t \equiv E[y_t(1) - y_t(0) | d = 1], t = q, q+1, \ldots, T \tag{5.21}$$

I start with the case without covariates as it is more easily seen how identification works. There are two assumptions that restrict the relationship between potential outcomes and the treatment assignment. The first has been called the "no anticipation" assumption, which, in the panel data literature, is technically similar to a "strict exogeneity" assumption. There are two implications of the assumption. First, because the potential outcomes before $t = q$ are defined in terms of the treatment d determined later, it means that units do not change their behavior in

anticipation of the treatment in ways that would affect the outcome. Another implication is that the mechanism used in deciding the treatment and control groups does not base it on systematic differences in potential outcomes prior to the intervention. When $T = 2$, Heckman, Ichimura, and Todd (1997) explicitly state the assumption as $y_1(1) = y_1(0)$ and the extension here is

$$y_t(1) = y_t(0), t < q \tag{5.22}$$

As in Callaway and Sant'Anna (2021) and Sun and Abraham (2021), we can get by with a weaker assumption.

Assumption NA (**No Anticipation**): For $t < q$,

$$E[y_t(1) - y_t(0)|d = 1] = 0. \square \tag{5.23}$$

Notice that the quantities in (5.23) are the ATTs *prior* to the intervention. As pointed out by Callaway and Sant'Anna (2021), with more than one pre-intervention period the NA assumption can be relaxed by allowing (5.23) to fail, say, with $t = q - 1$. For example, maybe units just prior to the intervention change their behavior in anticipation but that does not happen two periods prior. In that case, one can drop the $t = q - 1$ period is dropped from the analysis. In what follows I use the assumption as stated in (5.23); occasionally, for the purposes of simplifying the discussion, I will act as if (5.22) holds.

The second restriction is a standard version of the common (or parallel) trends assumption; it is used by Heckman, Ichimura, and Todd (1997), Abadie (2005), and Sant'Anna and Zhao (2021) in the $T = 2$ setup.

Assumption CT (**Common Trend**): With the (eventually) treated indicator given by d,

$$E[y_t(0) - y_1(0)|d] = E[y_t(0) - y_1(0)] \equiv \theta_t, \; t = 2, \ldots, T. \square \tag{5.24}$$

The assumption is given by the first equality, which says that the average trend in the control state, in every period relative to the initial period, does not depend on treatment status. Once we assume a common trend, we define it to be θ_t. It is easily seen that an equivalent assumption is to impose common trends across any two periods:

$$E[y_t(0) - y_{t-1}(0)|d] = E[y_t(0) - y_{t-1}(0)],\ t = 2,\ldots,T.$$

Now write the observed outcome as

$$y_t = y_t(0) + d \cdot [y_t(1) - y_t(0)] = y_t = y_t(0) + d \cdot te_t \tag{5.25}$$

and take the expectation conditional on d:

$$\begin{aligned} E(y_t|d) &= E[y_t(0)|d] + d \cdot E(te_t|d) \\ &= E[y_t(0)|d] + d \cdot \tau_t \end{aligned} \tag{5.26}$$

because $d \cdot E(te_t|d) = d \cdot E(te_t|d = 1)$. [This follows from the simple representation $E(te_t|d) = (1 - d) \cdot E(te_t|d = 0) + d \cdot E(te_t|d = 1)$.] Next, write $y_t(0)$ in terms of its initial period and the gain over the period:

$$\begin{aligned} y_t(0) &= y_1(0) + g_t(0),\ t = 2,\ldots,T \\ g_t(0) &\equiv y_t(0) - y_1(0) \end{aligned} \tag{5.27}$$

Now we impose the CT assumption, which can be written as

$$E[g_t(0)|d] = E[g_t(0)] \equiv \theta_t,\ t = 2,\ldots,T \tag{5.28}$$

Finally, because d is binary, we can always write

$$E[y_1(0)|d] = \lambda + \xi d \tag{5.29}$$

Combining (5.26) through (5.29) gives, with $\theta_1 \equiv 0$,

$$E(y_t|d) = \lambda + \xi d + \theta_t + d \cdot \tau_t,\ t = 1,\ldots,T \tag{5.30}$$

Under Assumption NA, $\tau_t = 0$ for $t < q$, and so

$$\begin{aligned} E(y_t|d) &= \lambda + \xi d + \theta_t, t < q \\ &= \lambda + \xi d + \theta_t + d \cdot \tau_t, t = q, \ldots, T \end{aligned} \tag{5.31}$$

It is easily seen that these equations identified all of the parameters – particularly the τ_t. The number of parameters in (5.30) is $2 + (T-1) + (T-q+1) = 2(T+1) - q$. There are $2T$ cell means corresponding the two treatment groups over T time periods. If $q = 2$ – its smallest allowable value – then there are the same number of parameters as means. Essentially, this is a way of seeing that we cannot test the combination of Assumptions NA and CT when $q = 2$. When $q > 2$, the NA and CT assumptions impose $q - 2$ restrictions on the $2T$ cell means determined by (d,t).

For implementation, it is useful to write the estimating equation in terms of time dummies as

$$E(y_t|d) = \lambda + \xi d + \theta_2 f2_t + \cdots + \theta_T fT_t + \tau_q(d \cdot fq_t) + \cdots + \tau_T(d \cdot fT_t), t = 1, \ldots, T \tag{5.32}$$

Equivalently, for a random draw i and letting $w_{it} = d_i \cdot p_t$, we can write

$$E(y_{it}|d_i) = \lambda + \xi d_i + \theta_2 f2_t + \cdots + \theta_T fT_t + \tau_q(w_{it} \cdot fq_t) + \cdots + \tau_T(w_{it} \cdot fT_t), t = 1, \ldots, T \tag{5.33}$$

Given a random sample of size N from the population, consistent estimators of τ_r are obtained from the pooled OLS regression

$$y_{it} \text{ on } 1, d_i, f2_t, \ldots, fT_t, w_{it} \cdot fq_t, \ldots, w_{it} \cdot fT_t, t = 1, \ldots, T;\ i = 1, \ldots, N \tag{5.34}$$

From Theorem 3.1, the regression in (5.34) produces estimates of the $\hat{\tau}_t$ equivalent to the TWFE estimator. Moreover, we need only include the time-averages and cross-sectional averages of $d_i fr_t$, $r = q, \ldots, T$, which means that only the time dummies $fq_t, \ldots, fT$ are needed in (5.34) along with d_i. This algebraic equivalance is useful for testing the common trend

assumption, something I take up in Section 7.

We can relax the common trend assumption if we have some constant (pre-intervention) covariates, and also allow the treatment effects to depend on the covariates. We also need to modify the no anticipation assumption so that the pre-intervention treatment effects are zero for all subpopulations defined by **x**.

Assumption CNA (**Conditional No Anticipation**): For treatment indicator d and covariates **x**,

$$E[y_t(1) - y_t(0)|d = 1, \mathbf{x}] = 0, t < q. \square \tag{5.35}$$

As before, this assumption is satisfied if we assume $y_t(1) = y_t(0), t < q$.

Assumption CCT (**Conditional Common Trends**): For treatment indicator d and covariates **x**,

$$E[y_t(0) - y_1(0)|d, \mathbf{x}] = E[y_t(0) - y_1(0)|\mathbf{x}],\ t = 2, \ldots, T. \square \tag{5.36}$$

To derive an estimating equation, again write y_t as in (5.25), but now take the expectation conditional on $(d, \mathbf{x})$:

$$E(y_t|d, \mathbf{x}) = E[y_t(0)|d, \mathbf{x}] + d \cdot E(te_t|d, \mathbf{x})$$

Now, define the ATT conditional on **x** as

$$\tau_t(\mathbf{x}) \equiv E(te_t|d = 1, \mathbf{x})$$

so that

$$E(y_t|d, \mathbf{x}) = E[y_t(0)|d, \mathbf{x}] + d \cdot \tau_t(\mathbf{x}) \tag{5.37}$$

By iterated expecations, $\tau_t = E[\tau_t(\mathbf{x})|d = 1]$.

Next, writing $y_t(0)$ as in (5.27),

$$E[y_t(0)|d,\mathbf{x}] = E[y_1(0)|d,\mathbf{x}] + E[g_t(0)|d,\mathbf{x}],\ t = 2,\ldots,T$$

and by the CCT assumption $E[g_t(0)|d,\mathbf{x}] = E[g_t(0)|\mathbf{x}]$,

$$E[y_t(0)|d,\mathbf{x}] = E[y_1(0)|d,\mathbf{x}] + E[g_t(0)|\mathbf{x}] \tag{5.38}$$

and so

$$E(y_t|d,\mathbf{x}) = E[y_1(0)|d,\mathbf{x}] + E[g_t(0)|\mathbf{x}] + d \cdot \tau_t(\mathbf{x}) \tag{5.39}$$

In principle, there are no restrictions on $E[y_1(0)|d,\mathbf{x}]$ or $E[g_t(0)|\mathbf{x}]$. Here we assume linearity to lead to a simple analysis.

Assumption LIN (**Linear in Parameters**): For covariates $\mathbf{x}$, which many include any functions of underlying control variables,

$$E[y_1(0)|d,\mathbf{x}] = \eta + \lambda d + \dot{\mathbf{x}}\boldsymbol{\kappa} + (d \cdot \dot{\mathbf{x}})\boldsymbol{\zeta} \tag{5.40}$$

$$E[g_t(0)|\mathbf{x}] = \theta_t + \dot{\mathbf{x}}\boldsymbol{\pi}_t,\ t = 2,\ldots,T \tag{5.41}$$

$$\tau_t(\mathbf{x}) = \tau_t + \dot{\mathbf{x}}\boldsymbol{\rho}_t,\ t = q,\ldots,T \tag{5.42}$$

where

$$\dot{\mathbf{x}} \equiv \mathbf{x} - E(\mathbf{x}|d = 1) \equiv \mathbf{x} - \boldsymbol{\mu}_1.\ \square \tag{5.43}$$

Notice that by centering $\mathbf{x}$ about $E(\mathbf{x}|d = 1)$ we force the intercept in the equation for $\tau_t(\mathbf{x})$ to be the parameter of interest, τ_t.

Substituting the conditional mean expressions into We can now obtain an estimating equation:

$$E(y_t|d,\mathbf{x}) = \eta + \lambda d + \dot{\mathbf{x}}\boldsymbol{\kappa} + (d \cdot \dot{\mathbf{x}})\boldsymbol{\zeta} + \theta_t + \dot{\mathbf{x}}\boldsymbol{\pi}_t + \tau_t d + (d \cdot \dot{\mathbf{x}})\boldsymbol{\rho}_t,\ t = 1,\ldots,T \tag{5.44}$$

For $t < q$, Assumption NA implies $\tau_t = 0$, $\boldsymbol{\rho}_t = \mathbf{0}$. As before, we can combine the expressions

into one that is the basis for pooled OLS estimation, adding an i subscript for emphasis (and setting $\theta_1 = 0$):

$$\begin{aligned} E(y_{it}|d_i, \mathbf{x}_i) = \eta + \lambda d_i + \dot{\mathbf{x}}_i\boldsymbol{\kappa} + (d_i \cdot \dot{\mathbf{x}}_i)\boldsymbol{\zeta} + \theta_2 f2_t + \cdots + \theta_T fT_t \\ + (f2_t \cdot \dot{\mathbf{x}}_i)\boldsymbol{\pi}_2 + \cdots + (fT_t \cdot \dot{\mathbf{x}}_i)\boldsymbol{\pi}_T + \tau_q(d_i \cdot fq_t) + \cdots + \tau_T(d_i \cdot fT_t) \\ + (d_i \cdot fq_t \cdot \dot{\mathbf{x}}_i)\boldsymbol{\rho}_q + \cdots + (d_i \cdot fT_t \cdot \dot{\mathbf{x}}_i)\boldsymbol{\rho}_T \end{aligned} \tag{5.45}$$

As a consequence of Theorem 3.1, we only have to include the time dummies and their interactions with $\dot{\mathbf{x}}_i$ for periods q and later, and so the pooled OLS regression is

$$\begin{aligned} y_{it} \text{ on } 1, d_i, \dot{\mathbf{x}}_i, d_i \cdot \dot{\mathbf{x}}_i, fq_t, ..., fT_t, fq_t \cdot \dot{\mathbf{x}}_i, ..., fT_t \cdot \dot{\mathbf{x}}_i, \\ d_i \cdot fq_t, ..., d_i \cdot fT_t, d_i \cdot fq_t \cdot \dot{\mathbf{x}}_i, ..., d_i \cdot fT_t \cdot \dot{\mathbf{x}}_i, \end{aligned} \tag{5.46}$$

where the regression is operationalized by now taking $\dot{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}_1$. As before, one should, technically, should correct the standard errors due to the sampling variation in $\bar{\mathbf{x}}_1$. From Theorem 3.1, the pooled OLS regression produces estimates on all coefficients identical to the TWFE estimates where we drop the time-constant variables d_i, $\dot{\mathbf{x}}_i$, and $d_i \cdot \dot{\mathbf{x}}_i$. The POLS formulation has the benefit of producing coefficients on d_i, $\dot{\mathbf{x}}_i$, and $d_i \cdot \dot{\mathbf{x}}_i$ to determine that nature of selection intro treatment and to see whether the covariates have the anticipated signs. Also, for estimating the treatment effects it is only necessary to center the covariates when they are interacted with the treatment indicators $d_i \cdot fr_t$, $r = q, \ldots, T$, but centering the covariates in all interactions gives the coefficients on d_i and fq_t, ..., fT_t sensible interpretations; the coefficient on d_i allows one to study the nature of the selection bias into exposure.

One can typically use built-in software commands to estimate the ATTs and account for the sampling error in $\bar{\mathbf{x}}_1$. Stata's "margins" command can be used for these purposes. Then the regression should be run explicitly with the time-varying treatment indicator, $w_{it} = d_i \cdot p_t$, and without centering the covariates. Using the simple fact that $d_i \cdot fr_t = w_{it} \cdot fr_t$, the regression

without centering the covariates is

$$y_{it} \text{ on } 1, d_i, \mathbf{x}_i, d_i \cdot \mathbf{x}_i, fq_t, ..., fT_t, fq_t \cdot \mathbf{x}_i, ..., fT_t \cdot \mathbf{x}_i, \\ w_{it} \cdot fq_t, ..., w_{it} \cdot fT_t, w_{it} \cdot fq_t \cdot \mathbf{x}_i, ..., w_{it} \cdot fT_t \cdot \mathbf{x}_i \tag{5.47}$$

Then, the average partial effect is computed with respect to w and with the appropriate time dummy turned on to reflect the different post-treatment periods. Then, the averaging is over the treated subsample, $d_i = 1$. The appendix provides sample Stata commands. Because the covariates have not been centered before creating the interactions, the usual warning applies: the coefficients mostly will be meaningless, as the estimated effects will be for the subpopulation determined by $\mathbf{x} = \mathbf{0}$ (which is often impossible or, at a minimum, not very interesting).

6. Application to Staggered Interventions

We now turn to the more complicated setting of staggered interventions. As in Section 5, the first time of entry is q, but only some units are subjected to the intervention during period q. At period $q + 1$, more units join the treated group, and so on, until period T. We assume that $q > 1$ so there is at least one untreated period. We still assume that the treatment is irreversible, as in Callaway and Sant'Anna (2021), Sun and Abraham (2021), Borusyak, Jaravel, and Spiess (2021), and other work.

The TWFE estimator that imposes a constant effect has come under much scrutiny lately with staggered interventions. Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœuille (2021) use different characterizations of the parameter identified by the simple two-way FE estimator – when a single coefficient on the time-varying treatment indiciator, w_{it}, is estimated. Goodman-Bacon (2021) shows that the estimand can be written as a weighted average of several different 2 × 2 DiD parameters, some of which make no sense (for example, using an

earlier treated group as a control for a later treated group). de Chaisemartin and D'Haultfœuille (2021) write the estimated as a weighted average of certain causal effects and, in this representation, where some of the weights can be negative. Here I show how the usual TWFE estimator can be made much more flexible and, under standard parallel trend assumptions, identifies interesting average treatment effects. Essentially, I provide a regression-based alternative to Callaway and Sant'Anna (2021) under a slightly different but similar set of assumptions. One of the important conclusions is that there is nothing inherently wrong with TWFE as an estimation method. The problem is that it is often applied to a model that is too restrictive.

6.1. Assumptions and Estimating Equations

One way to view the staggered intervention is that it generates different levels of exposure to the treatment, as determined by the date of the intervention. The earlier a unit is subjected to the intervention, the longer the exposure in later time periods. We define treatment cohort dummies, d_q, ..., d_q, that indicate when a unit was first subjected to the intervention. Given a never treated group, there are $T - q + 2$ total treatment levels. The general discussion in what follows acts as if there is a never-treated cohort, although that is not necessary. However, as I discuss further in Section 6.7, if all units are eventually treated then the ATT of the group that enters in the final period is not identified.

Given that the staggered entry leads to different treatment effect intensities, it is natural to think in terms of an expanded set of potential outcomes. The notation here is a bit tricky because there are multiple treatment levels determined by when a unit is initially exposed to the intervention. For $r \in \{q, q+1, \ldots, T\}$, $y_t(r)$ $y_t(r)$ is the potential outcome during time period t if a unit enters the treated state in time period r. We also need a notation for the

potential outcome if a unit is not subjected to the intervention during the time periods under study, $\{1, 2, \ldots, T\}$. One possibility is $y_t(0)$, which, in the common timing case, simply denotes the no treated state. This notation has some drawbacks because then when we write $y_t(r)$ the amount of time in treatment is not decreasing in r. A more subtle issue is that there is a difference between the outcome in the untreated state and the never treated state (although the no anticipation assumption will essentially assume these are the same). Therefore, I borrow the notation from Athey and Imbens (2021) and use $y_t(\infty)$ to denote the potential outcome in time t if a unit is never treated (in the period under study). A notation such as $y_t(T+)$ might be technically more precise but is less appealing.

I initially focus on the treatment effects that are standard in the staggered assignment literature [Athey and Imbens (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)], the difference in potential outcomes first receiving treatment in period r and never receiving it:

$$te_t(r) = y_t(r) - y_t(\infty), r = q, \ldots, T \tag{6.1}$$

[Below I will discuss other treatment effects, such as $y_t(r) - y_t(r+1)$ for $t \geq r$, which is the incremental effect of having been first exposed in period r compared with period $r+1$]

As in Callaway and Sant'Anna (2021) and Sun and Abraham (2021), the treatment effects we hope to identify are the ATTs in periods where the cohorts are actually subjected to the intervention:

$$\tau_{rt} \equiv E[te_t(r)|d_r = 1],\ r = q, \ldots, T;\ t = r, \ldots, T. \tag{6.2}$$

The requirement that we have one untreated period means $q \geq 2$. If a unit enters treatment in period r then we can hope to estimate ATTs for this treatment cohort in periods

$\{r, r+1, \ldots, T\}$. If there is no new entry into treatment in some periods then some of the $\tau(r)_t$. If there all units are treated by period T then the τ_{rT}, $r = q, \ldots, T$ are not identified, although other treatment effects can be under suitable assumptions.

Given the analysis in the common timing case, it is clear we need to rule out anticipatory effects.

Assumption NA (**No Anticipation**, **Staggered**): For treatment cohorts $r = q,\ q+1,\ \ldots, T$,

$$E[y_t(r) - y_t(\infty)|\mathbf{d}] = 0,\ t < r.\ \square \tag{6.3}$$

The stronger form of NA, $y_t(r) = y_t(\infty)$ for $t < r$, means that regardless of when a unit is first exposed to the intervention the potential outcomes are the same prior to exposure. For example, if $T = 5$ and $q = 3$, $y_t(3) = y_t(4) = y_t(5) = y_t(\infty)$ for all $t < 3$, $y_t(4) = y_t(5) = y_t(\infty)$ for all $t < 4$, and $y_t(5) = y_t(\infty)$ for $t < 5$. This has important implications for how one effectively defines control groups. For example, if the first period of exposure is $t = 3$ then cohorts $r \in \{4, 5, \infty\}$ contain valid control units.

I state the common trends assumption for the staggered case as follows.

Assumption CTS (**Common Trend**, **Staggered**): With the treatment cohort dummies d_q, ..., d_T,

$$E[y_t(\infty) - y_1(\infty)|d_q, \ldots, d_T] = E[y_t(\infty) - y_1(\infty)] \equiv \theta_t,\ t = 2, \ldots, T.\ \square \tag{6.4}$$

As before, this is the same as assuming that the average trends in adjacent time periods do not change with exposure:

$$E[y_t(\infty) - y_{t-1}(\infty)|d_q, \ldots, d_T] = E[y_t(\infty) - y_{t-1}(\infty)],\ t = 2, \ldots, T \tag{6.5}$$

The latter statement is similar to Callaway and Sant'Anna (2021), although they define the condition relative to a never treated group or a yet-to-be-treated group. Here, the CT

assumption is stated entirely in terms of the potential outcome in the never treated state – whether or not we observe never treated units – and so is more in the spirit of the usual potential outcomes setting.

Initially, we assume that there is a never treated (NT) group and that there is nonzero probability of entry in each period. This means that the τ_{rt} will be identified for $r \in \{q, \ldots, T\}$ and $t \in \{r, \ldots, T\}$.

In deriving estimating equations, it is useful to write the observed outcome in any period t as

$$\begin{aligned} y_t &= y_t(\infty) + d_q \cdot [y_t(q) - y_t(\infty)] + d_{q+1} \cdot [y_t(q+1) - y_t(\infty)] + \cdots + d_T \cdot [y_T(T) - y_t(\infty)] \\ &= y_t(\infty) + d_q \cdot te_t(q) + \cdots + d_T \cdot te_t(T) \end{aligned} \tag{6.6}$$

Underlying this equation is that $d_q + \cdots + d_T$ is not identically one, so that $d_\infty \equiv 1 - (d_q + \cdots + d_T)$ is not identically zero. It follows that

$$E(y_t|\mathbf{d}) = E[y_t(\infty)|\mathbf{d}] + d_q \cdot E[te_t(q)|\mathbf{d}] + \cdots + d_T \cdot E[te_t(T)|\mathbf{d}] \tag{6.7}$$

Because of the mutually exclusive nature of the cohort dummies, and that every unit is in one and only one treatment group (including the never treated group), we can always write

$$d_r E[te_t(r)|d_q, \ldots, d_t] = d_r E[te_t(r)|d_r = 1] \tag{6.8}$$

and so

$$E(y_t|\mathbf{d}) = E[y_t(\infty)|\mathbf{d}] + d_q \cdot E[te_t(q)|d_q = 1] + \cdots + d_T \cdot E[te_t(T)|d_T = 1] \tag{6.9}$$

For $t < r$, Assumption NAS implies that $E[te_t(r)|d_r = 1] = 0$. Therefore, for $t < q$,

$$E(y_t|\mathbf{d}) = E[y_t(\infty)|\mathbf{d}]$$

and for $t \geq q$,

$$E(y_t|\mathbf{d}) = E[y_t(\infty)|\mathbf{d}] + d_q \cdot E[te_t(q)|d_q = 1] + \cdots + d_t \cdot E[te_t(t)|d_t = 1]$$
$$\equiv E[y_t(\infty)|\mathbf{d}] + d_q\tau_{qt} + \cdots + d_t\tau_{tt} \tag{6.10}$$

Next, write $y_t(\infty)$ in terms of the initial outcome in the NT state and the change relative to the first period as

$$y_t(\infty) = y_1(\infty) + g_t(\infty), t = 2, \ldots, T$$

Imposing Assumption CTS (and $\theta_1 = 0$ because η is already the intercept for the first time period),

$$E[y_t(\infty)|\mathbf{d}] = E[y_1(\infty)|\mathbf{d}] + E[g_t(\infty)|\mathbf{d}] \equiv \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \theta_t \tag{6.11}$$

where the equation

$$E[y_1(\infty)|\mathbf{d}] = \eta + \lambda_q d_q + \cdots + \lambda_T d_T \tag{6.12}$$

is definitional because the cohort indicators are mutually exclusive and, along with the NT group, exhaustive. Combining (6.10), (6.11), and (6.12), we have for $1 \leq t < q$,

$$E(y_t|\mathbf{d}) = \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \theta_t, \tag{6.13}$$

and for $q \leq t \leq T$,

$$E(y_t|\mathbf{d}) = \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \theta_t + \tau_{qt} d_q + \cdots + \tau_{tt} d_t \tag{6.14}$$

The CCT assumption imposes common coefficients on the d_q for $t < q$ but with an intercept shift given by θ_t relative to $t = 1$. To impose these common coefficients in estimation, it is useful to have an equation for any t that includes time period dummies fs_t:

$$E(y_t|\mathbf{d}) = \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \sum_{s=2}^{T} \theta_s fs_t + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}(d_r \cdot fs_t) \tag{6.15}$$

For a random draw i, we have

$$E(y_{it}|\mathbf{d}_i) = \eta + \lambda_q d_{iq} + \cdots + \lambda_T d_{iT} + \sum_{s=2}^{T} \theta_s fs_t + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}(d_{ir} \cdot fs_t)$$

It is also useful to write an equivalent equation that includes the binary, time-varying treatment indicator, w_{it}:

$$E(y_{it}|\mathbf{d}_i) = \eta + \lambda_q d_{iq} + \cdots + \lambda_T d_{iT} + \sum_{s=2}^{T} \theta_s fs_t + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t), \tag{6.16}$$

which emphasizes that the analysis here allows for substantial treatment effect heterogeneity beyond the approach that includes only w_{it} by itself and estimates a single treatment effect. The equivalence follows because $w_{it} = d_{iq} \cdot pq_t + d_{i,q+1} \cdot p(q+1)_t + \cdots + d_{iT} \cdot pT_t$, where pr_t is a dummy variable indicating $t \geq r$. It follows that $w_{it} \cdot d_{ir} \cdot fs_t = d_{ir} \cdot fs_t$ for $s \geq r$.

The coefficients in (6.15), including the τ_{rs}, are consistently estimated using the POLS regression

$$y_{it} \text{ on } 1, d_{iq}, ..., d_{iT}, f2_t, ..., fT_t, d_{iq} \cdot fq_t, ..., d_{iq} \cdot fT_t, ..., d_{iT} \cdot fT_t \tag{6.17}$$

In fact, the estimators are unbiased conditional on their being new units in each treatment cohort. In other words, we interact the cohort dummies with the time dummies corresponding to those periods where a cohort is subjected to the intervention. If new units enter in every period and there is a never-treated group, then there are $(T-q+1)(T-q+2)/2$ estimated effects. As before, there is an equivalent TWFE estimator obtained by just including the treatment effect interactions, $d_{iq} \cdot fq_t, ..., d_{iq} \cdot fT_t, ..., d_{iT} \cdot fT_t$, and dropping $\left(1, d_{iq}, \ldots, d_{iT}, f2_t, \ldots, fT_t\right)$. Also, the time period dummies prior to q can be dropped without affecting the estimates.

While the main purpose of this section is to discuss identification and the mechanics of

POLS and TWFE estimation, using asymptotic inference requires a sufficient number of treated units for a particular cohort within each time period. With few new treated units, one might have no choice but to impose restrictions on the treatment effects. I discuss this further below.

6.2. Other Treatment Effects

One can also ask about identification of treatment effects other than just ATTs relative to the never treated state. For example, for $r \geq q$ and $r + 1 \leq t < T$, consider

$$y_t(r) - y_t(r+1),$$

the "gain" in period t of having been exposed to the intervention for one more period: rather than being first exposed in period $r + 1$, exposure happens one period earlier. Define the ATT for the subgroup actually treated in the earlier period:

$$\tau_{(r:r+1),t} \equiv E[y_t(r) - y_t(r+1)|d_r = 1] \tag{6.18}$$

Now, write

$$y_t(r) - y_t(r+1) = [y_t(r) - y_t(\infty)] - [y_t(r+1) - y_t(\infty)]$$

and so

$$\tau_{(r:r+1),t} = \tau_{rt} - E[y_t(r+1) - y_t(\infty)|d_r = 1]$$

Next, write

$$y_t(r+1) - y_t(\infty) = [y_t(r+1) - y_1(r+1)] - [y_t(\infty) - y_1(\infty)] + [y_1(r+1) - y_1(\infty)].$$

By no anticipation, the last term is zero. [Stated in terms of conditional means we would have to strengthen the assumption to $E[y_1(r+1) - y_1(\infty)|d_r = 1] = 0$.] Therefore,

$$E[y_t(r+1) - y_t(\infty)|d_r = 1] = E[y_t(r+1) - y_1(r+1)|d_r = 1] \\ - E[y_t(\infty) - y_1(\infty)|d_r = 1]$$

By the CT assumption, the second term is the same as $E[y_t(\infty) - y_1(\infty)|d_{r+1} = 1]$. For the first term, we have to add the CT assumption

$$E[y_t(r+1) - y_1(r+1)|\mathbf{d}] = E[y_t(r+1) - y_1(r+1)], \tag{6.19}$$

which says CT holds not just for the never treated state but also for other treated states. Under this assumption,

$$E[y_t(r+1) - y_t(\infty)|d_r = 1] = E[y_t(r+1) - y_t(\infty)|d_{r+1} = 1] \equiv \tau_{r+1,t}.$$

Therefore, we have shown, under the stronger NA and CT assumptions,

$$\tau_{(r:r+1),t} = \tau_{rt} - \tau_{r+1,t}, \tag{6.20}$$

which is the natural definition of the marginal increment of being treated one period earlier (with or without an explicit potential outcomes framework).

In order to obtain the incremental effects as differences $\tau_{rt} - \tau_{st}$ for $r \leq s \leq t$, we need parallel trends to hold in every treated state except for $r = q$ – the first treated cohort. Because $r = q$ is never used as an initial state – units cannot be exposed to the treatment earlier – we need not restrict $y_t(q) - y_1(q)$.

6.3. Adding Covariates

If we simply add covariates $\mathbf{x}_i$ to (6.17), or even interactions $d_{ir} \cdot \mathbf{x}_i$, the estimated treatment effects do not change. As in the common intervention case, this algebraic result follows from Theorem 3.2. However, having access to covariates allows us to relax the common trends assumption even in the staggered case. We first write the no anticipation assumption as follows:

Assumption CNAS (**Conditional No Anticipation**, **Staggered**): For treatment cohorts $r = q,\ q+1,\ \ldots, T$ and cohort indicators d_r,

$$E[y_t(r) - y_t(\infty)|d_r = 1, \mathbf{x}] = 0,\ t < r.\ \square \tag{6.21}$$

Again, this assumption clearly holds of we simply assume $y_t(r) = y_t(\infty),\ t < r$.

Assumption CCTS (**Conditional Common Trends**, **Staggered**): For cohort indicators d_r and covariates $\mathbf{x}$,

$$E[y_t(\infty) - y_1(\infty)|\mathbf{d}, \mathbf{x}] = E[y_t(\infty) - y_1(\infty)|\mathbf{x}],\ t = 2, \ldots, T.\ \square \tag{6.22}$$

The interpretation of this assumption is essentially the same as in the common intervention case.

Now we initially focus on the ATTs conditional on the covariates,

$$\tau_{rt}(\mathbf{x}) \equiv E[te_t(r)|d_r = 1, \mathbf{x}] \tag{6.23}$$

and then average out across the distribution of $\mathbf{x}$ conditional on $d_r = 1$ to obtain the ATTs.

Again write y_t as in (6.6) so that

$$E(y_t|\mathbf{d}, \mathbf{x}) = E[y_t(\infty)|\mathbf{d}, \mathbf{x}] + d_q \cdot E[te_t(q)|\mathbf{d}, \mathbf{x}] + \cdots + d_T \cdot E[te_t(T)|\mathbf{d}, \mathbf{x}] \tag{6.24}$$

Because of the mutually exclusive nature of the cohort dummies, and that every unit is in one and only one treatment group (which includes a never treated group), we can always write

$$d_r E[te_t(r)|\mathbf{d}, \mathbf{x}] = d_r E[te_t(r)|d_r = 1, \mathbf{x}] \equiv d_r \tau_{rt}(\mathbf{x})$$

Therefore, we can write

$$E(y_t|\mathbf{d}, \mathbf{x}) = E[y_t(\infty)|\mathbf{d}, \mathbf{x}] + d_q \cdot \tau_{qt}(\mathbf{x}) + \cdots + d_T \cdot \tau_{Tt}(\mathbf{x}) \tag{6.25}$$

Note that by Assumption CNA, $\tau_{rt}(\mathbf{x}) = 0$ if $t < r$, and so $d_r \tau_{rt}(\mathbf{x}) = 0$ if $t < r$.

Next, write the outcome in the never treated state as the first period value plus the time

change:

$$y_t(\infty) = y_1(\infty) + g_t(\infty),\ t = 2,\ldots,T$$

By the CCTS assumption,

$$E[y_t(\infty)|\mathbf{d},\mathbf{x}] = E[y_1(\infty)|\mathbf{d},\mathbf{x}] + E[g_t(\infty)|\mathbf{d},\mathbf{x}] = E[y_1(\infty)|\mathbf{d},\mathbf{x}] + E[g_t(\infty)|\mathbf{x}] \tag{6.26}$$

Assumption LINS (**Linear in Parameters**. **Staggered**): For covariates $\mathbf{x}$,

$$E[y_1(\infty)|\mathbf{d},\mathbf{x}] = \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \mathbf{x}\boldsymbol{\kappa} + (d_q \cdot \mathbf{x})\boldsymbol{\zeta}_q + \cdots + (d_T \cdot \mathbf{x})\boldsymbol{\zeta}_T \tag{6.27}$$

$$E[g_t(\infty)|\mathbf{x}] = \theta_t + \mathbf{x}\boldsymbol{\pi}_t,\ t = 2,\ldots,T \tag{6.28}$$

$$\tau_{rt}(\mathbf{x}) = \tau_{rt} + \dot{\mathbf{x}}_r\boldsymbol{\rho}_{rt},\ r = q,\ldots,T;\ t = r,\ldots,T \tag{6.29}$$

where

$$\dot{\mathbf{x}}_r \equiv \mathbf{x} - E(\mathbf{x}|d_r = 1) \equiv \mathbf{x} - \boldsymbol{\mu}_r.\ \square \tag{6.30}$$

Note how in (6.30) the controls have been centered about the within-cohort mean. This ensures that the intercept is τ_{rt}, the parameter we want to estimate on the interactions between the cohort dummies and calendar time dummies.

Combining (6.25), (6.27), and the linear conditional expectations, and imposing the NA assumption, it follows that, for $t \geq q$,

$$\begin{aligned} E(y_t|d_q,\ldots,d_T,\mathbf{x}) = {} & \eta + \lambda_q d_q + \cdots + \lambda_T d_T + \mathbf{x}\boldsymbol{\kappa} + (d_q \cdot \mathbf{x})\boldsymbol{\zeta}_q + \cdots + (d_T \cdot \mathbf{x})\boldsymbol{\zeta}_T \\ & + \theta_t + \mathbf{x}\boldsymbol{\pi}_t + \tau_{qt} d_q + (d_q \cdot \dot{\mathbf{x}}_q)\boldsymbol{\rho}_{qt} + \tau_{q+1,t} d_{q+1} + (d_{q+1} \cdot \dot{\mathbf{x}}_{q+1})\boldsymbol{\rho}_{q+1,t} \\ & + \cdots + \tau_{tt} d_t + (d_t \cdot \dot{\mathbf{x}}_t)\boldsymbol{\rho}_{tt} \end{aligned} \tag{6.31}$$

We can write the estimating equation that includes all parameters by using the time period dummies:

$$E(y_t|d_q,\ldots,d_T,\mathbf{x}) = \eta + \sum_{r=q}^{T}\lambda_r d_r + \mathbf{x}\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_r \cdot \mathbf{x})\boldsymbol{\zeta}_r + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{x})\boldsymbol{\pi}_s \tag{6.32}$$
$$+ \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(d_r \cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(d_r \cdot fs_t \cdot \dot{\mathbf{x}}_r)\boldsymbol{\rho}_{rs}$$

For a random draw i, and with the treatment indicator w_{it},

$$E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i) = \eta + \sum_{r=q}^{T}\lambda_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_{ir} \cdot \mathbf{x}_i)\boldsymbol{\zeta}_r + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_s \tag{6.33}$$
$$+ \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it} \cdot d_{ir} \cdot fs_t \cdot \dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs},$$

which emphasizes that we are allowing the effect of treatment to change with cohort, calendar time, and control variables. When we operationalize the POLS regression, we use deviations from the cohort averages:

$$\dot{\mathbf{x}}_{ir} = \mathbf{x}_i - \bar{\mathbf{x}}_r = \mathbf{x}_i - N_r^{-1}\sum_{h=1}^{N} d_{hr}\mathbf{x}_h. \tag{6.34}$$

We can drop the terms fs_t and $fs_t \cdot \mathbf{x}_i$ for $s < q$ or include them without changing the ATT estimates or the coefficients that measure moderating effects of the covariates. One version of the regression, across all i and t, is

$$\begin{aligned} &y_{it} \text{ on } 1,\ d_{iq},\ \ldots,\ d_{iT},\ \mathbf{x}_i,\ d_{iq} \cdot \mathbf{x}_i,\ \ldots,\ d_{iT} \cdot \mathbf{x}_i, fq_t,\ \ldots, fT_t, fq_t \cdot \mathbf{x}_i,\ \ldots, fT_t \cdot \mathbf{x}_i, \\ &\quad w_{it} \cdot d_{iq} \cdot fq_t,\ \ldots,\ w_{it} \cdot d_{iq} \cdot fT_t,\ \ldots,\ w_{it} \cdot d_{i,q+1} \cdot f(q+1)_t,\ \ldots,\ w_{it} \cdot d_{i,q+1} \cdot fT_t,\ \ldots,\ w_{it} \cdot d_{iT} \cdot fT_t, \\ &\quad w_{it} \cdot d_{iq} \cdot fq_t \cdot \dot{\mathbf{x}}_{iq},\ \ldots,\ w_{it} \cdot d_{iq} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT},\ w_{it} \cdot d_{i,q+1} \cdot f(q+1)_t \cdot \dot{\mathbf{x}}_{iq},\ \ldots, \\ &\quad w_{it} \cdot d_{i,q+1} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT},\ \ldots,\ w_{it} \cdot d_{iT} \cdot fT_t \cdot \dot{\mathbf{x}}_{iT} \end{aligned} \tag{6.35}$$

Alternatively – although the coefficients themselves would be difficult to interpret – we do not demean the $\mathbf{x}_i$ in the last terms and instead compute the average partial effect with respect to w, restrict attention to the relevant cohort-time period pair, and then average over the $d_r = 1$

subsample. (See the appendix for how this can be done in Stata.)

The regression in (6.35) uses levels controls for covariates and, effectively, treatment intensity. By contrast, Callaway and Sant'Anna (2021) use differences of the outcome variable and use inverse probability weighting. Because the expectation in (6.33) is in levels, it immediately suggests nonlinear models when y_{it} is limited in some way. I touch on this possibility in Section

As in all previous cases, there is an equivalent TWFE estimator when all time constant variables are dropped from (6.33) and fs_t and $fs_t \cdot \mathbf{x}_i$ are kept along with the terms involving w_{it}. Compared with the basic TWFE analysis, which imposes a constant coefficient on w_{it}, TWFE estimation of (6.33) allows for considerable heterogeneity. Compared with Sun and Abraham (2021), I explicitly allow covariates in order to make the common trends assumption more plausible and to allow treatment effects to vary by observed covariates.

6.4. Statistical Properties

It is useful to summarize the statistical properties of the POLS/ETWFE estimator by stating somewhat formal results. The following statements assume random sampling across i and enough finite moments. Unbiasedness, consistency, and asymptotic normality follow essentially immediately from (6.33). Let $\mathbf{D}$ be the $NT \times (T - q + 1)$ matrix of cohort indicators and let $\mathbf{X}$ be the $NT \times K$ matrix of covariates. For the consistency statement, let $\mathbf{z}_{it}$ be the row vector of explanatory variables in (6.33) and $\mathbf{Z}_i$ the matrix obtained by stacking the $\mathbf{z}_{it}$ from $t = 1, \ldots, T$. Naturally, these results include estimation common timing as a special case.

THEOREM 6.1 (Unbiasedness and Consistency): Under Assumptions CNA, CCT, and LIN, the POLS estimator $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{rs})$ from (6.35) has the following properties:

(i) The $\hat{\boldsymbol{\tau}}$ is unbiased for the $\boldsymbol{\tau}$ conditional on $(\mathbf{D}, \mathbf{X})$ for any realization where the standard

rank condition holds. (This means that each treatment cohort has a sufficient number of units and there is no perfect collinearity among the covariates.)

(ii) For fixed T, $\hat{\boldsymbol{\tau}} \xrightarrow{p} \boldsymbol{\tau}$ provided $E(\mathbf{Z}_i'\mathbf{Z}_i)$ is nonsingular.

(iii) For fixed T, $\sqrt{N}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})$ is asymptotically normal with the asymptotic variance the usual sandwich form.

Unbiasedness follows because under random sampling implies that the conditional mean in (6.33) is also the mean conditional on $(\mathbf{D}, \mathbf{X})$. Consistency and asymptotic normality (with the sandwich form of the asymptotic variance) are standard; see, for example, Wooldridge (2010, Section 7.3).

THEOREM 6.2 (Efficiency of POLS): Under the assumptions in Theorem 6.1, write (6.33) in composite error form as

$$y_{it} = \eta + \sum_{r=q}^{T} \lambda_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T} (d_{ir} \cdot \mathbf{x}_i)\boldsymbol{\zeta}_r + \sum_{s=2}^{T} \theta_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_s + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T} (w_{it} \cdot d_{ir} \cdot fs_t \cdot \dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs} + c_i + u_{it}, t = 1, \ldots, T \quad (6.36)$$

where

$$E(a_i|\mathbf{d}_i, \mathbf{x}_i) = 0, E(\mathbf{u}_i|c_i, \mathbf{d}_i, \mathbf{x}_i) = \mathbf{0} \quad (6.37)$$

Assume in addition that

$$Var(c_i|\mathbf{d}_i, \mathbf{x}_i) = \sigma_a^2 \quad (6.38)$$

$$Var(\mathbf{u}_i|c_i, \mathbf{d}_i, \mathbf{x}_i) = \sigma_u^2\mathbf{I}_T \quad (6.39)$$

where $\mathbf{u}_i' \equiv (u_{i1}, u_{i2}, \ldots, u_{iT})$. Then the POLS estimator $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{rs})$ has the following properties:

(i) $\hat{\boldsymbol{\tau}}$ is the best linear unbiased estimator (BLUE) of $\hat{\boldsymbol{\tau}}$ conditional on $(\mathbf{D}, \mathbf{X})$ for any realization where the standard rank condition holds.

(ii) $\hat{\boldsymbol{\tau}}$ is asymptotically efficient in the class of estimators consistent under the assumptions of Theorem 6.1. □

The BLUE result is somewhat subtle but it follows from the fact that the POLS estimator is the same as random effects even if we could use the true variances σ_a^2 and σ_u^2, a result that follows from Corollary 3.3. If we define $v_{it} = a_i + u_{it}$ and $\mathbf{v}_i' \equiv (v_{i1}, v_{i2}, \ldots, v_{iT})$ then $Var(\mathbf{v}_i|\mathbf{d}_i, \mathbf{x}_i)$ has the random effects structure, and so the RE estimator is BLUE. The asymptotic result follows from standard asymptotic efficiency of GLS for fixed-T panel data asymptotics when the variance-covariance matrix is correctly specified. See Wooldridge (2020, Section 10.4).

BJS (2021), assuming that all randomness stems from the error term u_{it} in (6.36), show that their imputation estimators are also BLUE under an assumption similar to (6.38). The BJS estimators have a fixed effects flavor and it would not be surprising if some version is identical the to POLS/RE/ETWFE estimator studied here. As stated before, one benefit of the POLS/ETWFE approach is its simplicity in obtaining robust standard errors of any combination of the $\hat{\tau}_{rs}$.

6.5. Aggregating and Imposing Restrictions on the Treatment Effects

An equation such as (6.33) contains many parameters, some of which may not be estimable depending on the pattern of assignment. Even if we can obtain an estimator for each (r, t) combination with $r \in \{q, \ldots, T\}$, $t \in \{r, \ldots, T\}$, the number of units entering treatment in a particular cohort might not produce suitably precise estimates. And, there is always the worry that the asymptotic approximations to the confidence intervals and t statistics are poor. Using long differencing combined with regression adjustment and IPW, Callaway and Sant'Anna (2021) suggest aggregating the τ_{rt}. Here, such aggregation is much easier because it can be

done within standard regression analysis.

The case of common treatment timing is relatively straightforward, as the number of treated and control units in each post-intervention period is the same. Therefore, provided the initial intervention is applied to sufficiently many units, the usual asymptotic approximations for standard errors and confidence intervals should work well. In this setting, one might want to impose restrictions on how the covariates appear. For example, rather than have a full set of interactions $w_{it} \cdot fr_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$, $r = q, \ldots, T$, one might include, in the TWFE estimation, only the interactions $w_{it} \cdot p_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$ for the post-treatment dummy p_t. In this case, one should include (at least) $p_t \cdot (\mathbf{x}_i - \boldsymbol{\mu}_1)$, too.

In the case of staggered interventions, two related approaches are possible for reducing the number of treatment effects. First, one might aggregate the separate TEs into a small number. For example, in the model with a full set of TEs, define the parameter of interest to be

$$\bar{\tau} \equiv \frac{1}{(T-q+1)(T-q+2)/2} \sum_{r=q}^{T} \sum_{t=r}^{T} \tau_{rt} \tag{6.40}$$

After obtaining the $\hat{\tau}_{rt}$ using the TWFE with covariates or the pooled OLS equivalent, it is easy in standard regression packages to obtain the standard error of $\hat{\bar{\tau}}$ because it is just a linear function of the $\hat{\tau}_{rt}$. Or, we might compute the average effect by entry cohort:

$$\hat{\bar{\tau}}_r \equiv \frac{1}{(T-r+1)} \sum_{t=r}^{T} \hat{\tau}_{rt}, \tag{6.41}$$

so there is one effect per entry cohort.

Alternatively, one might directly impose restrictions on the τ_{rt}. For example, common effect within cohort is obtained by

$$\tau_{rt} = \tau_{rr}, t = r, r+1, \ldots, T; r = q, \ldots, T \tag{6.42}$$

Then there are only $T - q + 1$ different average treatment effects rather than $(T - q + 1)(T - q + 2)/2$ ATTs. This restriction is imposed using treatment indicators $w_{it} \cdot dr_i$ for $r = q, \ldots, T$ rather than the triple interactions $w_{it} \cdot dr_i \cdot fs_t$. Or, one can allow different effects across calendar time but impose homogeneity by cohort: $\tau_{rt} = \tau_{tt}$, $r = q, \ldots, t$, $t = q, \ldots, T$. In implementation, this means only including $w_{it} \cdot fs_t$, $s = q, \ldots, T$ in the estimating equation.

Another potentially attractive restriction that can substantially reduce the number of parameters is to assume that treatment effect differs only by treatment intensity. The restrictions can be written, for parameters τ_{t-r+1}, as

$$\tau_{rt} = \tau_{t-r+1}, t = r, \ldots, T; r = q, \ldots, T. \tag{6.43}$$

When $t = r$ a unit has been exposed for one period, and the restrictions in (6.43) mean that the immediate impact is the same across entry cohort. If $t - r = 1$ then a unit is in its second period of exposure. And so on. For example, with $T = 8$ and staggered interventions starting in period $q = 5$, there would be four treatment effect parameters corresponding to four intensity levels – rather than 10 in the unrestricted model. Imposing these restrictions is easily done by creating a set of treatment intensity indicators and including them in place of the indicators $w_{it} \cdot d_{ir} \cdot fs_t$.

With many treated time periods and N not especially large, one might partition the cohorts into something like "early treated," "middle treated," and "late treated."

It is also prudent in many cases to impose restrictions on the vectors $\boldsymbol{\rho}_{rs}$. A natural restriction is common effects by cohort, and so the interaction terms become $w_{it} \cdot dr_i \cdot \dot{\mathbf{x}}_{ir}$. One might even impose $\boldsymbol{\rho}_{rs} = \boldsymbol{\rho}$ for all r and s and then center $\mathbf{x}_i$ about the mean computed from all

(eventually) treated units.

Any restrictions imposed is easily tested by estimating the unrestricted model and then testing the corresponding linear restrictions using a robust Wald statistic that allows general serial correlation and heteroskedasticity. See the supplemental materials for examples.

6.6. Restricted Treatment Patterns

In many intervention analyses with staggered treatment the intervention pattern is not completely general. One common case is when entry stops at a certain point but the units (control and treated) are followed for subsequent periods. This is an easy case to handle because there are simply fewer entry cohorts, and so the cohort dummies after the final entry period are simply dropped from the analysis. In particular, equations (6.16) and (6.33) will have fewer treatment interactions as well as interactions with the covariates in the latter case. For example, suppose $T = 8$, $q = 5$, but entry occurs only in periods five and six. Then the interactions one can include without consider covariates are $d_5 f5_t$, $d_5 f6_t$, $d_5 f7_t$, $d_5 f8_t$, $d_6 f6_t$, $d_6 f7_t$, and $d_6 f8_t$. It is still possible to estimate dynamic effects of the intervention for the two entering cohorts.

6.7. What if all Units are Eventually Treated?

In some designs, all units eventually enter the treated state. It is straightforward to modify the previous analysis so that a conditional mean similar to (6.32) still identifies interesting treatment effects. Consequently, we can resolve the identification issue that has been raised by Callaway and Sant'Anna (2021) and BJS (2021). For example, BJS (2021) use potential outcomes $y_t(0)$ and $y_t(1)$ and so it is more difficult define the incremental effect of an earlier exposure to treatment relative to first exposure in the final period.

With $d_\infty \equiv 0$, we effectively take the cohort entering in the final period – at which points

are units are treated – as the base group. Now define treatment effects

$$\tau_{(r:T),t} = E[y_t(r) - y_t(T)|d_r = 1],\ r = q,\ldots,T-1,\ t = r,\ldots,T, \tag{6.44}$$

so that $\tau_{(r:T),t}$ is the ATT in period t when moving from treatment in the final period T to any earlier period $q \leq r \leq T-1$. For example, with $T = 6$ and $q = 4$, there are entry cohorts in periods four, five, and six. We cannot, in general, identify a treatment effect for the final entry cohort because, with a never treated group, there is no suitable control group for the final period cohort. But we can estimate the TEs given in (6.44) by simply stating the no anticipation and parallel trend assumptions for the $d_T = 1$ rather than $d_\infty = 1$ group. This is easily seen by writing

$$y_t = y_t(T) + d_q[y_t(q) - y_t(T)] + \cdots + d_{T-1}[y_t(T-1) - y_t(T)]$$

and so, letting $\mathbf{d} = (d_q,\ldots,d_{T-1})$,

$$\begin{aligned} E(y_t|\mathbf{d},\mathbf{x}) &= E[y_t(T)|\mathbf{d},\mathbf{x}] + d_q E[y_t(q) - y_t(T)|\mathbf{d},\mathbf{x}] + \cdots + d_{T-1}E[y_t(T-1) - y_t(T)|\mathbf{d},\mathbf{x}] \\ &= E[y_t(T)|\mathbf{d},\mathbf{x}] + d_q \cdot \tau_{(q:T),t}(\mathbf{x}) + \cdots + d_{T-1} \cdot \tau_{(T-1:T),t}(\mathbf{x}) \end{aligned} \tag{6.45}$$

where

$$\tau_{(r:T),t}(\mathbf{x}) \equiv E[y_t(r) - y_t(T)|d_r = 1,\mathbf{x}],\ r = q,\ldots,T-1,\ t = r,\ldots,T$$

Writing

$$y_t(T) = y_1(T) + g_t(T)$$

the conditional CT assumption is now

$$E[g_t(T)|\mathbf{d},\mathbf{x}] = E[g_t(T)|\mathbf{x}]$$

The no anticipation assumption is as in the case with a never treated group: replace "∞" with "T" and state it for entry cohorts $r = q,\ldots,T-1$. Then, with linearity of the conditional

expectations as in Assumption LINS, the analog of (6.32) is

$$E(y_t|d_q,\ldots,d_{T-1},\mathbf{x}) = \eta + \sum_{r=q}^{T-1}\lambda_r d_r + \mathbf{x}\boldsymbol{\kappa} + \sum_{r=q}^{T-1}(d_r \cdot \mathbf{x})\boldsymbol{\zeta}_r + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{x})\boldsymbol{\pi}_t \qquad (6.46)$$
$$+ \sum_{r=q}^{T-1}\sum_{s=r}^{T}\tau_{(r:T),t}(d_r \cdot fs_t) + \sum_{r=q}^{T-1}\sum_{s=r}^{T}(d_r \cdot fs_t \cdot \dot{\mathbf{x}}_r)\boldsymbol{\rho}_{rs}$$

and then the POLS regression looks like

$$\begin{aligned} &y_{it} \text{ on } 1,\ d_{iq},\ ...,\ d_{i,T-1},\ \mathbf{x}_i,\ d_{iq}\cdot\mathbf{x}_i,\ ...,\ d_{i,T-1}\cdot\mathbf{x}_i, fq_t,\ ...,fT_t, fq_t\cdot\mathbf{x}_i,\ ...,fT_t\cdot\mathbf{x}_i,\\ &\quad w_{it}\cdot d_{iq}\cdot fq_t,\ ...,\ w_{it}\cdot d_{iq}\cdot fT_t,\ ...,\\ &\qquad w_{it}\cdot d_{i,q+1}\cdot f(q+1)_t,\ ...,\ w_{it}\cdot d_{i,q+1}\cdot fT_t,\ ...,\ w_{it}\cdot d_{iT-1}\cdot f(T-1)_t,\ w_{it}\cdot d_{iT-1}\cdot f(T)_t \qquad (6.47)\\ &\quad w_{it}\cdot d_{iq}\cdot fq_t\cdot\dot{\mathbf{x}}_{iq},\ ...,\ w_{it}\cdot d_{iq}\cdot fT_t\cdot\dot{\mathbf{x}}_{i,T-1},\ w_{it}\cdot d_{i,q+1}\cdot f(q+1)_t\cdot\dot{\mathbf{x}}_{iq},\ ...,\\ &\quad w_{it}\cdot d_{i,q+1}\cdot fT_t\cdot\dot{\mathbf{x}}_{i,T-1},\ ...,\ w_{it}\cdot d_{i,T-1}\cdot f(T-1)_t\cdot\dot{\mathbf{x}}_{i,T-1},\ w_{it}\cdot d_{i,T-1}\cdot fT_t\cdot\dot{\mathbf{x}}_{i,T-1} \end{aligned}$$

Just as in the case with the never treated cohort, the regression is easy to describe and to implement. We simply drop all terms corresponding to the final cohort, which means all of the terms that depend on d_{iT}: the final entry cohort now acts as the control group. The coefficients on terms of the form $w_{it}\cdot d_{ir}\cdot fs_t$ are ATTs relative the the last entry cohort. As usual, it is sufficient to use the interactions $d_{ir}\cdot fs_t$ because $w_{it}\cdot d_{ir}\cdot fs_t = d_{ir}\cdot fs_t$.

This analysis shows that a lot still can be learned even if there is no never treated cohort. In fact, under no anticipation, if we want to imagine there *could* be never treated units then we can take $y_t(T) = y_t(\infty)$ for $t < T$ by the no anticipation assumption: prior to the last cohort being treated their outcome is the same as in the never treated state. If we take this stance then, for $r < T$, the treatment effects $\tau_{(r:T),t}$, $t = r,\ldots,T$, are exactly as we had before with a never treated cohort. The primary change is that in the last period we have no choice but to compare the earlier treated cohorts with the final treated cohort, and so we etimate $\tau_{(r:T),T}$ for $r = q,\ldots,T-1$. Compared with the case with a never treated group we lose only one ATT

parameter. This is different from other approaches when there is no NT cohort – for example, CS (2021) – where no treatment effects are estimated in the final period.

6.8. Some Simple Cases

It is helpful to consider some simple cases in order to understand how pooled OLS (extended TWFE) uses information in the data under no anticipation and common trends. First consider the case $T = 3$ where $t = 1$ is the untreated period and entry is staggered at $t = 2$ and $t = 3$. Initially assume that there is a never treated group, so there are three treatment effects that we defined earlier: τ_{22}, τ_{23} (the two effects for the first entry cohort) and τ_{33}. Without covariates, the POLS regression is

$$y_{it} \text{ on } 1, d_{i2}, d_{i3}, f2_t, f3_t, d_{i2} \cdot f2_t, d_{i2} \cdot f3_t, d_{i3} \cdot f3_t, t = 1,2,3;\ i = 1,2,\ldots,N \tag{6.48}$$

Consider the estimate $\hat{\tau}_{22}$, the coefficient on $d_{i2} \cdot f2_t$. It can be shown that this estimate is identical to dropping the last time period and running the regression

$$y_{it} \text{ on } 1, d_{i2}, f2_t, d_{i2} \cdot f2_t, t = 1,2;\ i = 1,2,\ldots,N \tag{6.49}$$

and, again, obtaining the coefficient on $d_{i2} \cdot f2_t$. This shorter regression is precisely the simple two-period DiD regression where both $d_\infty = 1$ and $d_3 = 1$ are used as the control group. Using both the never treated and yet-to-be treated is the the correct thing to do because, if we believe the NA and CT assumptions, the potential outcomes $y_t(3)$ and $y_t(\infty)$ are the same for $t \leq 2$.

The DiD estimate from (6.49) is the usual one:

$$\hat{\tau}_{22} = N_2^{-1} \sum_{t=1}^{N} d_{i2} \Delta y_{i2} - (N - N_2)^{-1} \sum_{t=1}^{N} (1 - d_{i2}) \Delta y_{i2}$$

which shows that any use of regression adjustment or IPW methods should be based on $\Delta y_{i2} = y_{i2} - y_{i1}$ with the control group including the never treated units and those first treated

in $t = 3$. In cases where a never treated group exists, Callaway and Sant'Anna (2021) separate the NT group from other potential controls and therefore does not use all of the information availabe in the identifying assumptions. In the simplest case, this would mean separate estimates of τ_{22} using the $d_{i\infty} = 1$ as the control group and then $d_{i3} = 1$ as the control group. This may seem more robust than combining them into a single control group, but Assumptions NA and CT are already combined and used for identification of the treatment effects. Pooled OLS incorparates the information in the maintained assumptions and in some cases is the most efficient estimator in terms of exact variance and asymptotically.

Now consider estimating $\tau_{33} = E[y_3(3) - y_3(\infty)|d_3 = 1]$. Because the $r = 2$ cohort has been already treated it is not used as part of the control group, so we are relying on the never treated group. In doing so, we should recognize that information is available in the usual one-period difference and the two-period difference. To see why, first write

$$y_3(3) - y_3(\infty) = [y_3(3) - y_2(3)] - [y_3(\infty) - y_2(\infty)] + [y_2(3) - y_2(\infty)]$$

Therefore,

$$\tau_{33} = E[y_3(3) - y_2(3)|d_3 = 1] - E[y_3(\infty) - y_2(\infty)|d_3 = 1]$$

The first of these is always identified and the second is identified by the CT assumption because

$$E[y_3(\infty) - y_2(\infty)|d_3 = 1] = E[y_3(\infty) - y_2(\infty)|d_\infty = 1]$$

The method of moments estimator is then just the usual two-period DiD from $t = 2$ to $t = 3$, with $r = 3$ the treatment group and $r = \infty$ the control group.

Using DiD for $t \in \{2,3\}$ ignores useful information available in the first period outcomes. To see why, now write

$$y_3(3) - y_3(\infty) = [y_3(3) - y_1(3)] - [y_3(\infty) - y_1(\infty)] + [y_1(3) - y_1(\infty)]$$

Again, imposing NA and CT,

$$\tau_{33} = E[y_3(3) - y_1(3)|d_3 = 1] - E[y_3(\infty) - y_1(\infty)|d_\infty = 1]$$

which shows that we can use a long difference, from period one to three, in a standard DiD analysis. But there is no need to separate them. What makes sense is to use a common timing scenario with $q = 3$ and the first two periods as control periods, with $r = 3$ the treated group and $r = \infty$ the control group. This is precisely what POLS/TWFE does. In fact, in the general case, the POLS estimates of the τ_{rr} are identical to using a "rolling DiD" method, where not-yet-treated units and periods $t = 1, \ldots, r - 1$ are used as the control groups. By Assumptions CNA and CCT, the pre-treatment time periods can (and should be) averaged into a single unit, and that is what POLS does.

If in the previous setup all units are treated by $t = 3$, there are only two parameters to estimate. We called these $\tau_{(2:3),2}$ and $\tau_{(2:3),3}$ in Section 6.2, where the former is also $\tau_{(2:\infty),2}$ under no anticipation. As shown in Section 6.4, the POLS estimator (ETWFE) provides unbiased and consisent estimators of these parameters. Methods that require a "not yet treated" group do not identify $\tau_{(2:3),3}$, the impact in period three of having been treated one period earlier. Pooled OLS identifies these effects in the general case and for all combinations of cohort/calendar time combinations that have causal interpretations.

A related point concerns adding more pre-treatment periods. Suppose $T = 4$ and $q = 3$, so there are two control periods. Estimating the conditional mean in (6.16) will properly use both time periods in the control group. This is true with or without covariates. Using the $t = 1$ and $t = 2$ time periods separately as controls, as in Callaway and Sant'Anna (2021), is inefficient

(regardless of the approach taken, whether pure regression adjustment or combining regression with inverse probability weighting). As the number of pre-treatment periods increases, POLS will become even more efficient compared with CS (2021).

7. Testing and Relaxing the Common Trends Assumption

The POLS/ETWFE framework gives a simple framework for testing the null hypothesis that the common trends assumption holds (assuming no anticipation). The algebraic equivalences we relied on earlier provide insight into the nature of the pre-testing problem – something investigated in depth by Roth (2020). It is also easy to allow for certain kinds of heterogeneous trends by expanding the earlier equations.

7.1. Testing the Null of Common Trends

First consider the simplest case where we can test the CT assumption: $T = 3$ and the only treated period is $t = 3$. Because there are two pre-treatment periods we can test the CT assumption. There are many approaches the produce the same statistic. One possibility is to perform a standard two-period DiD using periods one and two. In other words, let $\Delta y_{i2} = y_{i2} - y_{i1}$ and then run the simple regression Δy_{i2} on $1, d_i$. Under the null of CT, the coefficient on d_i is zero, and so we can use a heteroskedasticity-robust t statistic. In effect, we are computing a treatment effect before the treatment took place. This is a kind of placebo test. Note that it can reject due to failure of no anticipation or common trends. As emphasized by Roth (2020), subsequent inference on the ATT after "passing" the common trends test can be misleading.

An equivalent way to obtain the test statistic is to run the pooled regression

$$y_{it} \text{ on } 1, d_i, f2_t, d_i \cdot f2_t, f3_t, d_i \cdot f3_t, t = 1,2,3; i = 1,\ldots,N \tag{7.1}$$

and obtain the cluster-robust t statistic on $d_i \cdot f2_t$. The regression in (7.1) is saturated: we are estimating the six cell means determined by the pairs (d,t) for $d \in \{0,1\}, t \in \{1,2,3\}$. The typical approach in an empirical study would be to fail to reject CT if the t statistic is not

significant, at, say, the 5% level. The previous algebraic equivalences help us understand that the pre-testing problem is identical to pre-testing on an explanatory variable. If we drop $d_i \cdot f2_t$ from the regression (7.1) then the presence of $f2_t$ does not affect the coefficient on $d_i \cdot f3_t$, which is $\hat{\tau}_3$. Thus, dropping $d_i \cdot f2_t$ means we are back to the POLS regression that delivers the ATT:

$$y_{it} \text{ on } 1, d_i, f3_t, d_i \cdot f3_t, t = 1,2,3; i = 1,\ldots,N$$

A test of CT can be obtained in the general case with $q-1$ control periods and common entry at time q. The expanded equation for testing CT is

$$y_{it} = \eta + \lambda d_i + \theta_2 f2_t + \cdots + \theta_T fT_t + \sum_{s=2}^{q-1} \omega_s (d_i \cdot fs_t) + \sum_{s=q}^{T} \delta_s (d_i \cdot fs_t) + u_{it} \tag{7.2}$$

and the null hypothesis is

$$H_0 : \omega_s = 0, s = 2,\ldots,q-1 \tag{7.3}$$

for a total of $q-2$ restrictions. We would test these restrictions using a cluster-robust Wald statistic. Note that the coefficients on $d \cdot fs_t$ for the treated periods are not shown as the ATTs because if we allow nonzero ω_s then the δ_s are not ATTs in general. If we estimate the equation with all ω_s set to zero then we obtain the $\hat{\tau}_t$ from Section 5.2. This, again, illustrates that pre-testing for CT is the same as pre-testing a set of $q-2$ variables.

One of the main reasons for collecting data on covariates is to relax the unconditional common trends assumption. Therefore, one should add the terms

$$d_i \cdot fq_t \cdot \dot{\mathbf{x}}_i, \ldots, d_i \cdot fT_t \cdot \dot{\mathbf{x}}_i, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, \mathbf{x}_i, d_i \cdot \mathbf{x}_i$$

to equation (7.2) and thest the same null in (7.3). If the null is imposed then we are back to the original estimates of the time-varying ATTs.

An alternative test replaces the $q-2$ interactions with the single variable

$$d_i \cdot t \tag{7.4}$$

which takes the alternative to CT to be a different linear for the treated group. With $q = 3$ this will produce the same statistic as (7.3), but with $q > 3$ the tests differ.

In the staggered case the number of place treatment indicators that can be included varies with treatment cohort. The last cohort has the most number of such indicators (assuming a never treated group). One can include $d_{iT} \cdot f2_t$, ..., $d_{iT} \cdot f_t(T-1)$, whereas for the first treated cohort the terms consist of $d_{iq} \cdot f2_t$, ..., $d_{iq} \cdot f_t(q-1)$. (If $q = 2$ then no terms are included for the first treated cohort.) Generally, write the augmented equation as

$$\begin{aligned} y_{it} = {} & \eta + \lambda_q d_{iq} + \cdots + \lambda_T d_{iT} + \theta_2 f2_t + \cdots + \theta_T fT_t \\ & + \sum_{r=q}^{T} \sum_{s=2}^{r-1} \omega_{rs}(d_{ir} \cdot fs_t) + \sum_{r=q}^{T} \sum_{s=r}^{T} \delta_{rs}(d_{ir} \cdot fs_t) + u_{it} \end{aligned} \tag{7.5}$$

and the null hypothesis is

$$H_0 : \omega_{rs} = 0, r = q, \ldots, T, s = 2, \ldots, r-1 \tag{7.6}$$

When we have covariates and are relying on the conditional version of CT, we should modify both versions of the test. The general idea is straightforward once we realize we need to allow full flexibility in the pre-treatment trends for the never treated state. Therefore, take the regression in (6.35) and add the variables

$$f2_t, \ \ldots, f_t(q-1), f2_t \cdot \mathbf{x}_i, \ \ldots, f_t(q-1) \cdot \mathbf{x}_i, \tag{7.7}$$

$$d_{iq}f2_t, \ ..., d_{iq}f_t(q-1), d_{i,q+1}f2_t, \ ..., d_{i,q+1}fq_t, d_{iT}f2_t, ..., , ..., d_{iT}f_t(T-1) \tag{7.8}$$

and then, again, jointly test the $(T-q+1)(q-2)$ variables in (7.8). If these variables are dropped then adding the variables in (7.7) does not affect the estimated ATTs or the

coefficients measuring the moderating effects. This test does not check for alternatives where violations of CT may depend on the $\mathbf{x}_i$; that would be costly in terms of degrees of freedom, and we are seeking a test with reasonable power. Alternatively, replace the variables in (7.8) with the cohort-specific trends, $d_{ir} \cdot t$. In either case were are testing for pre-trends prior to the first intervention period even in the staggered case.

If there is no never treated group then, as discussion Section 6.7, we simply drop all terms involving d_T because last cohort to be treated becomes the comparision group; see equation (6.46). The terms $d_{iT}f2_t$, ..., $d_{iT}f_t(T-1)$ are dropped and the last set of terms is $d_{i,T-1}f2_t$, ..., $d_{i,T-1}f_t(T-2)$. Naturally, this reduces the number of restrictions being tested compared with when there is a never treated group.

7.2. Allowing Heterogeneous Trends

In the common treatment timing case with only two pre-treatment periods, we noted that using $d_{ir} \cdot t$ in place of $d_{ir} \cdot f2_t$, $r = q, \ldots, T$, gives the same statistic for testing the null hypothesis of common trends. But they can deliver very different estimates on the post-intervention interaction terms that are used to estimate the treatment effects. In fact, the estimates on the actual treatment indicators $d_{ir} \cdot fs_t$, $s \geq r$ can be very different. This observation is one way of saying that correcting for heterogenous trends is generally difficult even when there are only two pre-intervention periods because it relies on specific assumptions about the underlying trends.

As models for heterogeneous trends, using $d_{ir} \cdot f2_t$ versus $d_{ir} \cdot t$ are very different, with the latter being more realistic. To see why, consider the very simple case with a common intervention date of $q = 3$ and no covariates. Letting $g_{it}(0)$ denote the growth from period 1 to t in the control state. The two models of heterogeneous trends are

$$E[g_{it}(0)|d_i] = \theta_t + \varphi(d_i \cdot f2_t), t = 2,\ldots,T \tag{7.9}$$

$$E[g_{it}(0)|d] = \theta_t + \omega(d_i \cdot t), t = 2,\ldots,T \tag{7.10}$$

In the first case, we can write

$$E[g_{it}(0)|d_i = 1] - E[g_{it}(0)|d_i = 0] = \varphi, t = 2 \tag{7.11}$$
$$= 0, t > 2$$

which means that any violation of CT happens only in period $t = 2$ and then reverts back to common trends. By contrast, (7.10) implies

$$E[g_{it}(0)|d_i = 1] - E[g_{it}(0)|d_i = 0] = \omega \cdot t, t = 2,\ldots,T$$

which can also be written as

$$\{E[g_{i,t+1}(0)|d_i = 1] - E[g_{it}(0)|d_i = 1]\} - \{E[g_{i,t+1}(0)|d_i = 0] - E[g_{it}(0)|d_i = 0]\} = \omega, t = 2,\ldots,T, \quad (7.$$

which allows for a constant difference in trends between the treated units and the control units. While restrictive, (7.12) seems much more plausible than (7.11). Plus, with more intervention periods the equation in (7.10) can allow for more polynomials in t. While more terms also can be added to (7.9), any extension would still have the implication that the difference in trends in the untreated state is always zero post-intervention.

The linear trend specification also leads to appealing estimators in simple cases. Again take $T = 3$ where the intervention occurs only in $t = 3$. The regression that incorporates a heterogenous linear trend is

$$y_{it} \text{ on } 1, d_i, f2_t, f3_t, d_i \cdot t, d_i \cdot f3_t, t = 1,2,3; i = 1,\ldots,N$$

and the estimator of the treatment effect is the coefficient on $d_i \cdot f3_t$. It turns out that we can write that coefficient as

$$\hat{\tau}_{3,DDD} = N_1^{-1} \sum_{i=1}^{N} d_i \cdot \Delta^2 y_{i3} - N_0^{-1} \sum_{i=1}^{N} (1 - d_i) \cdot \Delta^2 y_{i3} \tag{7.14}$$

$$= [(\bar{y}_{3,tr_t} - \bar{y}_{2,trt}) - (\bar{y}_{3,con} - \bar{y}_{2,con})] - [(\bar{y}_{2,trt} - \bar{y}_{1,trt}) - (\bar{y}_{2,con} - \bar{y}_{1,con})] \tag{7.15}$$

where $\Delta^2 y_{i3} = (y_{i3} - y_{i2}) - (y_{i1} - y_{i1})$ is the second difference of the response variable in the last time period. Equation (7.14) shows that $\hat{\tau}_{3,DDD}$ is the coefficient on d_i in the simple cross-sectional regression

$$\Delta^2 y_{i3} \text{ on } 1, d_i, i = 1, \ldots, N$$

whereas equation (7.15) shows that $\hat{\tau}_{3,DDD}$ is a difference-in-difference-in-differences (DiDiD) estimator. The second term in brackets in (7.15) is precisely the estimator underlying the common trends test discussed earlier. Here, it is used to adjust the usual two period DiD estimator, $(\bar{y}_{3,trt} - \bar{y}_{2,trt}) - (\bar{y}_{3,con} - \bar{y}_{2,con})$, for differences in trends prior to the intervention.

With staggered intervention we can add the terms

$$d_{iq} \cdot t, d_{i,q+1} \cdot t, ..., d_{iT-1} \cdot t, d_{iT} \cdot t$$

along with the time dummies from the pre-intervention periods to any of the previous estimations, including the regression (6.35) that allows for substantial heterogeneity. As discussed earlier, a valid test of the null of conditional CT is a joint test of these terms. If there is no never treated group, d_{iT} is dropped everywhere with the final cohort serving as the base group.

With enough data, one might also include interactions $d_{ir} \cdot t \cdot \mathbf{x}_i$ in addition to $fs_t \cdot \mathbf{x}_i$ for $s < q$ to allow the heterogeneous trends to depend on the observed covariates. Centering the covariates is optional because doing so does not affect the ATTs for the different cohorts across calendar time.

A fixed effects approach would drop the time constant variables and include a

heterogeneous trend along with the intercept:

$$y_{it} = c_i + h_i t + \sum_{s=2}^{T} \theta_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_t + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{rs}(w_{it} \cdot d_{ir} \cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T} (w_{it} \cdot d_{ir} \cdot fs_t \cdot \dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs} + u_{it}, \tag{7.14}$$

which is an extension of (6.36). Given at least two pre-treatment periods, we can remove $c_i + h_t t$ by using unit-specific detrending of y_{it} and all of terms on the right-hand side. For example, for each i we regress y_{it} on 1, t and obtain the residuals, $\ddot{y}_{it}$. After doing the same for each term on the RHS, we obtain, say, $\ddot{\mathbf{z}}_{it}$. Then we estimate all of the coefficients from a pooled OLS regression $\ddot{y}_{it}$ on $\ddot{\mathbf{z}}_{it}$. If N is large relative to T and we assume independence across i, one simply clusters the standard errors for serial correlation and heteroskedasticity. With more pre-treatment periods one can include more flexible unit-specific trends. See Wooldridge (2005) and Wooldridge (2010, Chapter 11) for further discussion.

8. Additional Issues and Extensions

8.1. Comments on Unbalanced Panels

When a panel data set is unbalanced due to missing data on some variables – in the context of Sections 5 and 6, this would be missing data on y_{it}, since when one uses covariates that do not change over time you are dropping units ahead of time with missing data on $\mathbf{x}$ – the equivalences derived earlier no longer hold. The pooled OLS regressions do not, in general, properly account for the possibility that that missingness can be due to unobserved heterogeneity. Moreover, in the fixed effects formulations, a full set of time period dummies should be included everywhere, including interacting with covariates when covariate adjustment is used. Even in the common timing case it is no longer sufficient to control for

time effects by simply including a $post_t$ dummy indicator.

Once we are including a full set of time period dummies, we can obtain a POLS (or RE) estimator equivalent to two-way FE by applying Wooldridge (2019). For example, in the case of common treatment time, with $w_{it} = d_i \cdot p_t$, the underlying equation that allows for differential effects across time is

$$y_{it} = \beta_q(w_{it} \cdot fq_t) + \cdots + \beta_{q1}(w_{it} \cdot fT_t) + \theta_2 f2_t + \cdots + \theta_T fT_t + c_i + u_{it} \tag{8.1}$$

With s_{it} denoting the complete cases indicator ($s_{it} = 1$ if y_{it} is observed), one must include the terms

$$d_i \cdot \overline{fq}_i, ..., d_i \cdot \overline{fT}_i, \overline{f2}_i, ..., \overline{fT}_i \tag{8.2}$$

where

$$\overline{fr}_i = T_i^{-1} \sum_{t=1}^{T} s_{it} fr_t$$

and $T_i = \sum_{t=1}^{T} s_{it}$ is the number of complete cases for unit i. (Any unit with $T_i = 0$ is necessarily dropped, and $T_i = 1$ units do not contribute to the estimation.) The POLS regression becomes

$$y_{it} \text{ on } 1, (d_i \cdot fq_t), ..., (d_i \cdot fT_t), f2_t, fT_t, d_i \cdot \overline{fq}_i, ..., d_i \cdot \overline{fT}_i, \overline{f2}_i, ..., \overline{fT}_i \tag{8.3}$$

Of course, it is much easier to use TWFE on the unbalanced panel once the interaction terms $w_{it} \cdot fr_t$ have been created. This is even moreso in the case of staggered interventions.

8.2. Nonlinear Models

As mentioned in the introduction, once we understand the relationships among estimators in the linear case, some flexible strategies for nonlinear models immediately suggest themselves. When $y_{it} \geq 0$ – it can be discrete, continuous, or mixed – a particularly attractive

approach is to assume an exponential mean function. The analog of (6.16) is

$$E(y_{it}|d_{iq},\ldots,d_{iT}) = \exp\left[\eta + \lambda_q d_{iq} + \cdots + \lambda_T d_{iT} + \sum_{s=2}^{T}\theta_s fs_t + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it}\cdot d_{ir}\cdot fs_t)\right], \quad (8.4)$$

and a convenient, fully robust estimation method is the pooled Poisson quasi-MLE; see, for example, Wooldridge (2010, Chapter 13). The parameter τ_{rs} is the ATT for cohort r in calendar time t (so conditional on $d_r = 1$) for $t = r,\ldots,T$, $r = q,\ldots,T$. Because of the exponential functional form, τ_{rs} is (approximately) a proportionate effect. One could compute a more accurate proportional effect, especially when τ_{rs} is large, by evaluating the mean function at $w_{it} = 1$ and $w_{it} = 0$ and setting the cohort and time dummies equal to the appropriate combination of zeros and ones. If the intervention date is common, the cohort dummies are replaced with a single d_i, and then treatment terms are simply the interactions $w_{it}\cdot fq_t$, $w_{it}\cdot f(q+1)_t$, ..., $w_{it}\cdot fT_t$.

It turns out to be equivalent to drop the time-constant controls d_{ir} and introduce multiplicative heterogeneity:

$$E(y_{it}|d_{iq},\ldots,d_{iT},c_i) = c_i\exp\left[\sum_{s=2}^{T}\theta_s fs_t + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it}\cdot d_{ir}\cdot fs_t)\right]. \quad (8.5)$$

Specifically, it can be shown that when (7.7) is used with the fixed effects Poisson quasi-MLE – see Wooldridge (1999, 2010) – it produces the same estimates of the τ_{rs} as the pooled Poisson estimates that include $d_{iq},\ldots,d_{iT}$ as control variables. The FEP estimator is fully robust to distributional misspecification and serial dependence, and robust inference is readily available. Plus, nothing special needs to be done if the panel is unbalanced.

It is straightforward to add covariates, as in (6.33). The following mean can be estimated

using FE Poisson:

$$E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i,c_i) = c_i \exp\left[\sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t \cdot \mathbf{x}_i)\boldsymbol{\pi}_s + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it}\cdot d_{ir}\cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it}\cdot d_{ir}\cdot fs_t\cdot \dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs}\right] \tag{8.6}$$

where, as in the linear case, we demean the covariates that interact with treatment cohort r:

$$\dot{\mathbf{x}}_{ir} = \mathbf{x}_i - E(\mathbf{x}_i|d_{ir}=1), r = q,\ldots,T.$$

Alternatively, one can use pooled Poisson regression with mean function

$$E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i) = \exp\left[\eta + \sum_{r=q}^{T}\lambda_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_{ir}\cdot\mathbf{x}_i)\boldsymbol{\zeta}_r + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{x}_i)\boldsymbol{\pi}_s + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rs}(w_{it}\cdot d_{ir}\cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it}\cdot d_{ir}\cdot fs_t\cdot\dot{\mathbf{x}}_{ir})\boldsymbol{\rho}_{rs}\right] \tag{8.7}$$

Naturally, in order to identify ATTs, we must impose some sort of common trend assumption. In the case with covariates, that assumption is in terms of ratios rather than differences:

$$\frac{E[y_t(\infty)|d_q,\ldots,d_T,\mathbf{x}]}{E[y_1(\infty)|d_q,\ldots,d_T,\mathbf{x}]} = \frac{E[y_t(\infty)|\mathbf{x}]}{E[y_1(\infty)|\mathbf{x}]}, t = 2,\ldots,T \tag{8.8}$$

When y_{it} is binary or fractional, nonlinear models that respect the bounded nature of y_{it} are attractive. For example, with $\Lambda(\cdot)$ the logistic function, a natural alternative to the linear equation in (6.33) is

$$E(y_{it}|d_{iq},\ldots,d_{iT},\mathbf{x}_i) = \Lambda\Bigg[\eta + \sum_{r=q}^{T}\lambda_r d_{ir} + \mathbf{x}_i\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_{ir}\cdot\mathbf{x}_i)\boldsymbol{\zeta}_r + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{x}_i)\boldsymbol{\pi}_t + \sum_{r=q}^{T}\sum_{s=r}^{T}\beta_{rs}(w_{it}\cdot d_{ir}\cdot fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{it}\cdot d_{ir}\cdot fs_t\cdot\mathbf{x}_i)\boldsymbol{\rho}_{rs}\Bigg], \tag{8.9}$$

where I dropped the demeaning of the covariates and relabeled the coefficients on the interactions $w_{it}\cdot d_{ir}\cdot fs_t$ to emphasize that these parameters are not average treatment effects (even with covariate demeaning). Instead, one needs to compute the average partial effect using discrete differences at $w_{it} = 0$ and $w_{it} = 1$, and choosing the different time periods, and then averaging over the relevant cohort. Packages that compute average partial effects for logit, probit (and fractional versions of these) make such calculations, along with standard errors, relatively simple.

As in all DiD-type estimators, underlying (8.9) is a conditional common trends assumption, but it would be effectively stated on an underlying latent variable, say, y_{it}^* that follows a linear model. I leave the exact nature of such assumptions to future research. Adding terms such as $d_{ir}\cdot t$, and even $d_{ir}\cdot t\cdot\mathbf{x}_i$ to the mean functions in (8.7) and (8.9) provides even more flexibility when there are at least two pre-intervention periods. These suggestions should be formally studied to ensure that interesting average treatment effects are being recovered, but that should be the case under reasonable assumptions.

Incidentally, when y_{it} is binary one might think of using a fixed effects logit (conditional MLE) estimation strategy, where the time-constant variables d_{ir} and in $(d_{ir}\cdot\mathbf{x}_i)$ are dropped. This approach is not recommended, for two reasons. First, consistency of the parameter estimators relies heavily on the assumption of serial independence in underlying idiosyncratic errors. Just as important is that, even under (conditional) serial independence, the FE logit

approach does not identify average treatment effects unless T is large enough to provide precise estimates of the unobserved heterogeneity. And even then, inference is very difficult.

8.3. Comments on Standard Errors

For panel data structures with N reasonably large and T not very large, at a minimum one should compute standard errors that allow for arbitrary serial correlation and heteroskedasticity. A more difficult issue is deciding whether to cluster at a level other than the unit level i. If the assignment is fully or partly made at a group level, g, then there is a case to be made for clustering at the level of g. The case can be made using a model-based approach with group-level heterogeneous treatment effects [for example, Wooldridge (2003)] or a design-based approach [as in Abadie, Athey, Imbens, and Wooldridge (2017)]. As an example, suppose that a staggered intervention occurs at the school level, i, but the probability of a school's being subjected to the intervention varies by school district, g. Generally, clustering at the school level will not properly account for the assignment uncertainty (including in the case where we observe the entire population of schools in a state, say). Clustering at the higher level can be conservative but is simple. See AAIW (2017) for additional discusion in a simple cross-sectional setting.

9. Concluding Remarks

The equivalence between the TWFE estimator and the TWM regression increases our understanding of commonly used estimators, especially difference-in-differences estimators with complicated intervention patters. It also highlights that, provided one allows treatment effects to be suitably heterogeneous, there is nothing inherently wrong with using TWFE in situations such as staggered interventions – a point that is also clear from Sun and Abraham (2021). In fact, because we know that TWFE is consistent for unbalanced panels (as the

cross-sectional sample size grows with T fixed), even when selection is correlated with additive, unobserved heterogeneity, it has advantages over other estimators that include time-constant cohort indicators and time effects.

The point here is not to conclude that other recent approaches – such de Chaisemartin and D'Haultfœuille (2021), Callaway and Sant'Anna (2021), Borusyak, Jaravel, and Spiess (2021), , among others – are not valuable and cannot improve over flexible TWFE methods. But I am recommending not to abandon simple regression approaches because I have shown they identify the treatment effects of interest very generally and can be made very flexible. Restrictions on treatment effects are easy to test and impose. Other than linearity of the conditional means in the covariates, the POLS/ETWFE approach offers everything one would want for staggered designs: it is simple, flexible, and has exact and asymptotic efficiency properties under the "ideal" assumptions. Competing estimation methods, such as Callaway and Sant'Anna (2021), do not exploit all available restrictions that are used for identification.

Another nice feature of a flexible TWFE approach is that it is easily extended to allow for heterogeneous trends, which can help when one suspects the common trends assumption is violated. Future simulation studies, empirical research, and even a competition where researchers apply their methods to various problems where they do not know the way the data were generated could help shed light on the advantages of each approach.

The equivalence between pooled OLS and ETWFE estimators in the balanced, linear case suggests simple strategies for nonlinear models. The exponential case is fairly straightforward, as it is easy to define treatment effects in terms of percentage changes and a suitable common trends assumption is readily identified. Moreover, the fixed effects Poisson estimator is fully robust and, as with the linear FE estimator, allows selection to be correlated with the

multiplicative heterogeneity. An in-depth analysis of nonnegative responses, as well as when y_{it} is binary or fractional, is left for future research.

References

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies* 72, 1-19.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics* 49, 431-434.

Borusyak, K., X. Jaravel, and J. Spiess (2021), "Revisiting Event Study Designs: Robust and Efficient Estimation," working paper, University College London.

Callaway, B., and P.H.C. Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods," forthcoming, *Journal of Econometrics*.

de Chaisemartin, C., and X. D'Haultfœuille (2020), "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review* 110, 2964–2996.

Driscoll, J., and A. Kraay (1998), "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics* 80, 549-560.

Goodman-Bacon, A. (2021), "Difference-in-Differences with Variation in Treatment Timing," forthcoming, *Journal of Econometrics*.

Hansen, C. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when *T* is Large," *Journal of Econometrics* 141, 597-620.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251-1271.

Heckman, J.J., H. Ichimura, and P.E. Todd (1997) "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64, 605-654.

Mundlak, Y. (1978), "On the Pooling of Cross Section and Time Series Data,"

Econometrica 46, 69-85.

Robins, J. M., and A. Rotnitzky (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association* 90, 122–129.

Roth, J. (2020), "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," working paper, Brown University Deparatment of Economics.

Sant'Anna, P.H.C. and J. Zhao (2020), "Doubly Robust Difference-in-Differences Estimators," *Journal of Econometrics* 219, 101-122.

Sun, L. and S. Abraham (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," forthcoming, *Journal of Econometrics*.

Vogelsang, T.J. (2012), "Heteroskedasticity, Autocorrelation, and Spatial Correlation Robust Inference in Linear Panel Models with Fixed-Effects," *Journal of Econometrics* 166, 303-319.

White, H. (1984), *Asymptotic Theory for Econometricians*. Academic Press: Orlando, FL.

Wooldridge, J.M. (1999), "Distribution-Free Estimation of Some Nonlinear Panel Data Models," *Journal of Econometrics* 90, 77-97.

Wooldridge, J.M. (2005), "Fixed-Effect and Related Estimators for Correlated-Random Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics* 87, 385-390.

Wooldridge, J.M. (2010), "Econometric Analysis of Cross Section and Panel Data," second edition. MIT Press: Cambridge, MA.

Wooldridge, J.M. (2019), "Correlated Random Effects Models with Unbalanced Panels,"

Journal of Econometrics 211, 137-150.

Supplemental Material: Implementing the DiD Estimators in Stata

The Stata data files and associated "do" files show how to implement most of the estimators described in Sections 5, 6, and even 7. The time-varying treatment indicator, w_{it}, is included in the data set. In my view, this variable always should be included to allow for simple calculation of the average treatment effects and standard errors, and, as discussed in Section 6.3, to make it easy to impose restrictions on treatment effects. In the common entry date case, it is then easy to obtain d_i and in the staggered entry case it is easy to define the cohort treatment indicators.

The statistical analysis implicit in the estimation assumes independence across i and that N is large enough and T not so large so that clustering for serial correlation (and heteroskedasticity) can be justified. Moreover, in the staggered case, if there are few new entries into treatment for a given time period then the standard errors may be misleading. As discussed in Section 6.3, one might have to assume some homogeneity across intensity or time in order to obtain reliable standard errors.

Speaking of standard errors, the various procedures that produce identical estimates do not always produce identical standard errors – but they should be close. The differences are due to degrees-of-freedom adjustments when including extra time effects or different conventions across commands on how to adjust for degrees of freedom.

Common Entry Date

Suppose first that $T = 2$ and that the panel is balanced. The covariates do not change over time. The following commands all give the same estimates of the ATT as the coefficient on w. For simplicity, they are shown for a single covariate. Multiple covariates is easily handled by

including all of the terms for the single covariate.

```
xtset cid tid
sum x if d
gen x_dm = x - r(mean)
reg y i.w i.w#c.x_dm i.d i.f2 x
   i.d#x i.f2#c.x, vce(cluster cid)
reg D.y w i.w#c,x_dm x, vce(robust)
xtreg y i.w i.w#c.x_dm i.f2 i.f2#x, fe vce(cluster cid)
xtreg y i.w i.w#x_dm i.d i.f2 x
   i.d#c.x i.f2#c.x, re vce(cluster cid)
gen dy = D.y
teffects ra (dy x) (w), atet
```

The final `teffects` command has the virtue of providing a standard error that adjusts for the sample averages (over the treated subsample) of the covariates. Alternatively, one could forego the centering and use the `margins` command with the `vce(uncond)` option. For example,

```
xtreg y i.w i.w#c.x i.f2 i.f2#c.x, fe vce(cluster cid)
margins, dydx(w) vce(uncond) subpop(if d == 1)
```

With an unbalanced panel, TWFE is preferred because it allows correlation between sample selection and the unobserved heterogeneity.

For any T and any number of treatment periods, let $post_t$ be the binary indicator for the post intervention periods. The following commands are equivalent:

```
reg y i.w i.w#c.x_dm i.d i.post x i.d#c.x
```

```
   i.post#c.x, vce(cluster cid)
xtreg y i.w i.w#i.c.x_dm i.post i.post#c.x, fe vce(cluster
cid)
xtreg y i.w i.w#i.c.x_dm i.tid i.tid#c.x, fe vce(cluster cid)
xtreg y i.w i.w#i.c.x i.post i.post#xc, fe vce(cluster cid)
margins, dydx(w) vce(uncond) subpop(if d == 1)
```

In the supplemental materials, I generated a data set with $T = 4$ with two control and two treated periods, with common entry. The data are in the Stata file did_4.dta. There is nothing important about $T = 4$ or the fact that the number of control and treated time periods are the same. The Stata do file, did_4.do, can be used to verify equivalences among the various estimators discussed in Section 5. Moreover, it shows how to test the null of common trends.

In addition to the linear model – applied to the variable $\log(y_{it})$ – an exponential model is applied to y_{it}. In this case, using fixed effects Poisson and a pooled Poisson estimator are identical only when there are no covariates.

Staggered Entry

Assume that the first treated cohort is at $t = q$ and the cohorth dummies have been defined. Below, they are `dq`, `dqp1`, ..., `dT`. Again, for simplicity assume a single covariate. I show only the fixed effects version of the command where the covariates have been demeaned.

```
xtset cid tid
sum x if dq
gen x_dmq = x - r(mean)
...
sum x if dT
```

```
gen x_dmT = x - r(mean)
xtreg y c.dq#c.fq ... c.dq#c.fT
   c.dqp1#c.fqp1 ... c.dqp1#c.fT ... c.dT#c.fT
   c.dq#c.fq#c.x_dmq ... c.dq#c.fT#c.x_dmT
   c.dqp1#c.fqp1#c.x_dmqp1 ... dqp1#c.fT#c.x_dmqp1
   ... dT#c.fT#c.x_dmT
   fq ... fT c.fq#c.x ... c.fT#c.x, fe vce(cluster cid)
```

I generated a staggered intervention with six time periods, with three acting as the control and three as the treatment. In period four (2014), some units are treated for the first time. In periods five and six (2015 and 2016), additional units are treated. Many units are never treated. The Stata data set is staggered_6.dta, and this can be used to run the code in staggered_6.do. As in the previous case, both linear and exponential models are estimated depending on whether the dependent variable is $\log(y_{it})$ or $y_{it} > 0$. Also, a binary variable is generated to show how binary response models can easily replace the linear model estimated by pooled OLS. Fractional responses are handled similarly, using either the Stata `glm` command or `fracreg`.