Empirical Economics (2025) 69:2545–2587
https://doi.org/10.1007/s00181-025-02807-z

Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators

Jeffrey M. Wooldridge[1]

Received: 27 December 2024 / Accepted: 30 July 2025 / Published online: 27 August 2025

Abstract

I derive a result on 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—the two-way Mundlak (TWM) regression. The equivalence between TWFE and TWM implies that various estimators used for intervention analysis can be computed using pooled OLS that controls for time-constant treatment cohort indicators, time-period indicators, covariates, and interactions among them—allowing for considerable treatment effect heterogeneity. An extended version of TWFE (ETWFE) is equivalent to the POLS approach. I show that an imputation estimator, derived under no anticipation and parallel trends assumptions, is also equivalent to the POLS/ETWFE estimator. The equivalence among various estimators shows that average treatment effects on the treated are identified by flexible regression. The framework allows for event study estimators, which can be used to test for pre-trends, and flexible estimation that allows for cohort-specific trends.

Keywords Panel data · Two-way fixed effects · Two-way mundlak regression · Staggered interventions · Heterogeneous treatment effects · Heterogeneous trends

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-sectional 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 with explanatory variables that vary across i (the cross section dimension) and

✉ Jeffrey M. Wooldridge
wooldri1@msu.edu

[1] Department of Economics, Michigan State University, 486 West Circle Drive, East Lansing 48824-1038, Michigan, USA

t (the time series dimension) 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. 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) feasible GLS. Wooldridge (2019) provides a recent analysis, showing that an entire class of regressions— including pooled OLS—reproduces 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 makes it clear that the pre-testing problem in choosing between RE and FE is essentially the same as pre-testing on a set of regressors.

In exploiting the equivalence between the OWFE 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 this paper, I explicitly consider the two-way FE (TWFE) estimator and show that a simple extension of the Mundlak device reproduces the TWFE estimates. In particular, adding both unit-specific time series averages and period-specific cross-sectional averages in a POLS regression reproduces 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. After writing the first version of this paper, I became aware of the independent work of Yang (2022), who also established the equivalence of the TWFE and TWM estimators, including extensions to more than two sets of fixed effects. In the two-way case, my result is more general in that I allow for both time-constant controls and controls that only vary across time.

The TWFE–TWM equivalence has several applications. On a fundamental 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 and including covariates of much lower dimension. My main application of this equivalence is to very flexible models in the context of staggered interventions.

To be sure that POLS/TWFE identifies interesting parameters with staggered interventions, in Sect. 4 I show that the average treatment effects on the treated (ATTs) are identified under standard assumptions used in difference-in-differences designs with control variables. The assumptions motivate a two-step imputation estimator. A drawback to imputation is that inference is cumbersome due to two-step estimation. Moreover, we do not directly estimate how the treatment effects depend on the observed covariates—what are often called "moderating" effects in some literatures. Fortunately, in Sect. 5 I am able to demonstrate that the pooled OLS estimator obtained from the TWM estimation is numerically identical to the imputation estimates. An important conclusion is that several different estimation approaches lead to exactly the same set of estimates: imputation using treatment cohort dummies, pooled OLS using cohort dummies, random effects using cohort dummies, two-way fixed effects, and an imputation estimator based on two-way fixed effects. Section 5.5 provides details after all estimators have been defined.

An important consequence of the identification and equivalence results in this paper is that there is nothing wrong with the TWFE estimator in the context of staggered interventions provided one applies the TWFE estimator to a sufficiently flexible model. This straightforward point sometimes gets lost in the research on staggered interventions. I provide a careful discussion in Sect. 5.2

The common estimator developed in Sects. 4 and 5 is sometimes called a "lags only" estimator because it only includes contemporaneous and lagged treatment indicators (and interactions of these with covariates). So-called "event study," or "leads and lags," estimators include lead indicators along with contemporaneous and lagged treatment indicators. In Sect. 6, I extend the equivalence results to leads and lags estimators, with covariates appearing flexibly. Again, an imputation method, regression on cohort dummies, TWFE, and random effects are all the same estimator. In effect, the analysis shows how to flexibly allow controls in the Sun and Abraham (2021)[SA (2021)] TWFE approach. Moreover, this estimator turns out to be identical to the 2 × 2 regression-based estimators in Callaway and Sant'Anna (2021)[CS (2021)]. One benefit of the event study approach is that it allows testing for the presence of pre-trends; typically, a finding of differing trends between treated and control before the intervention begins is taken to indicate parallel trends is violated in the post-treatment period.

With many time periods and treatment cohorts, the number of estimated ATTs can be large, and they need not be precisely estimated for some treated cohorts. In Sect. 7, I show how the estimates can be aggregated to either a single estimated effect or effects estimated by exposure time. This is possible for both the lags only and leads and lags estimators.

The methods proposed through Sect. 7 allow for heterogeneous trends as a function of observed control variables. With enough pre-treatment period, we can use parametric models of heterogeneous trends, where each treated cohort is allowed to have its own deviation from the unrestricted baseline trend. Typically, one might use linear trends, at least as a robustness check. I show how to do this in Sect. 8. The equivalences among all estimators continues to hold in this case.

In Sect. 9, I apply three versions of the regression-based estimators to the problem of estimating the effects of locating a Walmart store on county retail employment. In this application, the parallel trends assumption is clearly violated, and allowing for heterogeneous trends by treatment cohort gives notably different estimates. Section 10 contains brief discussions of two additional topics: time-varying controls and unbalanced panels. Section 11 contains some concluding remarks. An appendix includes proofs of the results.

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. 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 key results in this section and the next are purely algebraic. In fact, there is no need to write down an underlying model; Eq. (2.1) is for motivational purposes only.

To describe the TWFE estimator, for each i define the set of binary unit indicators as ch_i, $h = 1, \ldots, N$, where $ch_i = 1$ if $h = i$, $ch_i = 0$ if $h \neq i$. 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

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

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 $f2_t$, ..., fT_t. For the purposes of this paper, we are not interested in these coefficients: the unit and time dummies act as controls.

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, for example, Baltagi (2021). To describe the procedure, define the unit-specific averages over time as

$$\bar{\mathbf{x}}_{i\cdot} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it} \tag{2.3}$$

and time-specific averages across i as

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

The overall average is $\bar{\mathbf{x}} = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it} = N^{-1}\sum_{i=1}^{N}\bar{\mathbf{x}}_{i\cdot} = T^{-1}\sum_{t=1}^{T}\bar{\mathbf{x}}_{\cdot t}$. Define

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

As shown in Baltagi (2021), $\hat{\boldsymbol{\beta}}_{FE}$ is the pooled OLS estimator from

$$y_{it} \text{ on } \ddot{\mathbf{x}}_{it}, t = 1, \ldots, T;\ i = 1, \ldots, N. \tag{2.6}$$

By standard OLS algebra, the same estimates are obtained by replacing y_{it} with $\ddot{y}_{it}$, where $\ddot{y}_{it} = y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{y}$.

3 The two-way Mundlak regression

Mundlak (1978) showed that, with $\mathbf{x}_{it}$ including explanatory variables that vary across (i, t) or t only (including 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 in that they produce the same coefficients on $\mathbf{x}_{it}$. The one of primary interest for this paper is the pooled OLS estimator from

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

However, here we want to explicitly separate the time-period dummies from the elements of $\mathbf{x}_{it}$. Henceforth, $\mathbf{x}_{it}$ includes only variables that have some variation across both i and t.

As shown by Wooldridge (2019), adding time-constant variables, say $\mathbf{z}_i$, does not change the coefficients on $\mathbf{x}_{it}$. Here, we are interested in extending the equivalence result to a *two-way Mundlak* (TWM) regression. The TWM regression adds the cross-sectional averages, $\bar{\mathbf{x}}_{\cdot t}$ to (3.1). For later applications, it is useful to allow for variables that do not change over time, $\mathbf{z}_i$, and those that change only across t, $\mathbf{m}_t$:

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

The following result extends Wooldridge (2010, Proposition 2.1) to the two-way FE setting.

Theorem 3.1 *Let $\hat{\boldsymbol{\beta}}_{TWM}$ be the $K \times 1$ vector of coefficients on $\mathbf{x}_{it}$ from the two-way Mundlak regression in (3.2), and let $\hat{\boldsymbol{\beta}}_{TWFE}$ be the TWFE estimator from the regression in (2.2) [equivalently, (2.6)], assuming that*

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

is nonsingular. Then, $\hat{\boldsymbol{\beta}}_{TWM} = \hat{\boldsymbol{\beta}}_{TWFE}$. Moreover, the coefficients on $\mathbf{x}_{it}$ do not change when any subset of $(\mathbf{z}_i, \mathbf{m}_t)$ is dropped from (3.2). □

The proof of Theorem 3.1 is given in the appendix; it differs from Yang (2022) by using partialling out arguments, and it also accommodates the additional variables

$(\mathbf{z}_i, \mathbf{m}_t)$. Yang (2022) also shows that applying a two-way random effects estimator that includes the time averages and cross-sectional averages also reproduces the TWFE estimator. In independent work, Baltagi (2023) reached the same conclusion. Baltagi and Wansbeek (2025) contain a detailed discussion of multi-way Mundlak regressions and various equivalences, and they use these to derive specification tests.

That adding $(\mathbf{z}_i, \mathbf{m}_t)$ does not change the equivalence between $\hat{\boldsymbol{\beta}}_{TWM}$ and $\hat{\boldsymbol{\beta}}_{TWFE}$ follows intuitively because the TWFE estimation eliminates $(\mathbf{z}_i, \mathbf{m}_t)$ and therefore their presence cannot affect the coefficients on $\mathbf{x}_{it}$. The coefficients on $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$ will change as elements of $(\mathbf{z}_i, \mathbf{m}_t)$ are included, and this is important for specification testing. The focus is typically on correlation between unit-specific heterogeneity—c_i in (2.1)—and the elements of $\mathbf{x}_{it}$, in which case on can obtain a cluster-robust Hausman (1978) test by testing significance of $\bar{\mathbf{x}}_{i\cdot}$.

The following corollary is an immediate application of Wooldridge (2019, Proposition 2.1).

Corollary 3.2 *The pooled OLS estimator* $\hat{\boldsymbol{\beta}}_{TWM}$ *from (3.2) is identical to the one-way random effects (RE) GLS estimator (with a cross-sectional "random effect") using the same regressors as in (3.2). Moreover, the remaining coefficients from the POLS and RE estimators are identical.* □

Wooldridge (2019, Proposition 2.1) shows that the equivalence holds even 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—given that the averages $\bar{\mathbf{x}}_{i\cdot}$ and $\bar{\mathbf{x}}_{\cdot t}$ are included. Of course, it is also asymptotically efficient (fixed T, $N \to \infty$) under those same assumptions; see Wooldridge (2010, Section 10.4).

Theorem 3.1 has simple but useful implications. Suppose, for example, that x_{itj} can be expressed as an interaction between a time-constant variable and time-varying variable:

$$x_{itj} = z_{ij} \cdot m_{tj}. \tag{3.4}$$

Then,

$$\bar{x}_{i\cdot j} = z_{ij} \cdot \bar{m}_j, \ \bar{x}_{\cdot tj} = \bar{z}_j m_{tj}, \tag{3.5}$$

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}). Moreover, in models where the only time-varying variables have the interactive form (3.4), POLS and RE will be identical when z_{ij} and m_{tj} are included in the equation. We will use these simple observations to establish equivalence between various estimators in the context of staggered interventions.

4 Identification with staggered interventions and imputation

We now turn to the problem of estimating average treatment effects in the context of staggered interventions. We are primarily thinking of cases where the total number of time periods, T, is small relative to the available cross section units. We impose the "no reversibility" assumption, as in much previous work, including Callaway and Sant'Anna (2021) and Borusyak,Jaravel and Spiess (2024).

Let $q \in \{2, \ldots, T\}$ be the first time of treatment, where some units are exposed to an intervention. At period $q+1$, more units join the treated group, and so on, through period T. We assume that $q > 1$ so there is at least one pre-treatment period. One way to view the staggered intervention is that it generates different levels of exposure to the treatment, as determined by the first date of the intervention. To that end, define treatment cohort dummies, dq, ..., dT, that indicate when a unit is first subjected to the intervention. Given a never treated group, there are $T-q+2$ total treatment levels. In Sect. 5.4, I discuss how to handle situations where all units are eventually treated.

Given staggered entry, it is natural to think in terms of potential outcomes $y_t(g)$. For $g \in \{q, q+1, \ldots, T\}$, $y_t\,(g)$ is the potential outcome during time period t if a unit enters the treated state in time period g. We also need a notation for the potential outcome if a unit is not subjected to the intervention during the time periods under study, $\{1, 2, \ldots, T\}$. I adopt the notation in Athey and Imbens (2022) and SA (2021) and use $y_t(\infty)$ to denote the potential outcome in time t if a unit is never treated (during the period under study).

For treated cohort g, we define treatment effects

$$te_t(g) = y_t\,(g) - y_t\,(\infty)\,, g = q, \ldots, T, \tag{4.1}$$

which we take to be random variables with distributions in the population. As in CS (2021), SA (2021), and much other work, the treatment effects we hope to identify are the ATTs in periods where the cohorts are actually subjected to the intervention:

$$\tau_{gt} \equiv E\left[te_t\,(g)\,|dg = 1\right], \;\; g = q, \ldots, T; \quad t = g, \ldots, T. \tag{4.2}$$

The requirement that we have at least one pre-treatment period means $q \geq 2$. If a unit enters treatment in period g then we can hope to estimate ATTs in periods $\{g, g+1, \ldots, T\}$.

The first assumption is common in environments that explicitly use a potential outcomes framework. It rules out spillover effects, and so it can be restrictive.

Assumption SUTVA (Stable Unit Value Treatment Assumption): The potential outcome of each unit in the population does not depend on the treatment assignment of other units in the population. □

Although inference is not a focus in this paper, later we assume that the potential outcomes, treatment indicators, and controls are obtained by independently sampling from the cross-sectional units from a population. That setting implies SUTVA, although SUTVA can hold without independent sampling.

As in CS (2021), we relax the strict form of the parallel trends assumption by allowing trends to depend on time-constant controls, $\mathbf{x}$. Assuming these covariates do not change over time is not completely general, but it helps to ensure that these controls are not influenced either by the intervention. Nevertheless, this assumption should be explicitly stated.

Assumption NBC (No Bad Controls): Letting $\mathbf{x}(g)$ be the covariates when the treatment cohort is $g \in \{q, \ldots, T, \infty\}$, assume that $\mathbf{x}(g) = \mathbf{x}(\infty)$, $g = q, \ldots, T$. In what follows, let $\mathbf{x} = \mathbf{x}(\infty)$. □

We also impose a form of "no anticipation," which rules out the possibility that units alter their behavior before being subjected to the intervention. In stating the assumption, define the vector of cohort indicators $\mathbf{d} = (dg, \ldots, dT)$.

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

$$E\left[y_t(g) - y_t(\infty) | \mathbf{d}, \mathbf{x}\right] = 0,\ t < g. \square \tag{4.3}$$

A stronger form of NA is often imposed: $y_t(g) = y_t(\infty)$ for $t < g$, which 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 $t < 3$.

To motivate the conditional parallel trends assumption, first consider the case without covariates. We can always write the cohort-specific means of the outcome in the first time period in the never treated state as

$$E\left[y_1(\infty) | \mathbf{d}\right] = \alpha + \sum_{g=q}^{T} \beta_g dg. \tag{4.4}$$

It is important to recognize that (4.4) is definitional: the intercept is $\alpha \equiv E\left[y_1(\infty) | d\infty = 1\right]$ and $\beta_g \equiv E\left[y_1(\infty) | dg = 1\right] - E\left[y_1(\infty) | d\infty = 1\right]$ for $g = q, \ldots, T$. Nonzero β_g allows the treatment cohort to be systematically related to the first-period outcome in the never treated state. Allowing selection into treatment based on the levels of pre-treatment outcomes is a hallmark of the difference-in-differences approach.

The (unconditional) parallel trends (PT) assumption can be stated as

$$E\left[y_t(\infty) - y_1(\infty) | dq, \ldots, dT\right] = E\left[y_t(\infty) - y_1(\infty)\right] \equiv \gamma_t, t = 2, \ldots, T. \tag{4.5}$$

In other words, while no restrictions are placed on the unconditional means $E\left[y_t(\infty)\right]$ across t, selection into treatment cannot be systematically related to the *trend* in the never treated state, $y_t(\infty) - y_1(\infty)$. Remember, $\beta_g \neq 0$ allows for selection based on the levels.

When we combine (4.4) and (4.5), we obtain

$$E\left[y_t(\infty) | dq, \ldots, dT\right] = \alpha + \sum_{g=q}^{T} \beta_g dg + \gamma_t,\ t = 1, \ldots, T,$$

where $\gamma_1 \equiv 0$ is a normalization because an overall intercept is included. Letting fs_t denote time-period dummies, so that $fs_t = 1$ if and only if $s = t$, we can write

$$E\left[y_t\left(\infty\right)|dq,\ldots,dT\right]=\alpha+\sum_{g=q}^{T}\beta_g dg+\sum_{s=2}^{T}\gamma_s fs_t,\ t=1,\ldots,T. \tag{4.6}$$

Equation (4.6) is very useful because it clarifies the nature of parallel trends. The cohort dummies, dg, appear with unrestricted coefficients, β_g, and the time-period dummies, fs_t, also appear with unrestricted coefficients. The exclusion of interactions $dg \cdot fs_t$ is identical to the parallel trends assumption.

When covariates $\mathbf{x} = (x_1, \ldots, x_K)$ are available, the conditional parallel trends are stated as follows.

Assumption CPT (Conditional Parallel Trends): For $t = 2, \ldots, T$ and time-constant controls $\mathbf{x}$,

$$E\left[y_t(\infty)-y_1(\infty)|\mathbf{d},\mathbf{x}\right]=E\left[y_t\left(\infty\right)-y_1\left(\infty\right)|\mathbf{x}\right].\ \Box \tag{4.7}$$

The CPT assumption means that, conditional on the covariates, the cohort assignment, $\mathbf{d}$, is not systematically related to the trend, $y_t(\infty) - y_1(\infty)$. However, the trends can be a function of $\mathbf{x}$. A useful way to think of this assumption is to imagine partitioning the population into exhaustive and mutually exclusive strata based on $\mathbf{x}$. (This thought experiment is easier when $\mathbf{x}$ is discrete.) Then, we require PT to hold within each stratum. Across the entire population, PT need not hold if cohort assignment is correlated with $\mathbf{x}$.

Another useful characterization of Assumption CPT is as an unconfoundedness assumption, but it is in terms of *changes* in the potential outcome across time, not the level. Equation (4.7) means that $\mathbf{d}$ is unconfounded with respect to $y_t(\infty) - y_1(\infty)$ conditional on $\mathbf{x}$, for $t = 2, \ldots, T$. Importantly, assignment is allowed to be confounded in terms of the level, $y_t(\infty)$ (whether or not we condition on $\mathbf{x}$).

To make estimating equations straightforward, we assume linear functional forms in the control variables.

Assumption LIN (Linearity): For treatment cohort indicators dg, $g \in \{q, \ldots, T\}$, and control variables $\mathbf{x}$,

$$E\left[y_1\left(\infty\right)|\mathbf{d},\mathbf{x}\right]=\alpha+\sum_{g=q}^{T}\beta_g dg+\mathbf{x}\boldsymbol{\kappa}+\sum_{g=q}^{T}\left(dg\cdot\mathbf{x}\right)\boldsymbol{\xi}_g \tag{4.8}$$

$$E\left[y_t\left(\infty\right)|\mathbf{d},\mathbf{x}\right]-E\left[y_1\left(\infty\right)|\mathbf{d},\mathbf{x}\right]=\sum_{s=2}^{T}\gamma_s fs_t$$
$$+\sum_{s=2}^{T}\left(fs_t\cdot\mathbf{x}\right)\pi_s,\ t=2,\ldots,T.\ \Box \tag{4.9}$$

As noted above, when there are no controls, (4.8) is definitional and (4.9) is equivalent to PT. With covariates $\mathbf{x}$, (4.8) and (4.9) both generally impose functional form restrictions. Equation (4.9) implies CPT because $\mathbf{d}$ does not appear on the right-hand side.

One case where no functional form assumptions are imposed in (4.8) and (4.9) is when $\mathbf{x}$ only includes exhaustive and mutually exclusive dummy variables. For example, perhaps we measure the outcome at the school, city, or census tract level, and $\mathbf{x}$ includes only exhaustive (less one) and mutually exclusive indicators of some higher level of geographical location. Then, (4.8) is without loss of generality, and (4.9) imposes parallel trends conditional on location. The more "good" controls we include in $\mathbf{x}$, the more likely (4.9) is to hold. In general, we might have rich set of controls, including those that are essentially continuous. Naturally, $\mathbf{x}$ can include the usual functional forms that often included in regression analysis, such as squares, interactions, logarithms, and so on. As in typical regression frameworks, the key is linearity in the parameters.

Combining (4.8) and (4.9) gives a very useful equation for the conditional expectation in the never treated state across time:

$$\begin{aligned} E\left[y_t(\infty) \mid \mathbf{d}, \mathbf{x}\right] = \alpha + \sum_{g=q}^{T} \beta_g dg + \mathbf{x}\boldsymbol{\kappa} + \sum_{g=q}^{T} \left(dg \cdot \mathbf{x}\right) \boldsymbol{\xi}_g \\ + \sum_{s=2}^{T} \gamma_s f s_t + \sum_{s=2}^{T} \left(f s_t \cdot \mathbf{x}\right) \boldsymbol{\pi}_s, \quad t = 1, \ldots, T \end{aligned} \tag{4.10}$$

This equation is our basis for linear, regression-based approaches to DiD with staggered interventions, time-constant covariates, and lots of heterogeneity. The equation allows selection into treatment when $\beta_g \neq 0$ or $\boldsymbol{\xi}_g \neq \mathbf{0}$ and it allows for heterogenous trends in the never treated state when $\boldsymbol{\pi}_s \neq \mathbf{0}$. In Wooldridge (2023), I extend (4.10) to allow nonlinear functions of the linear index that are logically consistent with the nature of $y_t\left(\infty\right)$.

We can use equation (4.10) to identify the τ_{gt} when we add Assumption CNA. To see how, write

$$\begin{aligned} \tau_{gt} &= E\left[y_t\left(g\right)|dg = 1\right] - E\left[y_t\left(\infty\right)|dg = 1\right] \\ &= E\left(y_t|dg = 1\right) - \left[\alpha + \beta_g + \gamma_t + E\left(\mathbf{x}|dg = 1\right) \cdot \left(\boldsymbol{\kappa} + \boldsymbol{\xi}_g + \boldsymbol{\pi}_t\right)\right]. \end{aligned} \tag{4.11}$$

Note how (4.11) also uses Assumption NBC because (4.11) implicitly assumes $E\left[\mathbf{x}\left(\infty\right)|dg = 1\right] = E\left[\mathbf{x}\left(g\right)|dg = 1\right] = E\left(\mathbf{x}|dg = 1\right)$. Without this assumption, we would not be able to estimate $E\left[\mathbf{x}\left(\infty\right)|dg = 1\right]$. The first term in (4.11) follows from the fact that $y_t = y_t\left(g\right)$ when $dg = 1$; $E\left(y_t|dg = 1\right)$ is identified because it is the mean of the observed response y_t for treatment cohort g in time period t. Also, $E\left(\mathbf{x}|dg = 1\right)$ is identified because it is the average of the covariates over cohort g. Consequently, identification of τ_{gt} holds if the parameters α, β_g, γ_t, $\boldsymbol{\kappa}$, $\boldsymbol{\xi}_g$, and $\boldsymbol{\pi}_t$ are identified.

To see how all parameters in (4.10) are identified, define the time-varying binary treatment variable

$$\begin{aligned} w_t &= dq\cdot(fq_t+\cdots+fT_t)+d\left(q+1\right)\cdot\left[f\left(q+1\right)_t+\cdots+fT_t\right]+\cdots+dT\cdot fT_t \\ &= dq\cdot pq_t+d\left(q+1\right)\cdot p\left(q+1\right)_t+\cdots+dT\cdot fT_t, \end{aligned}$$

where $ps_t = fs_t + \cdots + fT_t$ is a post-intervention indicator for treatment that starts in period s. Note that w_t is the sum of mutually exclusive treatment dummies of the form $dg \cdot fs_t$, indicating treatment cohort g in period s. Also, unlike the cohort indicators dg, w_t varies over time for a treated unit. Because there is no reversibility, $\{w_t : t = 1, 2, \ldots, T\}$ is a sequence of zeros followed by a sequence of ones, with the first one appearing for cohort g when $t = g$.

By the conditional no anticipation assumption,

$$E\left[y_t\left(g\right)|\mathbf{d},\mathbf{x},w_t=0\right]=E\left[y_t\left(\infty\right)|\mathbf{d},\mathbf{x},w_t=0\right].$$

In other words, for the observed outcome y_t,

$$\begin{aligned} E\left(y_t \mid \mathbf{d},\mathbf{x},w_t=0\right) &= \alpha+\sum_{g=q}^{T}\beta_g dg+\mathbf{x}\boldsymbol{\kappa}+\sum_{g=q}^{T}\left(dg\cdot\mathbf{x}\right)\boldsymbol{\xi}_g \\ &\quad+\sum_{s=2}^{T}\gamma_s fs_t+\sum_{s=2}^{T}\left(fs_t\cdot\mathbf{x}\right)\boldsymbol{\pi}_s \end{aligned} \tag{4.12}$$

Because (4.12) is a conditional expectation linear in parameters, and all variables are observed, (4.12) shows identification of the parameters if we rule out perfect collinearity in

$$\left(1, dq, \ldots, dT, \mathbf{x}, dq\cdot\mathbf{x}, \ldots, dT\cdot\mathbf{x}, f2_t, \ldots, fT_t, f2_t\cdot\mathbf{x}, \ldots, fT_t\cdot\mathbf{x}\right),$$

which essentially requires that we have some units in every treated cohort and no perfect collinearity in $\mathbf{x}$. We also need to have a sampling scheme that allows us to consistently estimate the population parameters. While other sampling schemes would suffice, including stratifies sampling and cluster sampling, for simplicity we assume random sampling in the cross section. In other words, let $\{[(y_{it}, t = 1, \ldots, T), \mathbf{d}_i, \mathbf{x}_i] : i = 1, \ldots, N\}$ be an independent, identically distributed sample of size N. We do not restrict the dependence over time. Then, if (4.12) holds, the ordinary least squares (OLS) estimators using only the control observations $w_{it} = 0$ will be unbiased (conditional on the treatment assignments and covariates) and consistent with fixed T and $N \to \infty$.

As mentioned earlier, $E(y_t|dg = 1)$ is always identified, and a consistent estimator is the cohort average in each treated period t:

$$\bar{y}_{gt} = N_g^{-1}\sum_{i=1}^{N} dg_i \cdot y_{it}, \tag{4.13}$$

where $N_g = \sum_{i=1}^{N} dg_i$ is the number of units in treatment cohort g. This estimator is even unbiased conditional on having at least one treated unit in cohort g. We can combine this estimator with an imputation estimator for $E\left[y_t(\infty)\,|dg = 1\right]$, which requires estimating the parameters in (4.12) and then using the sample analog of Eq. (4.12). We call this "cohort imputation" because the regression includes cohort dummy variables.

Procedure 4.1 ***(Cohort Imputation):***

(i) Using the control observations, $w_{it} = 0$, run an OLS regression

$$\begin{aligned} y_{it} \text{ on } 1,\, dq_i, \ldots, dT_i,\, \mathbf{x}_i,\, dq_i \cdot \mathbf{x}_i, \ldots, dT_i \cdot \mathbf{x}_i, \\ f2_t, \ldots, fT_t,\, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i \end{aligned} \tag{4.14}$$

to obtain the parameter estimates

$$\left(\tilde{\alpha}, \tilde{\beta}_q, \ldots, \tilde{\beta}_T, \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\xi}}_q, \ldots, \tilde{\boldsymbol{\xi}}_T, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_T, \tilde{\boldsymbol{\pi}}_2, \ldots, \tilde{\boldsymbol{\pi}}_T\right). \tag{4.15}$$

(ii) Impute the missing outcomes in the NT state, $y_{it}(\infty)$, as

$$\begin{aligned} \tilde{y}_{it}(\infty) = \tilde{\alpha} + \sum_{g=q}^{T} \tilde{\beta}_g dg_i + \mathbf{x}_i \tilde{\boldsymbol{\kappa}} + \sum_{g=q}^{T} (dg_i \cdot \mathbf{x}_i)\, \tilde{\boldsymbol{\xi}}_g \\ + \sum_{s=2}^{T} \tilde{\gamma}_s f s_t + \sum_{s=2}^{T} (f s_t \cdot \mathbf{x}_i)\, \tilde{\boldsymbol{\pi}}_s \end{aligned} \tag{4.16}$$

and the unit-specific TEs as

$$\widetilde{te}_{it} = y_{it} - \tilde{y}_{it}(\infty). \tag{4.17}$$

(iii) Obtain the estimated ATT for cohort g in period t as

$$\begin{aligned} \tilde{\tau}_{gt} &= N_g^{-1} \sum_{i=1}^{N} dg_i \widetilde{te}_{it} = \bar{y}_{gt} - N_g^{-1} \sum_{i=1}^{N} dg_i \cdot \tilde{y}_{it}(\infty) \\ &= \bar{y}_{gt} - \left[\left(\tilde{\alpha} + \tilde{\beta}_g + \tilde{\gamma}_t\right) + \bar{\mathbf{x}}_g \cdot \left(\tilde{\boldsymbol{\kappa}} + \tilde{\boldsymbol{\xi}}_g + \tilde{\boldsymbol{\pi}}_t\right)\right], \end{aligned} \tag{4.18}$$

where

$$\bar{\mathbf{x}}_g = N_g^{-1} \sum_{i=1}^{N} dg_i \cdot \mathbf{x}_i \tag{4.19}$$

is the $1 \times K$ row vector of cohort-specific averages. □

Under Assumptions CNA, NBC, CPT, LIN, and random sampling across i (which implies SUTVA), $\tilde{\tau}_{gt}$ is a consistent estimator of τ_{gt} as $N \to \infty$ because

$$\left(\tilde{\alpha} + \tilde{\beta}_g + \tilde{\gamma}_t\right) + \bar{\mathbf{x}}_g \cdot \left(\tilde{\boldsymbol{\kappa}} + \tilde{\boldsymbol{\xi}}_g + \tilde{\boldsymbol{\pi}}_t\right) \xrightarrow{p} E\left[y_t(\infty)\,|dg = 1\right]; \tag{4.20}$$

see Eq. (4.11). These estimators are also unbiased conditional on the treatment indicators and covariates because POLS is conditionally unbiased when the linear (in parameters) conditional mean is correctly specified (and we rule out perfect collinearity in the sample). Because the estimates are obtaining using control observations, the $\widetilde{te}_{it}$ are out-of-sample residuals (or prediction errors). One can average over other subgroups to obtain other ATTs. For example, if the $\widetilde{te}_{it}$ are averaged over all treated observations $w_{it} = 1$, the resulting average is a weighted average of all $\tilde{\tau}_{gt}$, where the weight for $\tilde{\tau}_{gt}, t = g, \ldots, T$ depends on the cohort share, N_g/N.

Without covariates $\mathbf{x}_i$, this cohort imputation method is the same as that proposed by Gardner (2022)—which he calls "two-stage difference-in-differences." When controls are added, notice that the first step of the imputation method does not simply add the $\mathbf{x}_i$, but also interacts each element with the cohort dummies dg_i and the time-period dummies fs_t, allowing for flexible selection into treatment and violation of parallel trends as a function of $\mathbf{x}_i$.

A practical drawback of Procedure 4.1 is that calculation of analytical standard errors for the $\tilde{\tau}_{gt}$ is complicated by the multiple estimation steps. Of course, (4.18) can be used to obtain a valid asymptotic variance using the law of large numbers and central limit theorem, but the calculations must account for correlations among the first-step OLS estimators and the sample average $\bar{y}_{gt}$. Moreover, we often want to perform inference on aggregated effects—for example, by exposure time—which requires, at least implicitly, estimating the asymptotic covariances among the estimators—a process further hampered by separately averaging over the (g, t) pairs. Fortunately, as shown in the next section, inference can be obtained using standard regression packages.

5 Pooled OLS and extended two-way fixed effects

5.1 A pooled OLS estimator using all data

Rather than using an imputation approach based on the potential outcomes setting, suppose we bypass the potential outcomes framework and specify a flexible conditional mean for all time periods, using the observed outcome y_{it}:

$$
\begin{aligned}
E\left(y_{it} | dq_i, \ldots, dT_i, \mathbf{x}_i\right) = \alpha &+ \sum_{g=q}^{T} \beta_g dg_i + \mathbf{x}_i \boldsymbol{\kappa} + \sum_{g=q}^{T} \left(dg_i \cdot \mathbf{x}_i\right) \boldsymbol{\xi}_g \\
&+ \sum_{s=2}^{T} \gamma_s f s_t + \sum_{s=2}^{T} \left(f s_t \cdot \mathbf{x}_i\right) \boldsymbol{\pi}_s \\
&+ \sum_{g=q}^{T} \sum_{s=g}^{T} \tau_{gs} \left(w_{it} \cdot dg_i \cdot f s_t\right) \\
&+ \sum_{g=q}^{T} \sum_{s=g}^{T} \left(w_{it} \cdot dg_i \cdot f s_t \cdot \dot{\mathbf{x}}_{ig}\right) \boldsymbol{\delta}_{gs}
\end{aligned}
\tag{5.1}
$$

where $\dot{\mathbf{x}}_{ig} \equiv \mathbf{x}_i - E\left(\mathbf{x}_i | dg_i = 1\right)$. The first two lines in (5.1) are the same as (4.12), which led directly to Procedure 4.1. The presence of the treatment indicator w_{it} in the second line is redundant given that $w_{it} \cdot dg_i \cdot fs_t = dg_i \cdot fs_t$, but including it is useful for interpreting the equation. This equation is in the spirit of Wooldridge (2005), but I did not fully explore the implications of heterogeneous coefficients in the context of staggered interventions. Note that (5.1) uses τ_{gs} as the coefficients on the $w_{it} \cdot dg_i \cdot fs_t$, but we do not yet know if (5.1) identifies the ATTs under the assumptions imposed in Sect. 4.

To operationalize (5.1), we replace $E\left(\mathbf{x}_i | dg_i = 1\right)$ with $\bar{\mathbf{x}}_g$ and, reusing notation,

$$\dot{\mathbf{x}}_{ig} = \mathbf{x}_i - \bar{\mathbf{x}}_g, g = q, \ldots, T. \tag{5.2}$$

Procedure 5.1 ***(Pooled OLS on Cohort Dummies):***

(i) Demean the covariates about their cohort averages, as in (5.2).

(ii) With the treatment variables listed first, run the pooled OLS regression, across all i and t,

$$\begin{aligned}
&y_{it} \text{ on } w_{it} \cdot dq_i \cdot fq_t, \ldots, w_{it} \cdot dq_i \cdot fT_t, \ldots, \\
&\quad w_{it} \cdot d\left(q+1\right)_i \cdot f\left(q+1\right)_t, \ldots, w_{it} \cdot d\left(q+1\right)_i \cdot fT_t, \ldots, \\
&\quad w_{it} \cdot dT_i \cdot fT_t, \\
&\quad w_{it} \cdot dq_i \cdot fq_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, w_{it} \cdot dq_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iq}, \\
&\quad w_{it} \cdot d\left(q+1\right)_i \cdot f\left(q+1\right)_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, w_{it} \cdot d\left(q+1\right)_i \cdot fT_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, \\
&\quad w_{it} \cdot dT_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iT} \\
&\quad 1, \quad f2_t, \ldots, fT_t, \ f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, dq_i, \ldots, dT_i, \mathbf{x}_i, dq_i \cdot \mathbf{x}_i, \ldots, dT_i \cdot \mathbf{x}_i.
\end{aligned} \tag{5.3}$$

w_{it} is the time-varying treatment indicator
dg_i are the time-constant cohort indicators
fs_t are the time-period indicators
$\mathbf{x}_i$ are the time-constant control variables

(iii) Obtain $\hat{\tau}_{gs}$ as the coefficient on $w_{it} \cdot dg_i \cdot fs_t$ for $s = g, \ldots, T$, $g = q, \ldots, T$. □

If we use only the $w_{it} = 0$ observations in (5.3) then we obtain the regression in the first step of the imputation in Procedure 4.1. Therefore, it is not surprising that the common OLS coefficients in (5.3) and (4.14) are the same. More importantly, the estimates of the ATTs are the same.

Proposition 5.2 ***(Equivalence Between Cohort Imputation and POLS on Cohort Dummies):*** *Assuming no perfect collinearity among the regressors in (4.14),*

$$\hat{\tau}_{gt} = \tilde{\tau}_{gt}, g = q, \ldots, T; \ t = g, \ldots, T. \tag{5.4}$$

Moreover, the coefficients on the control variables, 1, $f2_t, \ldots, fT_t, f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, dq_i, \ldots, dT_i, \mathbf{x}_i, dq_i \cdot \mathbf{x}_i, \ldots, dT \cdot \mathbf{x}_i$, are identical across the two procedures. □

The appendix includes a general result that includes Proposition 5.1 as a special case.

Note that, for the coefficients on $w_{it} \cdot dg_i \cdot fs_t = dg_i \cdot fs_t$ to be estimates of the τ_{gs} (and equal to the imputation estimates), the covariates in $w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$ have been centered around $\bar{\mathbf{x}}_g$. Without the centering, the main effects would be attempts at estimating the ATTs at $\mathbf{x} = \mathbf{0}$—which is unlikely to be interesting (or even identified in a nonparametric sense).

The algebraic equivalence of the imputation approach and POLS using all of the data is convenient for several reasons. First, inference based on the POLS regression in (5.3) is straightforward because standard econometrics packages allow clustering standard errors at the unit i level to allow for unrestricted serial correlation and heteroskedasticity. If the data are obtained from cluster sampling, or if the treatment (cohort) assignment is clustered—as in Abadie, Athey, Imbens and Wooldridge (2023)—then one can cluster at a higher level. Moreover, in packages such as Stata, it is easy to account for the sampling variation in the $\bar{\mathbf{x}}_g$. By contrast, valid inference after step (iii) of Procedure 4.1 requires analytical adjustments for the multi-step estimation, or bootstrapping all steps of the procedure.

Because we can obtain a robust variance–covariance matrix estimator of $\hat{\boldsymbol{\tau}}$—the vector of all treatment ATT estimators—we can obtain valid standard errors and confidence intervals for various linear combinations of the τ_{gt}. One thing that is clear from both the imputation and pooled OLS approach is that τ_{gt} can be estimated only if there are sufficient units entering treatment in cohort g. In some cases, there may be cohorts where no additional units are treated, in which case $w_{it} \cdot dg_i \cdot fs_t$ simply gets dropped for all s and we do not estimate ATTs for that cohort. We also need to be cautious in using the usual cluster-robust variance matrix estimator in situations with few treated units per treated cohort.

The framework here provides formal justification for commonly used practices in empirical work. For example, the inclusion of $fs_t \cdot \mathbf{x}_i$ allows for unrestricted heterogenous trends at higher levels of aggregation than the unit i. In studies with annual county-level data, one often sees state × year interactions (or "state-by-year fixed effects") provided the policy assignment varies within state; with census tract data, one could include county × year interactions. This is achieved by including the appropriate dummy variables in $\mathbf{x}_i$.

Another attractive feature of (5.3) is that we obtain coefficients on $w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$, which show whether, and how, the ATTs change with the covariates. There is, however, a subtle point. The equivalence of $\hat{\tau}_{gt}$ and $\tilde{\tau}_{gt}$ in Proposition 5.1 is algebraic, and the assumptions in Sect. 4 are silent about the functional forms $E\left[te_t\left(g\right)|dg = 1, \mathbf{x}\right]$. In order for the $\hat{\boldsymbol{\delta}}_{gt}$ to consistently estimate the heterogeneous (or "moderating") effects, we would add the assumption that $E\left[te_t\left(g\right)|dg = 1, \mathbf{x}\right]$ is linear in $\mathbf{x}$. As linearity is often a good approximation—especially if the elements in $\mathbf{x}$ are chosen as flexible functions—in practice this seems like a minor requirement if the goal is to get a general idea about treatment effect heterogeneity. In any case, we have already assumed linearity of the conditional mean in the NT state.

I wrote the regression in (5.3) to include w_{it} because it emphasizes how flexible (5.3) is relative to a regression that includes w_{it} only on its own, estimating one treatment effect across all cohorts and time periods. Plus, (5.3) is very flexible in the covariates.

Interestingly, if the $\mathbf{x}_i$ only appear by themselves, or even along with $dg_i \cdot \mathbf{x}_i$, the ATT estimates will be the same as if $\mathbf{x}_i$ appears nowhere. Including the covariates changes the ATT estimates only when $fs_t \cdot \mathbf{x}_i$ or $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$ are included. When both sets are included along with $\mathbf{x}_i$ and $dg_i \cdot \mathbf{x}_i$, the resulting estimates of τ_{gt} are the same as the imputation estimates, and this is the recommended procedure unless the cohort groups are small.

As a practical matter, it can be convenient in applications to define a set of cohort-time-specific treatment dummies,

$$dgfs_{it} \equiv dg_i \cdot fs_t. \tag{5.5}$$

These are included in the regression and interacted with the appropriate $\dot{\mathbf{x}}_{ig}$. As mentioned above, including w_{it} is redundant for obtaining the point estimates, but having it makes aggregating the ATTs and obtaining proper standard errors relatively easy.

As already discussed, to force the coefficients on the $w_{it} \cdot dg_i \cdot fs_t$ treatment indicators to be the estimated ATTs, the covariates have been centered about the cohort means when interacting them with $w_{it} \cdot dg_i \cdot fs_t$. Often, one might want to use the same centering when interacting with just the cohort dummies, dg_i. Replacing $dg_i \cdot \mathbf{x}_i$ with $dg_i \cdot \dot{\mathbf{x}}_{ig}$ will make the coefficients on dg_i more meaningful. In effect, the coefficients on the dg_i will be estimates of an average “selection effect.” Naturally, centering will not affect the coefficients on $dg_i \cdot \mathbf{x}_i$ (or the estimated ATTs); it is purely for interpreting the coefficients on the dg_i.

There is less of a case for centering $\mathbf{x}_i$ before interacting with the time dummies, fs_t. However, to obtain an estimate of the average trend in the never treated state, one can use $fs_t \cdot (\mathbf{x}_i - \bar{\mathbf{x}})$ in place of $fs_t \cdot \mathbf{x}_i$, where $\bar{\mathbf{x}}$ is the overall sample average (because all units contributed to estimation in the untreated state under no anticipation). Alternatively, one might use $\bar{\mathbf{x}}_\infty$, the average of the never treated units.

5.2 An extended TWFE estimator

We can now apply the algebraic equivalences from Sect. 3 to show that other estimators are equivalent to POLS (and, therefore, cohort imputation). As motivation, consider an unobserved effects model with the variables rearranged so that the treatment variables appear first:

$$\begin{aligned} y_{it} = {} & \sum_{g=q}^{T}\sum_{s=g}^{T} \tau_{gs}\,(w_{it} \cdot dg_i \cdot fs_t) + \sum_{g=q}^{T}\sum_{s=g}^{T} \left(w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}\right) \boldsymbol{\delta}_{gs} \\ & + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\, \boldsymbol{\pi}_s + c_i + u_{it},\ t = 1, \ldots, T, \end{aligned} \tag{5.6}$$

where c_i is the unobserved unit effect. If this equation is estimated by fixed effects, the c_i are swept away by the within transformation. Equivalently, the two-way fixed effects characterization comes from the pooled OLS regression

$$\begin{aligned} &y_{it} \text{ on } w_{it}\cdot dq_i\cdot fq_t, \ldots, w_{it}\cdot dq_i\cdot fT_t, \ldots, w_{it}\cdot dT_i\cdot fT_t,\\ &w_{it}\cdot dq_i\cdot fq_t\cdot \dot{\mathbf{x}}_{iq}, \ldots, w_{it}\cdot dq_i\cdot fT_t\cdot \dot{\mathbf{x}}_{iq}, \ldots, dT_i\cdot fT_t\cdot \dot{\mathbf{x}}_{iT},\\ &f2_t\cdot \mathbf{x}_i, \ldots, fT_t\cdot \mathbf{x}_i,\ f2_t, \ldots, fT_t, c1_i, c2_i, \ldots, cN_i, \end{aligned} \tag{5.7}$$

where, again, the ch_i are unit-specific dummy variables. The variables multiplied by w_{it} are the same treatment variables as (5.3). The third line includes the terms that allow for heterogeneous trends as unrestricted linear functions of $\mathbf{x}_i$. This line also includes the time and unit fixed effects. With the inclusion of $c1_i$, $c2_i$, …, cN_i, the time-constant variables dq_i, …, dT_i, $\mathbf{x}_i$, $dq_i\cdot\mathbf{x}_i$, …, $dT_i\cdot\mathbf{x}_i$, and the constant in (5.3) are dropped because they are now redundant.

From Theorem 3.1, the two-way Mundlak regression produces identical estimates on all variables that have some variation across i and t—that is, all coefficients in the first six lines in (5.3). It is easily seen that the TWM regression that corresponds to TWFE estimation of (5.7) is identical to the regression in (5.3). This is easiest to see when we drop the redundant variable w_{it} from the equation. Then, the average across time of $dg_i\cdot fs_t$ is dg_i/T and, likewise, the average of $dg_i\cdot fs_t\cdot\dot{\mathbf{x}}_{ig}$ across t is $dg_i\cdot\dot{\mathbf{x}}_{ig}/T$. But dg_i and $dg_i\cdot\dot{\mathbf{x}}_{ig}$ already appear in (5.3). The average across i for each t of $dg_i\cdot fs_t$ is just a constant times fs_t, and fs_t is already in the TWM regression. And the cross-sectional average of $dg_i\cdot fs_t\cdot\dot{\mathbf{x}}_{ig}$ is just a vector of constants times fs_t. The time average of $fs_t\cdot\mathbf{x}_i$ is $\mathbf{x}_i/T$, and $\mathbf{x}_i$ already appears in (5.3). The cross-sectional average of $fs_t\cdot\mathbf{x}_i$ is just a vector of constants times fs_t. In other words, regression (5.3) accounts for all of the Mundlak terms required to apply Theorem 3.1, and so the coefficients on $w_{it}\cdot dg_i\cdot fs_t$, $w_{it}\cdot dg_i\cdot fs_t\cdot\dot{\mathbf{x}}_{ig}$, and $fs_t\cdot\mathbf{x}_i$ from the TWFE estimation (5.7) are identical to those in (5.3).

The equivalence between (5.3) and (5.7) has interesting implications. First, it shows that we need not control for unit-specific heterogeneity via $c1_i$, …, cN_i provided we include the cohort dummies dg_i, the controls $\mathbf{x}_i$, and the interactions between these. Without controls, it suffices to control for dg_i, …, dT_i rather than $c1_i$, …, cN_i. This is essentially a consequence of the parallel trends assumption. In many applications, the number of units in the dataset, N, is much larger than the number of treated cohorts. Even with many controls in $\mathbf{x}_i$, the number of regressors in (5.3) can be many fewer than in (5.7). For example, with $N = 1,000$, five treated cohorts, and ten control variables, (5.3) includes 60 time-constant controls (including an intercept) compared with 1000 in (5.7). And yet the estimated ATTs, moderating effects, and estimates on the heterogeneous trends ($fs_t\cdot\mathbf{x}_i$) are identical.

In Wooldridge (2023)—research that came after the first working paper version of the current paper—I used the equivalence between (5.3) and (5.7) as motivation to extend the pooled OLS approach on cohort dummies to nonlinear models, thereby avoiding the incidental parameters problem that usually arises in adding unit fixed effects in nonlinear models. Under a modification of the conditional PT assumption, Procedure 5.1 extends immediately to logit, fractional logit, and Poisson regression with an exponential mean—among other nonlinear models.

The identification analysis in Sect. 3 coupled with the equivalences in Sects. 4 and 5 lead to an important conclusion: the two-way fixed effects estimator is an attractive estimator provided it is applied to a suitably flexible model—namely (5.6). The work by De Chaisemartin and d'Haultfoeuille (2020), Callaway and Sant'Anna (2021), Goodman-Bacon (2021), and others use the phrase "two-way fixed effects" to refer to a particular, restrictive model, namely,

$$y_{it} = \tau \cdot w_{it} + \sum_{s=2}^{T} \gamma_s fs_t + c_i + u_{it}, t = 1, \ldots, T. \tag{5.8}$$

For simplicity, I have dropped the control variables. Equation (5.8) imposes a constant treatment effect regardless of when the first treatment occurred and also across calendar time period. As shown by these earlier authors, the resulting estimate, $\hat{\tau}_{TWFE}$, depends on 2×2 difference-in-differences estimates, several of which are based on "forbidden comparisons." Even when (5.6) excludes the controls, the extended equation is still considerably more flexible than (5.8):

$$y_{it} = \sum_{g=q}^{T} \sum_{s=g}^{T} \tau_{gs} \left(w_{it} \cdot dg_i \cdot fs_t \right) + \sum_{s=2}^{T} \gamma_s fs_t + c_i + u_{it}. \tag{5.9}$$

Under random sampling across i, no anticipation, and the unconditional PT assumption, estimation of (5.9) by fixed effects—TWFE with the inclusion of the fs_t—produces unbiased and consistent estimators of the ATTs τ_{gt}. We can note that $w_{it} = w_{it} \left(\sum_{g=q}^{T} \sum_{s=g}^{T} dg_i \cdot fs_t \right)$, which shows that (5.8) collapses the $(T - q + 1)(T - q + 2)/2$ treatment indicators $w_{it} \cdot dg_i \cdot fs_t$ into the single indicator w_{it}. Viewed from this perspective, it is not surprising that estimation of (5.8) need not produce an interesting weighted average of causal effects. With the availability of $\mathbf{x}_i$, just adding them to (5.8) does nothing because they are constant across t. Even if we interact $\mathbf{x}_i$ with w_{it}, or the time dummies fs_t, the resulting extension of (5.8) is not, generally, sufficiently flexible because it still would not allow separate effects τ_{gs}. We prefer (5.6) when covariates are available [and (5.9) when they are not].

To distinguish estimation of (5.8) [or even (5.9)] by TWFE from estimation of (5.6) by TWFE, we can refer to the latter as "extended TWFE." This label is not perfect, as it is not the estimator that is being "extended." We are simply applying the usual TWFE estimator to an extended equation. Generally, this is a good reminder that useful estimation methods can be applied to a wide range of models. That much empirical research has applied TWFE to (5.8) is not the fault of the estimation method; it is the fault of using a model that is too restrictive.

The estimating equation simplifies under common timing. Then, we need only introduce a single "cohort" dummy, d_i, with $d_i = 1$ if unit i is first treated in period q. Equation (5.6) becomes

$$y_{it} = \sum_{s=q}^{T} \tau_s \left(w_{it} \cdot d_i \cdot fs_t \right) + \sum_{s=q}^{T} \left(w_{it} \cdot d_i \cdot fs_t \right) \cdot \dot{\mathbf{x}}_{i1} \cdot \boldsymbol{\delta}_s \tag{5.10}$$

$$+ \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} \left(fs_t \cdot \mathbf{x}_i \right) \boldsymbol{\pi}_s + c_i + u_{it}, \ t = 1, \ldots, T,$$

where $\dot{\mathbf{x}}_{i1} = \mathbf{x}_i - \bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_1 = N_1^{-1} \sum_{i=1}^{N} d_i \cdot \mathbf{x}_i$ is the average of the covariates over the treated units. Naturally, there is a pooled OLS version that drops the unit FEs, c_i, and instead adds the single dummy variable d_i— the "ever treated" indicator— along with $\mathbf{x}_i$ and $d_i \cdot \mathbf{x}_i$. Without $\mathbf{x}_i$, the equivalence between POLS and ETWFE shows that it suffices to control for a single binary source of heterogeneity, d_i, rather than N different sources, ch_i, $h = 1, \ldots, N$.

5.3 Equivalence with random effects and efficiency considerations

It follows immediately from Corollary 3.2 that the previous estimators are also equivalent to the random effects estimator applied to an equation underlying the POLS estimator using cohort dummies. One can motivate the RE estimator starting from (5.6) and using a version of the Mundlak device:

$$c_i = \alpha + \sum_{g=q}^{T} \beta_g dg_i + \mathbf{x}_i \boldsymbol{\kappa} + \sum_{g=q}^{T} (dg_i \cdot \mathbf{x}_i)\, \boldsymbol{\xi}_g + a_i, \tag{5.11}$$

where, by construction, a_i has zero mean and is uncorrelated with all variables on the right-hand side. Plugging this equation into (5.6) gives

$$\begin{aligned} y_{it} = {} & \sum_{g=q}^{T}\sum_{s=g}^{T} \tau_{gs}\, (w_{it} \cdot dg_i \cdot fs_t) + \sum_{g=q}^{T}\sum_{s=g}^{T} \left(w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}\right) \boldsymbol{\delta}_{gs} \\ & + \sum_{s=2}^{T} \gamma_s fs_t + \sum_{s=2}^{T} (fs_t \cdot \mathbf{x}_i)\, \boldsymbol{\pi}_s + \alpha + \sum_{g=q}^{T} \beta_g dg_i \\ & + \mathbf{x}_i \boldsymbol{\kappa} + \sum_{g=q}^{T} (dg_i \cdot \mathbf{x}_i)\, \boldsymbol{\xi}_g + a_i + u_{it} \end{aligned} \tag{5.12}$$

In a traditional panel data analysis, both components of the error term, a_i and u_{it}, are uncorrelated with all explanatory variables in all time periods, justifying estimation of (5.12) by generalized least squares. The random effects estimator is a very specific GLS (or feasible GLS) estimator. But, again, the "truth" of (5.12) is irrelevant for the algebraic equivalences among the estimators.

Knowing that the imputation/POLS/TWFE estimator also can be obtained by RE estimation of (5.12) is convenient for discussing efficiency. Wooldridge (2010, Section 10.4) discusses the assumptions under which random effects is a true GLS estimator. Briefly, the second moment assumptions rule out heteroskedasticity in a_i (conditional on $\mathbf{d}_i$, $\mathbf{x}_i$) and heteroskedasticity and serial correlation in u_{it} (conditional on $\mathbf{d}_i$, $\mathbf{x}_i$, and a_i). None of those second moment assumptions is needed for consistency—which is obvious in the current setting because the RE estimator is the same as POLS and FE. In fact, the actual values of σ_a^2 and σ_u^2 we plug into the RE variance–covariance structure are irrelevant: we always obtain the POLS estimator. Therefore, we can do the thought experiment of plugging in the *unknown* population values, giving us a true

GLS estimator. It follows that we have found both the best linear unbiased estimator and the asymptotically efficient estimator (for fixed T, $N \to \infty$).

If there is heteroskedasticity in a_i or u_{it}, or serial correlation in u_{it}, then the POLS/TWFE/RE estimator is neither BLUE nor asymptotically efficient. Other estimators can be more efficient, especially if they exploit strong serial correlation in u_{it}.

5.4 All units eventually treated

The POLS/ETWFE approach extends immediately to the case where all units are treated in the final period T. To handle this situation, one initially defines new treatment effects,

$$te\,(g{:}T) = y_t\,(g) - y_t\,(T)\,, \quad g = q, \ldots, T-1; \quad t = g, \ldots, T, \tag{5.13}$$

which can be interpreted as the gain in period t from first being treated in the earlier period g rather than first being treated in the last period T. If the CNA and CPT assumptions are stated with $y_t\,(T)$ replacing $y_t\,(\infty)$, the identified parameters are

$$\tau_{(g:T),t} \equiv E\left[y_t\,(g) - y_t\,(T)\,|dg = 1\right], \; g = q, \ldots, T-1; \quad t = g, \ldots, T. \tag{5.14}$$

Assuming that it still makes sense to think of the potential outcome in the NT state, the no anticipation assumption implies

$$E\left[y_t\,(T)\,|dg = 1\right] = E\left[y_t\,(\infty)\,|dg = 1\right], t < T,$$

which means

$$\tau_{(g:T),t} = \tau_{gt}, g = q, \ldots, T-1; t = g, \ldots, T-1. \tag{5.15}$$

In other words, in every treated period before T, the NA assumption implies that the identified ATTs have the same interpretation as when there is an NT group. In the final period, we identify the ATTs $\tau_{(g:T),T}$, and there is no identified ATT for treatment cohort T because there is no control group for that cohort. Mechanically, all variables in regression (5.3) (or its TWFE version) involving dT_i get dropped, effectively forcing treatment cohort T to play the role of the never treated group.

5.5 Summary of estimator equivalences

In addition to the equivalences discussed so far, there is another equivalence worth pointing out. Borusyak et al. (2024)[BJS (2024)] propose an imputation estimator based on TWFE estimation in the imputation step. In the current framework, that means replacing the time-constant variables in (4.14) with the unit-specific dummies, leading to the regression y_{it} on $c1_i, \ldots, cN_i, f2_t, \ldots, fT_t$, and $f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i$ using the $w_{it} = 0$ observations. The out-of-sample residuals, $\widehat{te}_{it}$, are generally different from those in Procedure 4.1. Nevertheless, averaging these by cohort and time-period produces the same estimates of the ATTs, $\hat{\tau}_{gt}$. As discussed in BJS (2024), these unit-specific treatment effects can be averaged across other subsamples to obtain ATTs for different groups. But inference is complicated, and BJS (2024) only provide conservative standard errors. Inference is made easier by implementing the estimator as pooled OLS, random effects (equivalently TWFE); averaging over different subsamples is accomplished by including appropriate dummy variables in $\mathbf{x}_i$.

To summarize, we have the following algebraic equivalences:

$$\text{Cohort Imputation} = \text{POLS} = \text{TWFE} = \text{RE} = \text{BJS Imputation}, \tag{5.16}$$

where the imputation methods use only the control observations and TWFE sweeps away all time-constant variables. For general panel data applications, POLS, RE, and TWFE are all different. That they are the same in the staggered DiD setting when applied to an equation derived under standard no anticipation, parallel trends, and linear conditional expectations assumptions is compelling. These equivalences also hold in the absence of a never treated group.

Another valuable implication of the algebraic equivalence is that the POLS, TWFE, and RE provide moderating effects without additional work. Finally, the POLS estimator (and RE estimator) using cohort dummies allows us to see coefficients on the dg_i (and also on $\mathbf{x}_i$ and $dg_i \cdot \dot{\mathbf{x}}_{ig}$), which provides direct evidence on the nature of selection into the different treatment cohorts. Aesthetically, the POLS estimator from (5.3) produces the lengthiest output; by doing so, it opens up the black box on all of the equivalent estimators. If we want to know about selection, heterogeneous trends, and moderating effects, we should use the estimation method that provides information on these issues.

The equivalences in (5.16) do assume that we have a balanced panel and that the covariates are not time-varying. I discuss relaxing these restrictions in Sect. 10.

6 Event study estimators

The estimation method presented in Sect. 5, whether we call it imputation, pooled OLS, TWFE, or random effects, employs the conditional PT assumption in effectively using all pre-treatment periods as control observations for any treated cohort g. The procedure does not provide evidence on whether the (conditional) PT assumption holds in periods before the intervention.

Without controls and using the current notation, the so-called leads and lags estimator can be implemented as

$$\begin{aligned}
&y_{it} \text{ on } dq_i \cdot f1_t, \ldots, dq_i \cdot f(q-2)_t, dq_i \cdot fq_t, \ldots, dq_i \cdot fT_t\\
&\quad d(q+1)_i \cdot f1_t, \ldots, d(q+1)_i \cdot f(q-1)_t,\\
&\qquad d(q+1)_i \cdot f(q+1)_t, \ldots, d(q+1)_i \cdot fT_t,\\
&\qquad \ldots, dT_i \cdot f1_t, \ldots, dT_i \cdot f(T-2)_t, \ldots, dT_i \cdot fT_t,\\
&\quad f2_t, \ldots, fT_t, 1, dq_i, \ldots, dT_i,
\end{aligned} \tag{6.1}$$

where interactions $dg_i \cdot fs_t$ for $s < g$ are the pre-treatment (or "leads") indicators. Notice that (6.1) excludes the pre-treatment dummies in the period just before the intervention, namely, $dg_i \cdot f(g-1)_t$ for treatment cohort g. This choice forces $g-1$ to be the reference period for treated cohort g. As before, the variables in the last line of (6.1) are the control variables, allowing for secular changes in the never treated state and selection into treatment cohort. The equivalence result in Theorem 3.1 can be used to show that the estimates from (6.1) are the same as the TWFE estimates where, in the last line, $1, dq_i, \ldots, dT_i$ is replaced with the unit indicators $c1_i, \ldots, cN_i$. The TWFE version of the estimator is the one proposed by Sun and Abraham (2021). Now we can conclude that pooled OLS on cohort dummies and RE estimation using the same variables in (6.1) are numerically the same.

Without controls (and dropping w_{it}), regression (5.3) becomes

$$\begin{aligned} & y_{it} \text{ on } dq_i \cdot fq_t, \ldots, dq_i \cdot fT_t, \\ & \quad d(q+1)_i \cdot f(q+1)_t, \ldots, d(q+1)_i \cdot fT_t, \\ & \quad dT_i \cdot fT_t, f2_t, \ldots, fT_t, 1, dq_i, \ldots, dT_i \end{aligned} \tag{6.2}$$

and so all of the leads have been dropped from (6.2). The estimates of the ATTs, τ_{gt}, from (6.1) and (6.2) are not the same. So what is the benefit of adding the pre-treatment indicators? One is that, in addition to the estimated ATTs, which we will label $\breve{\tau}_{gs}$, (6.1) produces estimates $\breve{\theta}_{gs}$, for $s = 1, \ldots, g-2$, which are estimates of "pre-trends" in the never treated state. Combined, for each cohort g we have a set of estimates

$$\breve{\theta}_{gs}, s = 1, \ldots, g-2; \breve{\tau}_{gs}, s = g, \ldots, T. \tag{6.3}$$

This set of estimates can be displayed in an event study graph. Under the ideal second moment conditions described in Sect. 5.3, the $\breve{\tau}_{gs}$ are less efficient than the $\hat{\tau}_{gs}$ from (6.2). In effect, (6.1) adds redundant regressors if no anticipation and parallel trends hold, and that causes inefficiency if the ideal RE assumptions hold. However, including the pre-treatment indicators can actually improve efficiency in the presence of strong, positive serial correlation— more on this below.

With covariates, symmetry suggests including all possible interactions $dg_i \cdot fs_t$ with the controls, including those for pre-treatment periods:

$$\begin{aligned} & y_{it} \text{ on } dq_i \cdot f1_t, \ldots, dq_i \cdot f(q-2)_t, dq_i \cdot fq_t, \ldots, dq_i \cdot fT_t \\ & \quad d(q+1)_i \cdot f1_t, \ldots, d(q+1)_i \cdot f(q-1)_t, \\ & \quad d(q+1)_i \cdot f(q+1)_t, \ldots, d(q+1)_i \cdot fT_t, \ldots, \\ & \quad dT_i \cdot f1_t, \ldots, dT_i \cdot f(T-2)_t, \ldots, dT_i \cdot fT_t, \\ & \quad dq_i \cdot f1_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, dq_i \cdot f(q-2)_t \cdot \dot{\mathbf{x}}_{iq}, dq_i \cdot fq_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, dq_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iq}, \\ & \quad d(q+1)_i \cdot f1_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, d(q+1)_i \cdot f(q-1)_t \cdot \dot{\mathbf{x}}_{i,q+1}, \\ & \quad d(q+1)_i \cdot f(q+1)_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, d(q+1)_i \cdot fT_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, \\ & \quad dT_i \cdot f1_t \cdot \dot{\mathbf{x}}_{iT}, \ldots, dT_i \cdot f(T-2)_t \cdot \dot{\mathbf{x}}_{iT}, dT_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iT}, \\ & \quad f2_t, \ldots, fT_t, \ f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i, \end{aligned}$$

$$1, dq_i, \ldots, dT_i, \mathbf{x}_i, dq_i \cdot \dot{\mathbf{x}}_{iq}, \ldots, dT_i \cdot \dot{\mathbf{x}}_{iT}. \tag{6.4}$$

As before, the reference period for cohort g is $g - 1$, and so $dg_i \cdot f(g-1)_t$ and $dg_i \cdot f(g-1)_t \cdot \dot{\mathbf{x}}_{ig}$ do not appear in (6.4). As with the other regressions, the POLS estimator is equivalent to RE estimation, and also to TWFE estimation with the last row of control variables dropped. These follow almost immediately from Theorem 3.1 and Corollary 3.2.

Regression (6.4) is a natural extension of (5.3) in that it includes the "leads" of the treatment indicators and the interactions of these lead indicators with the controls. It can also be viewed as a fully flexible extension of the Sun and Abraham (2021) TWFE estimator.

The regression in (6.4) is fully saturated, and therefore, the estimates can be characterized as many 2×2 difference-in-differences regressions. In particular, for treatment cohort g, one uses a 2×2 DiD where $g - 1$ is the reference period and s as the "treatment" period. The control group is the never treated group and the treated group is cohort g. If $s \geq g$, the 2×2 regression produces $\breve{\tau}_{gs}$ (along with moderating effects, selection effects, and heterogeneous trend effects). By contrast, the "lags only" regression in (5.3) uses all pre-treatment periods when implicitly constructing a control—not just period $g - 1$.

Another useful representation connects (6.4) with the recent literature. The coefficients can be obtained using a sequence of cross-sectional regressions using "long" differences. Specifically, consider the regression where all variables have been differenced relative to time period $g - 1$:

$$y_{is} - y_{i,g-1} \text{ on } 1, dg_i, \dot{\mathbf{x}}_{ig}, dg_i \cdot \dot{\mathbf{x}}_{ig} \text{ using } dg_i = 1 \text{ or } d\infty_i = 1. \tag{6.5}$$

Then, if $s \geq g$, the coefficient on dg_i is $\breve{\tau}_{gs}$; if $s \leq g - 2$, (6.4) recovers $\breve{\theta}_{gs}$. Moreover, the coefficients on $\dot{\mathbf{x}}_{ig}$ in (6.5) are the same as the moderating effects on $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$ in (6.4). The regressions in (6.5) are identical to the linear regression-adjustment estimators proposed by CS (2021) for the panel data case when the never treated units are used as the controls. This equivalence shows that, while the long regression in (6.4) appears somewhat intimidating, it is equivalent to a currently popular method that obtains the estimates by breaking the long regression into a series of shorter, cross-sectional regressions based on differences. [CS (2021) propose several other estimators, including doubly robust estimators.] In using (6.4), we can look at the confounding effects, which are not available in (6.5) because, in effect, (6.5) differences away the additive heterogeneity that determines selection. As we will see in the next section, it is also easy to aggregate the estimates from (6.4) to obtain a weighted event study plot.

Harmon (2024) shows that the CS (2021) regression estimator is efficient when the idiosyncratic errors—u_{it} in (5.6) —follow a random walk (and are conditionally homoskedastic). The equivalence with the long regression in (6.4) shows that this is a case where including redundant regressors (under PT) can improve efficiency when the "ideal" assumptions fail.

Interestingly, the estimates $\breve{\tau}_{gs}$ and $\breve{\boldsymbol{\delta}}_{gs}$—the coefficients on $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$ for $s = g$, $g + 1$, …, T— also can be obtained by an imputation method by first including all pre-treatment indicators, interactions of those with the $\dot{\mathbf{x}}_{ig}$, and all of the controls in (6.4) (the last two lines). Any variable involving $dg_i \cdot fs_t$ with $s \in \{g, g+1, \ldots, T\}$ is dropped. The equivalence follows from the general result in the appendix on imputation estimation.

6.1 Testing the null of parallel trends

One of the main uses of the leads and lags regression is to test for the presence of pre-trends, which is indicated by nonzero population coefficients on $dg_i \cdot fs_t$ for $s \leq g-2$. The equivalence between imputation and the full regression (6.4) implies it does not matter whether we use the fully sample, as in (6.4), or the untreated observations ($w_{it} = 0$) in the first step of the imputation: the coefficients are the same. Therefore, provided we use the fully flexible regression in (6.4), there is no issue of "contamination bias" in testing for pre-trends. Of course, if we fail to find statistically significant evidence of pre-trends, we cannot know if parallel trends is violated as the treatment period begins for the different cohorts. See Roth (2022), Rambachan and Roth (2023), and Roth et al. (2023) for important discussions.

One can argue that, if the goal is purely to test for pre-trends, one might omit $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}$ for $s \leq g-2$. A case for including these terms is that it reproduces the CS (2021) regression-based estimates from (6.4) because it is the same as including $dg_i \cdot \dot{\mathbf{x}}_{ig}$ in (6.5).

Additionally, one might wonder whether the outcome of the pre-trends test depends on which pre-intervention period is used as the reference period. It turns out that any set of pre-treatment periods can be used. For example, if we use $t = 1$ as the reference period for each cohort g (rather than period $g-1$), the test of pre-trends is numerically the same. In regression (6.4) we would add $dg_i \cdot f(g-1)_t$ and $dg_i \cdot f(g-1)_t \cdot \dot{\mathbf{x}}_{ig}$ and drop $dg_i \cdot f1_t$ and $dg_i \cdot f1_t \cdot \dot{\mathbf{x}}_{ig}$. However, the estimates of the treatment effects do change. As a simple example, suppose there are no controls with $T = 3$ and treatment occurs at $q = 3$. The regression in (6.4) becomes

$$y_{it} \text{ on } 1, d_i, f2_t, f3_t, d_i \cdot f1_t, d_i \cdot f3_t, t = 1, 2, 3; i = 1, \ldots, N, \tag{6.6}$$

and the test statistic is the cluster-robust t statistic on $d_i \cdot f1_t$. The regression in (6.6) is saturated: we are estimating the six cell means determined by the pairs (d, t) for $d \in \{0, 1\}$, $t \in \{1, 2, 3\}$ and computing linear combinations. The typical approach in an empirical study would be to fail to reject PT if the cluster-robust t statistic on $d_i \cdot f1_t$ is not significant, at, say, the 5% level. The estimated treatment effect, the coefficient on $d_i \cdot f3_t$, is the two-period difference-in-differences estimator $(\bar{y}_{13} - \bar{y}_{12}) - (\bar{y}_{03} - \bar{y}_{02})$, where the first index indicates treated or control and the second indicates time period. If we replace $d_i \cdot f1_t$ with $d_i \cdot f2_t$, the coefficient on $d_i \cdot f2_t$ is simply the negative of that on $d_i \cdot f1_t$ in (6.6)—resulting in the same two-sided t test. But the coefficient on $d_i \cdot f3_t$ becomes $(\bar{y}_{13} - \bar{y}_{11}) - (\bar{y}_{03} - \bar{y}_{01})$, which uses the longer difference over time; in other words, the reference period is first period for the 2 × 2 DiD rather than the second period.

Under the assumptions in Sect. 4, all estimators—the lags only, the leads and lags, and any other variation of the event study approach are consistent. The lags only estimator has the benefit of using all pre-treatment periods in effectively determining a suitable control. Even when it is not the efficient estimator, it does eliminate the decision on which pre-treatment period to choose as the reference period.

6.2 Relaxing parallel trends before the intervention

Because the leads and lags regression (6.4) includes more regressors than the lags only regression (5.3), it is worthwhile asking whether it has more robustness to violation of parallel trends. I turns out that, along with linearity, the CPT assumption underlying regression (6.4) is

$$\begin{aligned} E\left[y_t(\infty)\,|\mathbf{d},\mathbf{x}\right] &= \alpha + \sum_{g=q}^{T}\beta_g dg + \mathbf{x}\kappa + \sum_{g=q}^{T}(dg\cdot\mathbf{x})\,\boldsymbol{\xi}_g \\ &\quad + \sum_{s=2}^{T}\gamma_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{x})\,\boldsymbol{\pi}_s \\ &\quad + \sum_{g=q}^{T}\sum_{s=1}^{g-2}\theta_{gs}(dg\cdot fs_t) + \sum_{g=q}^{T}\sum_{s=1}^{g-2}\left(dg\cdot fs_t\cdot\dot{\mathbf{x}}_g\right)\boldsymbol{\lambda}_{gs}, \end{aligned} \tag{6.7}$$

where, as before, $\dot{\mathbf{x}}_g = \mathbf{x} - E(\mathbf{x}|dg = 1)$ and $E(\mathbf{x}|dg = 1)$ is replaced with $\bar{\mathbf{x}}_g$ in estimation. The last line in (6.7) adds additional terms compared with Eq. (4.10). These extra terms allow violation of the parallel trends assumption in periods before the intervention occurs. It is this equation that motivates the imputation procedure that is then the same as the long OLS regression in (6.4) (and the RE and FE equivalents). In the first stage of imputation in Procedure 4.1, the terms $dg_i\cdot fs_t$ and $dg_i\cdot fs_t\cdot\mathbf{x}_{ig}$ are added to (4.14).

It is tempting to think that the presence of extra controls in (6.7) always makes the leads and lags estimator more resilient to violations of parallel trends. If the violation of PT is only before period $g-1$ for each cohort g, then the leads and lags estimator will be unbiased and consistent whereas the lags only will not be. However, it seems a bit too convenient to assume that the PT violation disappears just when we need it to. If the violation of PT carries into the treated periods, including the pre-treatment indicators can actually exacerbate the bias compared with not including them. Simulation results, where each treated cohort differs from the baseline trend for the NT group by a linear trend, bear this out. The robustness issue is also discussed in De Chaisemartin and d'Haultfoeuille (2023) and Roth et al. (2023).

7 Aggregating the estimated ATTs

Consider again the flexible regression in Eq. (5.3). Even with moderate T, if there are many treated cohorts, the number of ATTs, τ_{gt}, can be large. For example, with

$T = 10$ and the first intervention at $q = 6$, there are $5(5+1)/2 = 15$ different treatment effects. In addition to potentially producing more information than desired, some of the $\hat{\tau}_{gt}$ may be based on a small number of observations. Consequently, some estimates may be imprecise, and the reported clustered standard errors may not accurately reflect the uncertainty in the estimates. For these reasons, it often makes sense to aggregate the estimated ATTs. As a reminder, because the estimates from regression (5.3) or (6.4) have been shown to be consistent under the assumptions in Sect. 4, there are no "bad comparisons" underlying the $\hat{\tau}_{gt}$. Therefore, aggregating these estimates makes more sense than imposing a constant effect, or imposing effects that vary only by exposure time.

If we want to report a single number that represents an average effect across all cohorts and treatment periods, one possibility is to define a straight average of the coefficients:

$$\bar{\tau} \equiv \frac{1}{(T-q+1)(T-q+2)/2} \sum_{g=q}^{T} \sum_{t=g}^{T} \tau_{gt}, \tag{7.1}$$

which is easily estimable by averaging the $\hat{\tau}_{gt}$ to obtain $\widehat{\bar{\tau}}$. Typically, one might prefer a weighted average of the τ_{gt}, where the weights are the cohort shares in the population:

$$\bar{\tau}_{\omega} = \sum_{g=q}^{T} \sum_{t=g}^{T} \omega_g \tau_{gt}. \tag{7.2}$$

The estimate is

$$\widehat{\bar{\tau}}_{\omega} = \sum_{g=q}^{T} \sum_{t=g}^{T} \hat{\omega}_g \hat{\tau}_{gt}, \tag{7.3}$$

where the weights are the same within a treated cohort:

$$\hat{\omega}_g \equiv N_g / \left[(T-q+1) N_q + \cdots + 2N_{T-1} + N_T \right], t = g, \ldots, T; \; g = q, \ldots, T. \tag{7.4}$$

It turns out that popular statistical packages, such as Stata, provide a simple way to obtain (7.3) and obtain a valid standard error that accounts for the sampling error not only in the $\bar{\mathbf{x}}_g$ but also in the cohort shares, $\hat{\omega}_g$. After running the regression (5.3), compute the partial effect with respect to w_{it}, and then average over the treated observations $w_{it} = 1$. This average partial effect (APE) over the treated units automatically weights the $\hat{\tau}_{gt}$ by cohort shares. Often there is an option that applies formulas from generalized method of moments (GMM) estimation to account for all estimation uncertainty.

Also common is to aggregate the estimated effects by exposure time. Define the exposure time, e, as the number of periods exposed to the treatment. By definition, each cohort has an exposure of $e = 0$ because they are all subjected to the intervention for at least one period. (We define the exposure time starting with $e = 0$ because that is common in empirical work.) The first treated cohort, in period q, has exposure times $e \in \{0, 1, \ldots, T-q\}$ and the final treated cohort only has $e = 0$.

To obtain a weighted average of exposure effects,

$$\hat{\tau}_{\omega,e} = \sum_{g=q}^{T-e} \hat{\omega}_{ge} \hat{\tau}_{g,g+e} \tag{7.5}$$

where the weights are now

$$\hat{\omega}_{ge} = N_g / \left(N_q + \cdots + N_{T-e} \right); \tag{7.6}$$

these are positive and sum to one for each exposure time. The first two weighted estimates by exposure time, along with the weights, are

$$\begin{aligned}
\hat{\tau}_{\omega,0} &= \hat{\omega}_{q0} \cdot \hat{\tau}_{qq} + \hat{\omega}_{q+1,0} \cdot \hat{\tau}_{q+1,q+1} + \cdots + \hat{\omega}_{T0} \cdot \hat{\tau}_{TT} \\
\hat{\omega}_{g0} &= N_g / \left(N_q + \cdots + N_T \right), g = q, \ldots, T \\
\hat{\tau}_{\omega,1} &= \hat{\omega}_{q1} \cdot \hat{\tau}_{q,q+1} + \hat{\omega}_{q+1,1} \cdot \hat{\tau}_{q+1,q+2} + \cdots + \hat{\omega}_{T-1,0} \cdot \hat{\tau}_{T-1,T} \\
\hat{\omega}_{g1} &= N_g / \left(N_q + \cdots + N_{T-1} \right), g = q, \ldots, T-1.
\end{aligned}$$

As with the overall weighted average, these weighted exposure time effects are easily computed in statistical packages that compute average partial effects along with valid standard errors. One first defines exposure time dummies, say, es_{it} for $s = 0, \ldots, T - q$. After running regression (5.3), to obtain (7.5) compute the APE for w_{it} now averaged over $es_{it} = 1$ for the specified exposure time. Proper standard errors are computed provided one accounts for heteroskedasticity, serial correlation, and sampling error in the $\bar{\mathbf{x}}_g$ and $\hat{\omega}_{ge}$. [One might cluster at a level higher than i if the data were obtained via cluster sampling or, more likely, the intervention is assigned at a higher level of aggregation; see Abadie, Athey, Imbens and Wooldridge (2023).] The weighted estimates by exposure time are often plotted along with 95% confidence intervals.

Conveniently, the same computational trick can be extended to the leads and lags (event study) estimation. It is easiest is to define a "not treated" indicator $nw_{it} = 1 - w_{it}$. Then, run the regression in (6.4) as

$$\begin{aligned}
&y_{it} \text{ on } nw_{it} \cdot dq_i \cdot f1_t, \ldots, nw_{it} \cdot dq_i \cdot f(q-2)_t, w_{it} \\
&\quad \cdot dq_i \cdot fq_t, \ldots, w_{it} \cdot dq_i \cdot fT_t \\
&\quad nw_{it} \cdot d(q+1)_i \cdot f1_t, \ldots, nw_{it} \cdot d(q+1)_i \cdot f(q-1)_t, \\
&\quad w_{it} \cdot d(q+1)_i \cdot f(q+1)_t, \ldots, w_{it} \cdot d(q+1)_i \cdot fT_t, \ldots \\
&\quad nw_{it} \cdot dT_i \cdot f1_t, \ldots, nw_{it} \cdot dT_i \cdot f(T-2)_t, \ldots, w_{it} \cdot dT_i \cdot fT_t, \\
&\quad nw_{it} \cdot dq_i \cdot f1_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, nw_{it} \cdot dq_i \cdot f(q-2)_t \cdot \dot{\mathbf{x}}_{iq}, w_{it} \cdot dq_i \cdot fq_t \cdot \dot{\mathbf{x}}_{iq}, \ldots, \\
&\quad w_{it} \cdot dq_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iq}, \qquad\qquad (7.7) \\
&\quad nw_{it} \cdot d(q+1)_i \cdot f1_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, nw_{it} \cdot d(q+1)_i \cdot f(q-1)_t \cdot \dot{\mathbf{x}}_{i,q+1}, \\
&\quad w_{it} \cdot d(q+1)_i \cdot f(q+1)_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, w_{it} \cdot d(q+1)_i \cdot fT_t \cdot \dot{\mathbf{x}}_{i,q+1}, \ldots, \\
&\quad nw_{it} \cdot dT_i \cdot f1_t \cdot \dot{\mathbf{x}}_{iT}, \ldots, nw_{it} \cdot dT_i \cdot f(T-2)_t \cdot \dot{\mathbf{x}}_{iT}, w_{it} \cdot dT_i \cdot fT_t \cdot \dot{\mathbf{x}}_{iT},
\end{aligned}$$

$$f2_t, \ldots, fT_t,\ f2_t \cdot \mathbf{x}_i, \ldots, fT_t \cdot \mathbf{x}_i,$$
$$1, dq_i, \ldots, dT_i, \mathbf{x}_i, dq_i \cdot \dot{\mathbf{x}}_{iq}, \ldots, dT_i \cdot \dot{\mathbf{x}}_{iT}.$$

The weighted averages of the $\breve{\tau}_{gt}$ are computed exactly as in (7.5). Now, we can also obtain weighted averages of the $\breve{\theta}_{gt}$, the coefficients on the pre-treatment dummies, by time until treatment. Remember, by the choice of $g-1$ as the reference period for each cohort g , $\breve{\theta}_{g,g-1} \equiv 0$, and so the weighted ES estimates are zero when $e = -1$. For the cohort where first treatment is at g, the pre-treatment "exposure" times run from $-(g-1)$ through -2. The weighted effects are obtained by computing the partial effect with respect to nw_{it} and, again, averaging over the subsample $es_{it} = 1$. When we combine these effects with the estimates in the treatment periods, we obtain a single, weighted event study plot. I provide an example in Sect. 9.

Naturally, in the common timing case the same calculations produce the usual event study estimates, but there is no weighting because there is only one treated cohort. An application to the effects of increased police presence on car thefts is provided in Wooldridge (2023), and compared with an exponential mean model.

8 Heterogeneous cohort trends

In some applications, the parallel trends assumption may fail even after conditioning on a set of controls included in $\mathbf{x}_i$. With a sufficient number of pre-treatment periods, we can allow cohort-specific trends to differ from the unrestricted baseline trend in parametric ways.

A straightforward extension of Eq. (4.10) allows the difference in trends from baseline by cohort status in the NT state:

$$\begin{aligned} E\left[y_t(\infty)\,|\mathbf{d},\mathbf{x}\right] = \alpha + \sum_{g=q}^{T}\beta_g dg + \mathbf{x}\kappa + \sum_{g=q}^{T}(dg\cdot\mathbf{x})\,\boldsymbol{\xi}_g \\ + \sum_{s=2}^{T}\gamma_s fs_t + \sum_{s=2}^{T}(fs_t\cdot\mathbf{x})\,\boldsymbol{\pi}_s + \sum_{g=q}^{T}\eta_g\cdot(dg\cdot t)\,, t = 1,\ldots,T. \end{aligned} \tag{8.1}$$

As before, the presence of $\sum_{s=2}^{T}\gamma_s fs_t$ allows an unrestricted trend in the never treated state, but this part of the trend does not vary by treatment status. The terms in $\sum_{s=2}^{T}(fs_t\cdot\mathbf{x})\,\boldsymbol{\pi}_s$ allow for violation of parallel trends, but only as a function of observed controls. Remember, with enough data and variation in treatment status, $\mathbf{x}_i$ can include binary indicators for a higher level of aggregation—for example, state dummies if i is a county or school district dummies if i is a school. The new terms in (8.1), $\eta_g\cdot(dg_i\cdot t)$, allow for trends to differ from the baseline in a linear fashion, with a different slope for each treated cohort g. Unlike the final line in (6.7), Eq. (8.1) allows violation of CPT into the treatment periods.

Following (4.10) and (6.7), Eq. (8.1) suggests a natural imputation method. And, as before, the ATTs from imputation can be conveniently obtained by a single pooled regression. In the (long) regression (5.3), simply add the interactions $dg_i\cdot t$, $g = q$,

…, T. It is shown in the appendix that all of the coefficients and the ATT estimates, and the moderating effects, are the same as the imputation approach that only uses the $w_{it} = 0$ observations in the first stage. Consequently, one need not restrict estimation of the trend using only the control observations and then extrapolate into the future; the pooled OLS regression using all data does exactly that. Moreover, RE applied to the same underlying equation and FE applied to the equation that drops the time-constant controls are still equivalent to POLS. In order to allow separate linear trends for each treated cohort, we need at least two pre-treatment periods per cohort. Adding these linear trends creates collinearity with the treatment indicators $w_{it} \cdot dg_i \cdot fs_t = dg_i \cdot fs_t$. These indicators are zero for the first $g - 1$ periods and then switch to one for $s \geq g$. The linear trends $dg_i \cdot t$ are also increasing. With two or more pre-treatment periods, the $dg_i \cdot fs_t$ and $dg_i \cdot t$ are not perfectly correlated, but there is positive correlation. Collinearity can be expected to increase the standard errors, but that can be offset be reduced residual variance if the trends help to explain y_{it}. Also, including the trends could reduce serial correlation or heteroskedasticity. It is not clear ahead of time which effect will dominate. Of course, if the cohort-specific trends are not needed, and the "ideal" assumptions for efficiency hold, then adding the $dg_i \cdot t$ only induces unnecessary collinearity and is less efficient than dropping the trends—just as when adding the leads in the event study estimation.

The linear trend specification leads to appealing estimators in simple cases. When $T = 3$ and the intervention occurs only in $t = 3$, the heterogenous linear trend regression is

$$y_{it} \text{ on } 1, d_i, f2_t, f3_t, d_i \cdot t, d_i \cdot f3_t, t = 1, 2, 3; i = 1, \ldots, N. \tag{8.2}$$

Using OLS algebra, the coefficient on $d_i \cdot f3_t$ can be written as

$$\begin{aligned}\hat{\tau}_{3,ddd} &= N_1^{-1} \sum_{i=1}^{N} d_i \cdot \Delta^2 y_{i3} - N_0^{-1} \sum_{i=1}^{N} (1 - d_i) \cdot \Delta^2 y_{i3} \\ &= [(\bar{y}_{13} - \bar{y}_{12}) - (\bar{y}_{03} - \bar{y}_{02})] - [(\bar{y}_{12} - \bar{y}_{11}) - (\bar{y}_{02} - \bar{y}_{01})], \end{aligned} \tag{8.3}$$

where the first index is for treated (one) or control (zero) and the second index is for time period. Note that $\hat{\tau}_{3,ddd}$ is a difference-in-difference-in-differences estimator, where the double differencing in this case is across time. The first term in (8.3) is the usual DiD estimator using periods two and three. The second term is a DiD estimator using periods two and one, which measures the difference in trends before the intervention. Often $(\bar{y}_{12} - \bar{y}_{11}) - (\bar{y}_{02} - \bar{y}_{01})$ is used as a placebo check, but in (8.3) it is used to adjust the 2 × 2 DiD estimator for the presence of pre-trends.

If we drop $d_i \cdot t$ from (8.2) we obtain the lags only estimator as the coefficient on $d_i \cdot f3_t$. If we replace $d_i \cdot t$ with $d_i \cdot f1_t$, we obtain the typical leads and lags estimator—which, for the only ATT, is a 2 × 2 DiD estimator where $t = 2$ is used as the pre-treatment reference period. [In other words, the first term in (8.3).] If we replace $d_i \cdot t$ with $d_i \cdot f2_t$, the coefficient on $d_i \cdot f3_t$ is the 2 × 2 DiD using $t = 1$ as the reference period. The four estimates are all different in general, with the regression in (8.2) the only one allowing for violation of PT into the post-treatment period. Interesting, the cluster-robust t statistics on $d_i \cdot t$, $d_i \cdot f1_t$ and $d_i \cdot f2_t$ are all the same in absolute value, leading to the same test of the null hypothesis of parallel trends. However, the estimates of τ_3 can be very different.

Generally, we can make the regression even more flexible by adding the terms $dg_i \cdot t \cdot \dot{\mathbf{x}}_{ig}$, where centering the covariates makes the coefficients on $dg_i \cdot t$ more readily interpretable but does not change the estimated ATTs. With many pre-treatment periods, we can add higher order polynomials in t—such as t^2 and t^3—interacted with dg_i and $dg_i \cdot \dot{\mathbf{x}}_{ig}$. Naturally, adding more parametric trend functions requires more pre-treatment periods, and could easily reduce the precision of the estimates of τ_{gt} even further.

9 Effects of Walmart openings on county retail employment

In this section, I use data compiled by and used in Brown and Butts (2025) for a subset of counties in the USA over the years 1977 through 1999. These data are also used in Lee and Wooldridge (2023). The goal is to estimate the effects of locating a Walmart store—an early example of a "big box" store—on county-wide retail employment. In the dataset, none of the 1,288 counties had a Walmart prior to 1986. In 1986, a Walmart opened in 69 of the counties. By 1999, 893 of the counties had a Walmart. I only use information on whether the county has at least one Walmart, and so this is a staggered intervention without reversibility and a binary treatment. There are fourteen different treated cohorts. The outcome variable is the log of retail employment, and so the coefficients can be turned into (approximate) percentage effects by multiplying by 100.

Three time-constant controls are used in the fully flexible estimation, all dated in 1980: the share of the population employed in manufacturing, the share of the population above the poverty line, and the share of the population with a high school degree.

The estimates by exposure time are reported in Table 1. The standard errors are clustered at the county level, and the sampling variation in the controls and the weights used to obtain the exposure time effects have been accounted for using the margins command in Stata 18. Column (1) is the estimator developed in Sections 4 and 5, which does not estimate pre-treatment effects ("lags only"). Column (2) is the leads and lags estimator from Sect. 6, which is the fully flexible version of the Sun and Abraham (2021) estimator and also the regression-based Callaway and Sant'Anna (2021) estimator.

The estimates in the first two columns are similar, with those in column (2) being lower by, on average, about two percentage points. Many of the estimates are practically large. For example, eight years after a Walmart opens in a county, it is estimated to have retail employment that is between 11 and 13.2 percent higher than if no Walmart were opened, with t statistics above nine. The effects are even larger in later years. The weighted averages of the 14 estimates are 9.4% and 7.3% for columns (1) and (2), respectively. Again, both are very statistically significant.

Unfortunately, the estimates in columns (1) and (2) cannot be relied upon as good causal estimates because, as shown in Fig. 1, there is clear evidence of differences in

Table 1 Effects of Walmart Sitings on County Retail Employment

	(1) Lags Only	(2) Leads and Lags	(3) Heterogeneous Trends
$\tau_{\omega,0}$	0.0414 (0.0057)	0.0232 (0.0028)	0.0060 (0.0039)
$\tau_{\omega,1}$	0.0732 (0.0066)	0.0543 (0.0041)	0.0315 (0.0052)
$\tau_{\omega,2}$	0.0730 (0.0076)	0.0537 (0.0050)	0.0244 (0.0064)
$\tau_{\omega,3}$	0.0753 (0.0088)	0.0548 (0.0061)	0.0213 (0.0078)
$\tau_{\omega,4}$	0.0805 (0.0098)	0.0592 (0.0072)	0.0203 (0.0090)
$\tau_{\omega,5}$	0.0912 (0.0108)	0.0669 (0.0084)	0.0220 (0.0106)
$\tau_{\omega,6}$	0.1010 (0.0119)	0.0772 (0.0096)	0.0257 (0.0121)
$\tau_{\omega,7}$	0.1192 (0.0129)	0.0954 (0.0107)	0.0431 (0.0138)
$\tau_{\omega,8}$	0.1324 (0.0141)	0.1104 (0.0117)	0.0491 (0.0155)
$\tau_{\omega,9}$	0.1371 (0.0160)	0.1196 (0.0134)	0.0468 (0.0180)
$\tau_{\omega,10}$	0.1582 (0.0192)	0.1366 (0.0161)	0.0351 (0.0223)
$\tau_{\omega,11}$	0.1655 (0.0235)	0.1452 (0.0195)	0.0082 (0.0276)
$\tau_{\omega,12}$	0.1668 (0.0301)	0.1532 (0.0244)	0.0035 (0.0349)
$\tau_{\omega,13}$	0.2062 (0.0432)	0.1910 (0.0345)	0.0282 (0.0515)
τ_{ω}	0.0935 (0.0098)	0.0728 (0.0073)	0.0260 (0.0096)

trends for the treated cohorts and the never treated cohort before a Walmart is sited. This figure is derived from the estimation underlying column (2) of Table 1. Many of the pre-treatment estimates are quite large, more than 10% going back 15 or more years before the opening. Remember, these estimates use the period just before the intervention (Walmart siting) as the reference period. The graph is consistent with the idea that retail employment was trending higher in counties that eventually would get a Walmart—well before the Walmart actually opened. Although the estimated pre-treatment coefficients are aggregated across treatment cohort, the figure is suggestive that allowing trend deviations by cohort to be linear linear might fit the pre-treatment periods pretty well. Column (3) in Table 1 contains the estimates that allow each treatment cohort to have a trend that deviates linearly from the baseline trend, and also includes the full interactions with the heterogeneous trends and the three control variables. The picture is now much different, with the estimated ATTs being much more modest. Except for the immediate effect, the effects in the first six years range between two and three percent, much lower than in columns (1) and (2). The effect peaks at just under 5% nine years out, but then effectively becomes zero. The average effect across all horizons is about 2.6% (t statistic ≈ 2.71).

In terms of precision, the leads and lags estimator has uniformly smaller standard errors, probably because there is substantial positive serial correlation in the underlying idiosyncratic errors. Interestingly, for the short exposure times, the standard errors for the heterogeneous trends estimates are actually below those for the lags only, perhaps reflecting the fact that the cohort-specific trends explain nontrivial variation in log retail sales.

When including the cohort-specific trends, it is unclear what is the proper event study plot, because choosing a reference period is not obvious. Nevertheless, it is

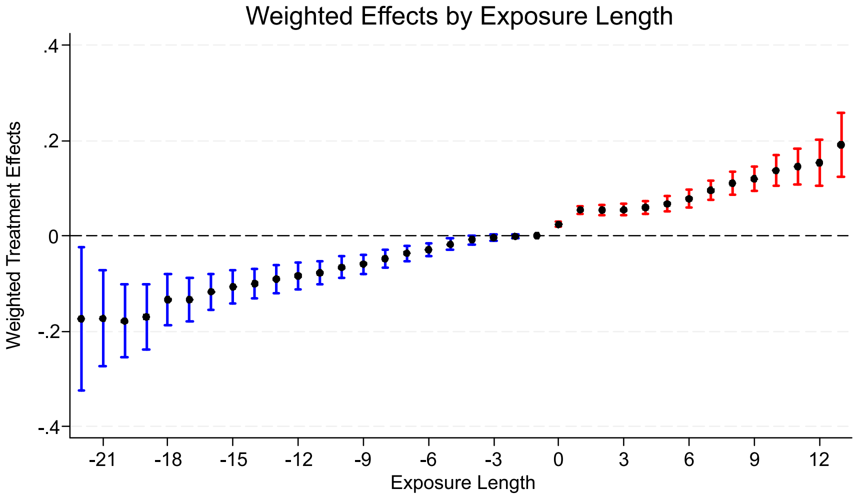

Fig. 1 Event Study (Leads and Lags) Estimates on Log of Retail Employment

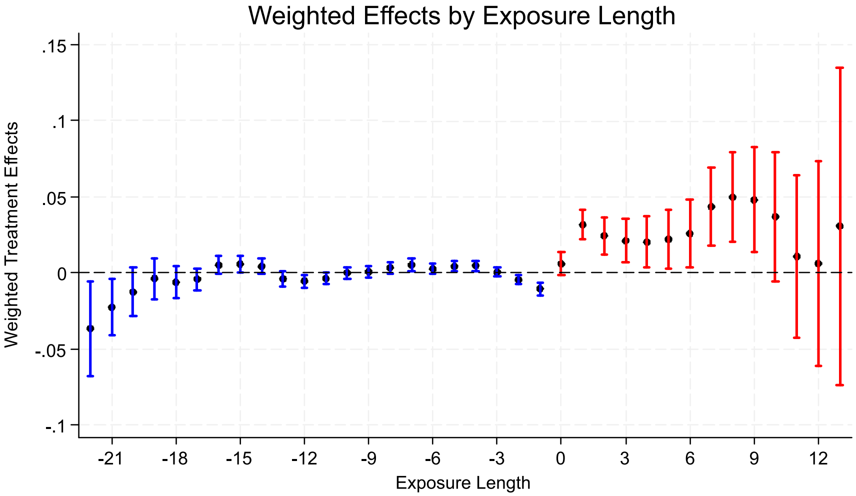

Fig. 2 Pre-Treatment and Treatment Effects with Cohort-Specific Linear Trends

straightforward to use the characterization as an imputation estimator to determine if allowing the cohort-specific trends alleviates the violation of (conditional) PT. Figure 2 plots the average residuals across each exposure time—both pre-treatment and post-treatment. For the pre-treatment period, the estimates come from the first-step imputation regression by averaging the residuals by time until exposure. The post-treatment estimates come from the second imputation step—or, equivalently, from the long regression. The standard errors used to obtain the 95% confidence are obtained using 1,000 bootstrap replications (rather than working through the more complicated analytical standard errors).

From Fig. 2, the cohort-specific linear trends do a pretty good job of eliminating the pre-trends, although it is not perfect. Going back 20 periods before the intervention, the deviations from zero are small, although some are statistically different from zero. At lags 21 and 22, the estimated pre-treatment effects are larger in magnitude. A more complete analysis could drop some of the early years in the dataset—for example, start in 1980—to determine the robustness of the findings. The important consideration is how were Walmart executives choosing counties to open a Walmart store. It seems reasonable that both level and trend of historical retail employment could come into play—but over what period? Here, my purpose is to show that heterogeneous trends can be used within the general framework when PT is clearly violated. The estimated effects are much smaller than either the lags only or leads and lags estimates.

10 Additional issues and extensions

In this section, we discuss two issues that arise in the context of both common and staggered interventions.

10.1 Time-varying covariates

It is common in the recent DiD literature to assume that controls do not change over time—as I have done in previous sections. The assumption that controls are dated prior to the first intervention period is typically imposed to avoid, or at least minimize, the possibility of "bad controls"– covariates whose values change due to the intervention. Conceptually, it makes sense that if the controls are included to partially account for selection into treatment, then they should be dated before the treatment period. Moreover, when thinking of modeling heterogeneous trends, it makes sense to think of those trends depending on pre-treatment covariates. Nevertheless, sometimes one wishes to include time-varying controls that predict the outcome and can be assumed to be unaffected by the intervention. For example, seasonal dummy variables and (perhaps) weather variables. Caetano and Callaway (2024) provide a recent analysis of time-varying controls in the context of staggered designs.

It is easy to see that the imputation approach in Procedure 4.1 does not go through if $E\left[\mathbf{x}\left(g\right)|d\infty=1\right] \neq E\left[\mathbf{x}\left(\infty\right)|d\infty=1\right]$. Naturally, the same is true with time-varying controls, and it is easiest to extend Assumption NBC to

$$\mathbf{x}_t\left(g\right) = \mathbf{x}_t\left(\infty\right), \quad g = q, \ldots, T; \quad t = 1, \ldots, T. \tag{10.1}$$

Then, we can let $\{\mathbf{x}_t : t = 1, \ldots, T\}$ denote the observed history of the covariates with the understanding that these do not vary by treatment status.

With time-varying covariates, the equivalence between imputation and pooled OLS continues to hold. For imputation, simply replace $\mathbf{x}_i$ with $\mathbf{x}_{it}$ in Procedure 4.1. For pooled OLS in Procedure 5.1, to keep a similar notation it is easiest to define

$$\dot{\mathbf{x}}_{itgs} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_{gs}, \tag{10.2}$$

where $\bar{\mathbf{x}}_{gs}$ is the average across treated cohort g in period s:

$$\bar{\mathbf{x}}_{gs} \equiv N_{gs}^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} dg_i \cdot fs_t \cdot \mathbf{x}_{it},\, N_{gs} \equiv \sum_{i=1}^{N} \sum_{t=1}^{T} dg_i \cdot fs_t. \tag{10.3}$$

Notice that in (5.3), the interactions for the moderating terms become $w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}$, and $dg_i \cdot fs_t = 0$ unless $s = t$ and so $\dot{\mathbf{x}}_{itgs}$ does not get used when $s \neq t$. In practice, it is easiest to define $\dot{\mathbf{x}}_{itgs}$ even when $s \neq t$ and then interact these with the appropriate treatment dummy $dg_i \cdot fs_t$. As in the case with time-constant controls, the POLS procedure using all of the data is more convenient for obtaining standard errors and aggregating effects. One can ensure that the ATTs have been properly obtained by using built-in command that compute average partial effects, and such packages usually have an option for accounting for the sampling variability in the $\bar{\mathbf{x}}_{gs}$.

With time-varying covariates, POLS on cohort dummies and POLS with unit dummies (TWFE) are no longer the same. When applying TWFE, the cohort dummies, dq_i, ..., dT_i, still drop out along with any time-constant controls and interactions between the dg_i and time-constant controls. But interactions $dg_i \cdot x_{itj}$ for time-varying covariates remain. Typically, we might think that TWFE is more robust because it controls for more heterogeneity. Nevertheless, we know that, with time-constant controls, there is no need to control for unit-specific effects under conditional parallel trends. Under the assumptions in Section 4, modified in the ways described above to handle time-varying variables, the POLS method is consistent. It all depends on whether parallel trends holds conditional on $\mathbf{x}_{it}$. Technically, with fixed-T asymptotics, consistency of the TWFE estimator requires that the sequence $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is strictly exogenous with respect to the implicit idiosyncratic shocks, u_{it}. As a practical matter, it seems cases where $\mathbf{x}_{it}$ is contemporaneously exogenous, does not react to treatment assignment, but is not strictly exogenous would be fairly rare.

With time-varying covariates comes additional control variables. For example, let $\mathbf{x}_{it} = (\mathbf{x}_{i1}, \mathbf{x}_{it2})$ where only $\mathbf{x}_{it2}$ has some time variation. Then, we can also control for the time average, $\bar{\mathbf{x}}_{i2} = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it2}$— which includes not only past but also future values of $\mathbf{x}_{it2}$ (for $t < T$). Such an extension is easy to accommodate using POLS (or TWFE) by using Procedure 5.1 with the expanded set of covariates $(\mathbf{x}_{i1}, \mathbf{x}_{it2}, \bar{\mathbf{x}}_{i2})$. Naturally, imputation and POLS continue to be the same, but it is no longer the same as RE or FE.

The same comments hold for the event study (leads and lags) estimator. When interacted with the pre-treatment indicators $dg_i \cdot fs_t$ for $s < g - 1$, the covariates are centered around the (g, s) averages. One can aggregate the estimated treatment effects, and pre-treatment effects, to construct a single event study graph—just as with time-constant controls. Incidentally, in other event study methods, particularly Callaway and Sant'Anna (2021), it is not clear how time-varying controls dated in the same period as the outcome should be handled. CS (2021) recommend using only the covariates dated in the earlier time period, although this makes less sense in estimating the pre-treatment effects.

10.2 Unbalanced panels

When a panel dataset is unbalanced, the equivalences derived earlier no longer hold. The POLS regression are still consistent if the missingness is related to $(dq_i, \ldots, dT_i, \mathbf{x}_i)$ (or replace $\mathbf{x}_i$ with $\mathbf{x}_{it}$). TWFE has an advantage with missing data because the missingness can be related to the unobserved effect, c_i, implicitly appearing in the equation.

It is possible to obtain an extended equation that can be estimated by POLS (or RE) using the results in Wooldridge (2019). For example, in the case of common treatment timing, with $w_{it} = d_i \cdot p_t$, the underlying equation that allows for differential effects across time is

$$y_{it} = \tau_q \left(w_{it} \cdot fq_t\right) + \cdots + \tau_T \left(w_{it} \cdot fT_t\right) + \theta_2 f2_t + \cdots + \theta_T fT_t + c_i + u_{it} \quad (10.4)$$

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

$$d_i \cdot \overline{fq}_i, \ldots, d_i \cdot \overline{fT}_i\ , \overline{f2}_i, \ldots, \overline{fT}_i \quad (10.5)$$

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 $T_i = 0$ has no usable data, and $T_i = 1$ units do not contribute to the estimation.) The POLS regression becomes

$$y_{it} \text{ on } 1,\ (d_i \cdot fq_t), \ldots, (d_i \cdot fT_t),\ f2_t, \ldots, fT_t, d_i \cdot \overline{fq}_i, \ldots, d_i \cdot \overline{fT}_i\ , \overline{f2}_i, \ldots, \overline{fT}_i \quad (10.6)$$

In effect, including the $\overline{fs}_i$ along with $d_i \cdot \overline{fs}_i$ accounts for selection issues related to additive unobserved heterogeneity. Of course, it is easier to use TWFE on the unbalanced panel once the interaction terms $w_{it} \cdot fs_t = w_{it} \cdot d_i \cdot fs_t$ have been created.

The POLS regression equivalent to FE is more complicated with covariates (even if time-constant) and staggered interventions. It is easiest to estimate the flexible Eq. (5.6) by TWFE using the complete cases. For the leads and lags estimator, add the pre-treatment terms

$$nw_{it} \cdot dg_i \cdot fs_t, w_{it} \cdot dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{ig}, 1 \le s \le g - 2.$$

Again, TWFE allows selection to depend on $(dq_i, \ldots, dT_i, \mathbf{x}_i, c_i)$. As with other estimation methods, selection is not allowed to depend systematically on u_{it}. With time-varying controls, one would demean the covariates using only the complete cases by (g, s) pair. Or, use built-in features of software packages that compute average marginal (partial) effects.

The regression-based methods offered here have notable advantages over the approaches due to CS (2021) and De Chaisemartin and d'Haultfoeuille (2024), which are based on differencing. The latter estimators require complete data in both time periods in order for a period to contribute to the estimation, whereas the regressions in levels use all available complete cases.

11 Concluding remarks

The equivalence between the TWFE estimator and the TWM regression increases our understanding of commonly used estimators. With staggered adoption, the identification results and equivalence between imputation and pooled OLS/extended TWFE highlight that, provided one allows treatment effects to be suitably heterogeneous, there is nothing inherently wrong with using TWFE—a conclusion reached by Sun and Abraham (2021) for the leads and lags estimator without controls.

The point here is not to conclude that other recent approaches— such as de De Chaisemartin and d'Haultfoeuille (2024), Callaway and Sant'Anna (2021), Borusyak et al. (2024), and others—are not valuable and cannot improve over flexible TWFE methods. However, we need not abandon simple approaches that have intuitive appeal, especially when they can be extended to allow nonstandard situations. Other than linearity of the conditional means in the covariates in the never treated state, 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. It can be applied to obtain "lags only" estimators of the ATTs, using all pre-treatment periods as part of the control, or the "leads and lags" event study estimators.

Another nice feature of the flexible POLS/ETWFE approach is that it is easily extended to allow for heterogeneous trends, which can help when one has evidence of violation of parallel trends or wants to determine the robustness of the estimates. The application to the effects of Walmart sitings on county retail employment show that accounting for trends that differ by cohort status can result in very different estimates.

If one has time-varying controls that are suitably exogenous, then these are easily incorporated. Imputation and pooled OLS using cohort and time dummies, and the same interactions as in the case of time-constant controls, continue to be identical. Extended TWFE and random effects are no longer identical to POLS unless one expands the Mundlak equation. But all procedures are consistent under no anticipation and conditional parallel trends provided the controls are strictly exogenous.

In the case of unbalanced panels, the different estimation approaches are no longer numerically identical, and ETWFE has the advantage of allowing selection to be correlated with additive, unobserved heterogeneity. Sample selection that is correlated with time-varying unobservables requires more advanced estimation methods.

Pooled OLS using cohort indicators in a flexible way extends to nonlinear models, and I have developed those in Wooldridge (2023) for binary, fractional, count, and corner solution outcomes.

Finally, the methods the imputation and pooled OLS methods can be extended to the case of repeated cross sections, provided we impose essentially the same assumptions in Sect. 4 along with a stable population over time. See Deb et al. (2024) for details.

Appendix A Mathematical Appendix

Proof of Theorem 3.1: It is easiest to first prove the result when $\mathbf{z}_i$ and $\mathbf{m}_t$ are not present, and then modify the 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

$$\mathbf{x}_{it} \text{ on } 1, \bar{\mathbf{x}}_{i\cdot}, \bar{\mathbf{x}}_{\cdot t},\ t = 1, \ldots, T;\ i = 1, \ldots, N. \tag{A1}$$

By a simple application of F-W, the residuals from (a.1) 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

$$\mathbf{x}_{it} - \bar{\mathbf{x}} \text{ on } \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}. \tag{A2}$$

It is easily seen that the two sets of regressors in (a.2) are orthogonal in sample:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})' (\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}) = \left[\sum_{i=1}^{N} (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})\right]' \sum_{t=1}^{T} (\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}) = \mathbf{0} \tag{A3}$$

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 (A3), we can focus on each term separately. We now show that the matrix of OLS coefficients on $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$, say $\hat{\mathbf{\Pi}}$, is $\mathbf{I}_K$:

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

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 (A3) is also $\mathbf{I}_K$. Therefore, the residuals from the regression (A3) are

$$\begin{aligned}\hat{\mathbf{r}}_{it} &\equiv (\mathbf{x}_{it} - \bar{\mathbf{x}}) - (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}})\,\mathbf{I}_K - (\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}})\,\mathbf{I}_K \\ &= \mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot t} + \bar{\mathbf{x}} = \ddot{\mathbf{x}}_{it}. \end{aligned} \tag{A4}$$

Now add the variables $\mathbf{z}_i$ and $\mathbf{m}_t$. It suffices to show that the $\ddot{\mathbf{x}}_{it}$ are still the residuals from

$$\mathbf{x}_{it} - \bar{\mathbf{x}} \text{ on } \bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}, \bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}, \mathbf{z}_i - \bar{\mathbf{z}}, \mathbf{m}_t - \bar{\mathbf{m}} \tag{A5}$$

where $\bar{\mathbf{z}} = N^{-1}\sum_{i=1}^{N}\mathbf{z}_i$ and $\bar{\mathbf{m}} = T^{-1}\sum_{t=1}^{T}\mathbf{m}_t$. We proceed by applying Frisch–Waugh multiple times. 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 without $(\mathbf{z}_i, \mathbf{m}_t)$ we know that the residuals are $\ddot{\mathbf{x}}_{it}$. Next, 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 $\mathbf{m}_t - \bar{\mathbf{m}}$. We already showed $\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_{\cdot t} - \bar{\mathbf{x}}$ are orthogonal in sample, so 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 as above, 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 (A5) are the same as those from

$$\ddot{\mathbf{x}}_{it} \text{ on } \ddot{\mathbf{e}}_i, \ddot{\mathbf{a}}_t, t = 1, \ldots, T; \quad i = 1, \ldots, N \tag{A6}$$

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

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

In fact, this shows that the matrix coefficients on $\mathbf{z}_i - \bar{\mathbf{z}}$ and $\mathbf{m}_t - \bar{\mathbf{m}}$ in (A5) are identically zero, and so the residuals from (A5) are the same as in (A2); namely, $\ddot{\mathbf{x}}_{it}$. □

The following general proposition can be used to establish the equivalence between the imputation estimators and their pooled OLS counterparts presented throughout the text.

Proposition A.1: Consider a panel dataset $\{(y_{it}, \mathbf{g}_{it}, \mathbf{h}_{it}, w_{it}) : t = 1, \ldots, T; i = 1, \ldots, N\}$ where $\mathbf{g}_{it}$ is $1 \times K$, $\mathbf{h}_{it}$ is $1 \times L$, and w_{it} is a binary indicator. Further, $w_{it}\mathbf{h}_{it} = \mathbf{h}_{it}$ for all i and t [so that $(1 - w_{it})\,\mathbf{h}_{it} = \mathbf{0}$]. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ be the POLS estimates from the regression

$$y_{it} \text{ on } \mathbf{g}_{it}, \mathbf{h}_{it}, t = 1, \ldots, T; i = 1, \ldots, N. \tag{A7}$$

Let $\tilde{\boldsymbol{\beta}}$ be the estimates from the short POLS regression using the $w_{it} = 0$ sample:

$$y_{it} \text{ on } \mathbf{g}_{it} \text{ if } w_{it} = 0 \tag{A8}$$

Define the residuals, for all (i, t), as

$$\tilde{v}_{it} = y_{it} - \mathbf{g}_{it}\tilde{\boldsymbol{\beta}} \tag{A9}$$

Let $\tilde{\boldsymbol{\gamma}}$ be the $L \times 1$ vector of coefficients from the regression

$$\tilde{v}_{it} \text{ on } \mathbf{h}_{it} \text{ if } w_{it} = 1 \tag{A10}$$

Suppose the OLS rank condition holds such that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are unique and, for all (i, t),

$$w_{it}\mathbf{g}_{it} = \mathbf{h}_{it}\mathbf{A} \tag{A11}$$

for an $L \times K$ matrix $\mathbf{A}$. Then

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \text{ and } \tilde{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}. \square \tag{A12}$$

Proof: The POLS problem defining $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ is

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{t=1}^{T} (y_{it} - \mathbf{g}_{it}\boldsymbol{\beta} - \mathbf{h}_{it}\boldsymbol{\gamma})^2$$

Because $(1 - w_{it}) \cdot \mathbf{h}_{it} = \mathbf{0}$, this is the same as

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{t=1}^{T} (1 - w_{it})\,(y_{it} - \mathbf{g}_{it}\boldsymbol{\beta})^2 + \sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\,(y_{it} - \mathbf{g}_{it}\beta - \mathbf{h}_{it}\boldsymbol{\gamma})^2$$

The first-order conditions (FOCs) can be written as

$$\sum_{i=1}^{N}\sum_{t=1}^{T} (1 - w_{it})\,\mathbf{g}_{it}'\left(y_{it} - \mathbf{g}_{it}\hat{\boldsymbol{\beta}}\right) + \sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{g}_{it}'\left(y_{it} - \mathbf{g}_{it}\hat{\boldsymbol{\beta}} - \mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right) = \mathbf{0}$$
$$\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{h}_{it}'\left(y_{it} - \mathbf{g}_{it}\hat{\boldsymbol{\beta}} - \mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right) = \mathbf{0}$$

Now substitute $w_{it}\mathbf{g}_{it} = w_{it}\mathbf{h}_{it}\mathbf{A}$ and use simple algebra to write the FOCs as

$$\sum_{i=1}^{N}\sum_{t=1}^{T} (1 - w_{it})\,\mathbf{g}_{it}'\left(y_{it} - \mathbf{g}_{it}\hat{\boldsymbol{\beta}}\right) + \mathbf{A}'\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{h}_{it}'\left(y_{it} - \mathbf{g}_{it}\hat{\boldsymbol{\beta}} - \mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right) = \mathbf{0} \tag{A13}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{h}_{it}'\left(y_{it}-\mathbf{g}_{it}\hat{\boldsymbol{\beta}}-\mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right)=\mathbf{0} \tag{A14}$$

The second term in (A13), being a linear combination of the left hand side of (A14), is redundant. In particular, $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ solves the two sets of equations

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-w_{it})\,\mathbf{g}_{it}'\left(y_{it}-\mathbf{g}_{it}\hat{\boldsymbol{\beta}}\right)=\mathbf{0} \tag{A15}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{h}_{it}'\left(y_{it}-\mathbf{g}_{it}\hat{\boldsymbol{\beta}}-\mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right)=\mathbf{0} \tag{A16}$$

But (A15) just says that $\hat{\boldsymbol{\beta}}$ is the OLS estimator of y_{it} on $\mathbf{g}_{it}$ using $w_{it}=0$—that is, $\hat{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}$. Then, (A16) can be written as

$$\sum_{i=1}^{N}\sum_{t=1}^{T} w_{it}\mathbf{h}_{it}'\left(\tilde{v}_{it}-\mathbf{h}_{it}\hat{\boldsymbol{\gamma}}\right)=\mathbf{0},$$

which, from (A10), says that $\hat{\boldsymbol{\gamma}}=\tilde{\boldsymbol{\gamma}}$. □

We can apply this result to prove Proposition 5.1 with $\hat{\boldsymbol{\gamma}}=\hat{\boldsymbol{\tau}}$ (the vector of ATTs)—and it is not much more difficult to tackle the case with time-varying controls. Take

$$\begin{aligned}\mathbf{g}_{it} &= \big(1, dq_i, \ldots, dT_i, f2_t, \ldots, fT_t, \mathbf{x}_{it},\\ &\qquad dq_i\cdot\mathbf{x}_{it}, \ldots, dT_i\cdot\mathbf{x}_{it}, f2_t\cdot\mathbf{x}_{it}, \ldots, fT_t\cdot\mathbf{x}_{it}\big) \end{aligned}\tag{A17}$$

$$\begin{aligned}\mathbf{h}_{it} &= \big(dq_i\cdot fq_t, \ldots, dq_i\cdot fT_t, \ldots, dT_i\cdot fT_t,\\ &\qquad dq_i\cdot fq_t\cdot\dot{\mathbf{x}}_{itqq}, \ldots, dq_i\cdot fT_t\cdot\dot{\mathbf{x}}_{itqT}, \ldots, dT_i\cdot fT_t\cdot\dot{\mathbf{x}}_{itTT}\big)\end{aligned}\tag{A18}$$

The condition $w_{it}\mathbf{h}_{it}=\mathbf{h}_{it}$ clearly holds because $w_{it}=\sum_{g=q}^{T}\sum_{s=g}^{T} dg_i\,fs_t$, the sum of mutually exclusive treatment dummies. Therefore, $w_{it}dg_i\,fs_t=dg_i\,fs_t$. Also, we can see that each element of $w_{it}\mathbf{g}_{it}$ is a fixed linear combination of elements in $\mathbf{h}_{it}$. Specifically, w_{it} is the sum of the treatment indicators in $\mathbf{h}_{it}$, $w_{it}dg_i=dg_i\,fg_t+\cdots+dg_i\,fT_t$, and $w_{it}fs_t=dq_i\,fs_t+\cdots+ds_i\,fs_t$. The terms involving $\mathbf{x}_{it}$ are handled similarly. (Note that removing the within-cohort averages does not change this fact because $dg_i\,fs_t\dot{\mathbf{x}}_{itgs}=dg_i\,fs_t\mathbf{x}_{it}-dg_i\,fs_t\bar{\mathbf{x}}_{gs}$ and $dg_i\,fs_t\bar{\mathbf{x}}_{gs}$ just consists of multiples of $dg_i\,fs_t$, which are already in $\mathbf{h}_{it}$.)

Given the definition of $\mathbf{g}_{it}$, it is immediate that $\tilde{v}_{it}$ in Proposition A.1 is $\widetilde{te}_{it}$ given in (4.17). All that remains is to characterize the coefficients in the regression

$$\widetilde{te}_{it} \text{ on } \mathbf{h}_{it} \text{ if } w_{it}=1 \tag{A19}$$

for the choice of $\mathbf{h}_{it}$ in (A18). Because every term in $\mathbf{h}_{it}$ is a treatment indicator or a treatment indicator multiplied by covariates, the $w_{it}=1$ restriction is redundant.

(We cannot have $dg_i \cdot fs_t = 1$ and $w_{it} = 0$.) Moreover, the treatment indicators are mutually exclusive, so we can separate the regression of $\widetilde{te}_{it}$ on $\mathbf{h}_{it}$ into regressions on several orthogonal pairs:

$$\widetilde{te}_{it} \text{ on } dg_i \cdot fs_t, dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}. \tag{A20}$$

Because the covariates are centered about $\bar{\mathbf{x}}_{gs}$, $\sum_{i=1}^{N} \sum_{t=1}^{T} (dg_i \cdot fs) \cdot \big(dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}\big) = \sum_{i=1}^{N} \sum_{t=1}^{T} dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs} = \mathbf{0}$—that is, $dg_i \cdot fs_t$ and $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}$ are orthogonal in sample. It follows that the coefficient on $dg_i \cdot fs_t$ in (A20) is the same as dropping $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}$. Because $dg_i \cdot fs_t$ is a dummy variable, the coefficient on it is simply the average of $\widetilde{te}_{it}$ over the $dg_i \cdot fs_t = 1$ subsample—precisely as in Eq. (4.18), extended to allow time-varying controls. Moreover, the coefficients from regression (5.3) on the $dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}$ are the same as those from (A20), which are clearly measures of moderating effects.

The previous argument is easily extended to the leads and lags estimator. There is no change in the definition of $\mathbf{h}_{it}$, but now $\mathbf{g}_{it}$ is expanded to include the terms

$$dg_i \cdot fs_t, dg_i \cdot fs_t \cdot \dot{\mathbf{x}}_{itgs}, s = 1, \ldots, g-2;\; g = q, \ldots, T. \tag{A21}$$

Because each of these involves pre-treatment indicators, multiplication by w_{it} leaves zero. Therefore, the requirement in (A11) is trivially satisfied for these new elements of $\mathbf{g}_{it}$ because zero is a linear combination of anything. The second-step regression for imputation is still (A20). This shows that, even with time-varying covariates, imputation and POLS in the leads and lags equation are identical for $\hat{\tau}_{gt}$ and the moderating effects.

The argument is also easily modified when the cohort-specific trends, $dg_i \cdot t$ and $dg_i \cdot t \cdot \mathbf{x}_{it}$., are included. These are added to $\mathbf{g}_{it}$, with $\mathbf{h}_{it}$ the same as before. The equivalence follows because $w_{it} \cdot dg_i \cdot t$ is a linear combination of $dg_i \cdot fg_t, \ldots,$ $dg_i \cdot fT$, which is in $\mathbf{h}_{it}$: the coefficients in the linear combination are g, $g+1$, ..., T. Similarly, $w_{it} \cdot dg_i \cdot t \cdot \mathbf{x}_{it}$ is a linear combination of $dg_i \cdot fg_t, \ldots, dg_i \cdot fT$, $dg_i \cdot fg_t \cdot \dot{\mathbf{x}}_{itgg}, \ldots, dg_i \cdot fT \cdot \dot{\mathbf{x}}_{itgT}$, which also is in $\mathbf{h}_{it}$. The second-step imputation regression is still (A20).

Acknowledgements Over the past several years, many people provided valuable input on early drafts of this paper and related talks. I would like to particularly thank Soren Anderson, Trang Hoang, Kamyar Nasseh, Whitney Newey, Jon Roth, Pedro Sant'Anna, Jeff Zabel, participants at the 15th New York Camp Econometrics, participants at the University of Michigan Two-Way Fixed Effects workshop, and seminar participants at Michigan State, Microsoft, University College London, Princeton, Brown, MIT/Harvard, Tulane, the Norwegian Business School, and the Agricultural and Applied Economics Association. I would also like to acknowledge the helpful comments of the guest editor, Badi Baltagi, and two anonymous referees.

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2023) When should you adjust standard errors for clustering? Quart J Econ 138(1):1–35

Athey S, Imbens GW (2022) Design-based analysis in difference-in-differences settings with staggered adoption. J Econom 226(1):62–79

Baltagi BH (2021) Econometric analysis of panel data 6e

Baltagi BH (2023) The two-way mundlak estimator. Economet Rev 42:240–246

Baltagi BH, Wansbeek T (2025) The basics of the mundlak and chamberlain projections. Seven decades of econometrics and beyond (pp. 395-415). Springer

Borusyak K, Jaravel X, Spiess J (2024) Revisiting event-study designs: Robust and efficient estimation. Rev Econ Stud 91:3253–3285

Brown NL, Butts K (2025) Dynamic treatment effect estimation with interactive fixed effects and short panels. J Econom 250:106013

Caetano C, Callaway B (2024) Difference-in-differences when parallel trends holds conditional on covariates. arXiv preprint arXiv:2406.15288

Callaway B, Sant'Anna PH (2021) Difference-in-differences with multiple time periods. J Econom 225:200–230

Deb P, Norton EC, Wooldridge JM, Zabel JE (2024) A flexible, heterogeneous treatment effects difference-in-differences estimator for repeated cross-sections (Tech. Rep.). National Bureau of Economic Research

De Chaisemartin C, d'Haultfoeuille X (2020) Two-way fixed effects estimators with heterogeneous treatment effects. Am Econom Rev 110(9):2964–2996

De Chaisemartin C, d'Haultfoeuille X (2023) Two-way fixed effects estimators with heterogeneous treatment effects: A survey. Economet J 26(3):C1–C30

De Chaisemartin C, d'Haultfoeuille X (2024) Difference-in-differences estimators of intertemporal treatment effects. Review of Economics and Statistics, forthcoming

Gardner J (2022) Two-stage differences in differences. arXiv:2207.05943

Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. J Econom 225(2):254–277

Harmon NA (2024) Difference-in-differences and efficient estimation of treatment effects. Working paper

Hausman JA (1978) Specification tests in econometrics. Econometrica 46(6):1251–1271

Lee SJ, Wooldridge JM (2023) A simple transformation approach to differencein- differences estimation for panel data. Available at SSRN 4516518 , ,

Mundlak Y (1978) On the pooling of time series and cross section data. Econometrica 46(1):69–85

Rambachan A, Roth J (2023) A more credible approach to parallel trends. Rev Econ Stud 90(5):2555–2591

Roth J (2022) Pretest with caution: Event-study estimates after testing for parallel trends. Am Econom Rev Insights 4(3):305–322

Roth J, Sant'Anna PH, Bilinski A, Poe J (2023) What's trending in differencein- differences? a synthesis of the recent econometrics literature. J Econom 235(2):2218–2244

Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J Econom 225(2):175–199

Wooldridge JM (2005) Fixed-effects and related estimators for correlated randomcoefficient and treatment-rffect panel data models. Rev Econ Stat 87(2):385–390

Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT press, Cambridge

Wooldridge JM (2019) Correlated random effects models with unbalanced panels. J Econom 211(1):137–150

Wooldridge JM (2023) Simple approaches to nonlinear difference-in-differences with panel data. Economet J 26(3):C31–C66

Yang Y (2022) A correlated random effects approach to the estimation of models with multiple fixed effects. Econ Lett 213:110408