A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data

Soo Jeong Lee
Department of Economics
Michigan State University

Jeffrey M. Wooldridge
Department of Economics
Michigan State University

This version: July 20, 2023

Abstract: In the case of panel data, we propose a simple time-series transformation that can be combined with various treatment effect estimators, including regression adjustment, matching methods, and doubly robust estimators. The approach is motivated by the fact that, in the common timing case, our transformation, when applied with linear regression adjustment, numerically reproduces the pooled OLS estimator in Wooldridge (2021). In the general staggered case, the transformation is at the unit level, and simply requires computing the average outcome prior to an intervention, subtracting it from a post-treatment outcome, and then carefully selecting the control units in each time period. We show formally that, allowing for staggered entry under no anticipation and parallel trends assumptions, the cohort treatment indicators satisfy the key unconfoundedness assumption with respect to the transformed potential outcome. Given identification, any number of treatment effect estimators can be applied for each treated cohort and calendar time pair where the average treatment effects on the treated are identified. In effect, we establish the consistency of intuitively appealing rolling methods. The doubly robust method of combining inverse probability weighting with linear regression works particularly well in terms of bias and efficiency. Long differencing methods, such as those proposed by Callaway and Sant'Anna (2021), can be considerably less efficient. We also show how to modify the transformation to account for unit-specific trends.

Keywords: Difference-in-differences; panel data; parallel trends, doubly robust estimators; heterogenous trends

JEL Classification Codes: C23, C54

1. Introduction

The two-way fixed effects (TWFE) estimator, applied to a linear panel model with a constant treatment effect, has been commonly applied in difference-in-differences settings. The TWFE estimator of a single effect is simple to understand and is taught in courses that cover panel data methods. Recently, several authors have pointed out shortcomings of the constant effect model, especially when the intervention is staggered. These include Borusyak and Jaravel (2018), de Chaisemartin and D'Haultfoeuille (2020), and Goodman-Bacon (2021), who establish different representations of the simple TWFE estimator when the treatment effects (TEs) are heterogeneous across treatment cohort or calendar time and the intervention is staggered.

Other authors have proposed more flexible estimation methods that uncover average treatment effects on the treated in the staggered intervention case. These include Callaway and Sant'Anna (2021) [CS (2021)], who propose long-differencing strategies and apply standard treatment effect estimators. Sun and Abraham (2021) [SA (2021)] propose a fixed effects estimator applied to a more flexible model. Both SA (2021) and CA (2021) are event-study-type estimators that use only the single period prior to the first intervention time as the control period. Wooldridge (2021) shows that a pooled OLS (POLS) strategy that includes cohort and calendar time interactions, as well as interactions of cohort dummies, time period dummies, and the treatment indicators with covariates, identifies the ATTs under standard no anticipation and parallel trends assumptions. These estimators effectively use all pre-treatment periods and all not-yet-treated units in the control group. Wooldridge (2021) also shows the POLS is equivalent to a TWFE estimator on an expanded equation that includes interactions of cohort and time dummies with each other and with covariates. Borusyak, Jaravel, and Spiess (2022) propose imputation estimators based on pooled OLS regressions that can include unit and time fixed effects. Wooldridge (2021) shows that, with time constant covariates, the imputation estimates are identical to POLS (and therefore TWFE) estimation of the flexible model using the entire sample.

An attractive feature of the CS (2021) approach, which builds on Abadie (2005) for the two-period case, is that it permits the application of treatment effects estimators beyond regression adjustment. However, as mentioned above, the CS (2021) method uses only the period just prior to the intervention in defining the control group, thereby discarding potentially useful information in earlier time periods. In fact, Wooldridge (2021) shows that, under the standard "error components" structure on the error, with a homoskedastic time-constant component and homoskedastic and serially uncorrelated idiosyncratic errors, the POLS estimator is both best linear unbiased (BLUE) and asymptotically efficient. These theoretical results imply that the CS (2021) estimators are inefficient under a standard set of assumptions. The simulations in Wooldridge (2021) bear this out, showing the CS approach can be very inefficient. Balanced against the loss in precision is that the CS approach can be less biased when parallel trends are violated. See de Chaisemartin and D'Haultfoeuille (2023) and Wooldridge (2021) for further discussion.

In this paper, we propose an alternative "rolling" approach that allows for the application of many different treatment effects estimators while maintaining much of the efficiency of regression-based methods. The idea is to use as many control observations as possible – in both the common timing and staggered cases – while permitting methods such as inverse probability weighting (IPW), doubly robust methods such as the one in Wooldridge (2007) that combines regression and IPW (IPWRA), and matching on covariates or the propensity score. Like CS (2021) in the panel data case, our approach is based on time series transformations at the unit level. Rather than using long differences, we show how to use all suitable control observations in transforming the outcome variable. This leads to significant improvements in efficiency compared with CS (2021) and allows one substantial flexibility in the choice of treatment effects estimators.

In the case of common timing – so that there is one average treatment effect per post-treatment period – we show that applying regression adjustment to our transformed outcome variable is equivalent to the regression adjustment estimator based on levels. As mentioned above, Wooldridge (2021) shows that this estimator is both BLUE and asymptotically efficient under standard assumptions. This provides strong motivation for applying estimators other than regression adjustment to the transformed variables in order to check robustness of findings. In the case of staggered entry, our approach identifies the average treatment effects on the treated (ATTs) by cohort and calendar time under the same no anticipation and parallel trends assumptions as in Wooldridge (2021). We show this in both the case of common timing and staggered interventions. Once identification is established, various estimation methods can be applied.

The remainder of the paper is organized as follows. Section 2 begins with the common timing case, defining the potential outcomes and parameters of interest, and establishing identification under no anticipation, conditional parallel trends, and overlap assumptions. In Section 3 we propose a general approach to estimation using a transformed outcome variable. We also show that regression adjustment applied to the new transformation is identical to pooled OLS estimator in Wooldridge (2021).

In Section 4 we extend the framework and identification argument to the staggered case, where there is more than one treatment cohort and each cohort may be treated in multiple periods. The transformation is applied by cohort, time period pair before applying standard treatment effect estimators. In Section 5, we show how we can account for heterogeneous trends, focusing on linear trends, to allow violation of the parallel trends assumption (even after we condition on covariates). Section 6 discusses how one might accommodate suspected failures of no anticipation, and how one modifies the procedure for unbalanced panels.

Section 7 contains simulations to show that the new approach works well in terms of both bias and precision. Section 8 contains some concluding remarks

2. Setup and Identification in the Common Timing Case

In this section we assume that the date of the intervention is the same for all treated units and then the intervention is in place through the final period. The time periods in the population are $t = 1, 2, \ldots, T$ and the date of the intervention is S, where $1 < S \leq T$; in other words, there is at least one pre-treatment period. The arguments in this section are based on an underlying population, and so we use $\{Y_t(0), Y_t(1) : t = 1, \ldots, T\}$ to denote the time series of outcomes in the control and treated states.

The binary time-constant treatment indicator is D, where $D = 1$ means treatment starting in period S and lasting through period T. A time-varying treatment indicator is $W_t = D \cdot p_t$, where p_t is a post-treatment period indicator: $p_t = 1$ if $t \geq S$ and equals zero of $t < S$.

Without treated units prior to time S we can, at most, hope to identify average treatment effects in periods S, $S + 1$, ..., T. Our focus here, like almost all of the other recent literature, is on the average treatment effect on the treated (ATT or ATET) in each treated period:

$$\tau_r = E[Y_r(1) - Y_r(0)|D = 1], \; r = S, \ldots, T \tag{2.1}$$

The methods we propose can, under stronger assumptions than we propose, recover the overall average treatment effects (ATEs), $E[Y_r(1) - Y_r(0)]$, and we will mention how that can be done.

The fundamental problem of identification of τ_r is that we only observe the treatment status, D, and the outcome

$$Y_r = (1 - D) \cdot Y_r(0) + D \cdot Y_r(1) \tag{2.2}$$

(Shortly we will introduce a vector of covariates). Importantly, when $D = 1$, $Y_r = Y_r(1)$, which means

$$E[Y_r(1)|D = 1] = E(Y_r|D = 1), r = S, \ldots, T \tag{2.3}$$

The expectation $E(Y_r|D = 1)$ can be estimated in a consistent,even unbiased, way under various sampling schemes. Under random sampling, the average of Y_r across the treated subsample is unbiased and consistent. Therefore, writing

$$\tau_r = E[Y_r(1)|D = 1] - E[Y_r(0)|D = 1] = E(Y_r|D = 1) - E[Y_r(0)|D = 1],$$

it is easily seen that the challenge is in identifying $E[Y_r(0)|D = 1]$.

If the treatment is randomly assigned with respect to $Y_r(0)$, then $E[Y_r(0)|D = 1]$ $= E[Y_r(0)|D = 0]$. Because $Y_r = Y_r(0)$ when $D = 0$, $E[Y_r(0)|D = 0] = E(Y_r|D = 0)$ is consistently estimated using the control units under various sampling schemes. Under random sampling, one would use the sample average of Y_r across the control units. The resulting estimator of τ_r would be the simple difference in sample means between the treated and control units in period r.

The assumption of random assignment is too strong for most applications. To see how to relax it, use simple algebra to write

$$\begin{aligned} Y_r(1) - Y_r(0) &= \left[Y_r(1) - \frac{1}{(S-1)} \sum_{q=1}^{S-1} Y_q(1) \right] - \left[Y_r(0) - \frac{1}{(S-1)} \sum_{q=1}^{S-1} Y_q(0) \right] \\ &\quad + \frac{1}{(S-1)} \sum_{q=1}^{S-1} [Y_q(1) - Y_q(0)] \\ &\equiv \dot{Y}_r(1) - \dot{Y}_r(0) + \frac{1}{(S-1)} \sum_{q=1}^{S-1} [Y_q(1) - Y_q(0)] \end{aligned} \tag{2.4}$$

where

$$\dot{Y}_r(1) \equiv Y_r(1) - \frac{1}{(S-1)} \sum_{q=1}^{S-1} Y_q(1) \tag{2.5}$$

and similarly for $\dot{Y}_r(0)$. Note that for each $r \in \{S, S+1, \ldots, T\}$, $\dot{Y}_r(1)$ is the time r potential outcome with the average of the pre-treatment period outcomes removed. The third term is the average of the difference of the pre-treatment period “treatment” effects.

Given the representation in (2.4), we can write

$$\tau_r = E[\dot{Y}_r(1)|D = 1] - E[\dot{Y}_r(0)|D = 1] + \frac{1}{(S-1)} \sum_{q=1}^{S-1} E[Y_q(1) - Y_q(0)|D = 1] \tag{2.6}$$

The first assumption, a weak version of "no anticipation," eliminates the third term in (2.6).

Assumption NAC (**No Anticipation**, **Common Timing**): For the eventually treated indicator D,

$$E[Y_t(1) - Y_t(0)|D = 1] = 0\ ,\ t = 1,\ldots,S-1.\ \square \tag{2.7}$$

The name of this assumption derives from the fact that $E[Y_t(1) - Y_t(0)|D = 1]$ for $t < S$ are average treatment effects on the treated prior to the intervention, and the assumption is that these are all zero. Assumption NAC is implied by an assumption commonly used in the literature, namely, $Y_t(1) = Y_t(0)$, $t = 1,\ldots,S-1$. This assumption is implicit in Heckman, Ichimura, and Todd (1997) and made explicit in Abadie (2005) and elsewhere. Because the variable indexing $Y_t(\bullet)$ is treatment status not yet assigned, the assumption rules out anticipatory changes in the potential outcomes, on average. If one is concerned about anticipation of a policy that is announced prior to its being implemented, one might drop a period or two just prior to the intervention – as a minimum, as a robustness check. Naturally, this will result in less precise estimators in general.

Given Assumption NAC, we can express τ_r as

$$\tau_r = E[\dot{Y}_r(1)|D = 1] - E[\dot{Y}_r(0)|D = 1]. \tag{2.8}$$

Estimating the first term in (2.8) is easy because we observe $\dot{Y}_r(1)$ when $D = 1$. More precisely, define the same transformation in the observed variable Y_r:

$$\dot{Y}_r \equiv Y_r - \frac{1}{S-1} \sum_{q=1}^{S-1} Y_q \equiv Y_r - \bar{Y}_{pre} \tag{2.9}$$

When $D = 1$, $\dot{Y}_r = \dot{Y}_r(1)$ and so $E[\dot{Y}_r(1)|D = 1] = E(\dot{Y}_r|D = 1)$ and the latter is trivially identified (as usual, under a suitable sampling scheme). Notice in the simple $T = 2$ case with

$S = 1$, $\dot{Y}_2 = Y_2 - Y_1$, the difference from period one to two.

The difficult term in identifying τ_r is $E[\dot{Y}_r(0)|D = 1]$. The unconditional parallel trends assumption implies that $E[\dot{Y}_r(0)|D = 1] = E[\dot{Y}_r(0)|D = 0]$. Here we allow a weaker version of parallel trends by assuming it holds conditional on observed (pre-treatment) covariates.

Assumption CPTC (**Conditional Parallel Trends**, **Common Timing**): For observed covariates $\mathbf{X}$,

$$E[Y_t(0) - Y_1(0)|D, \mathbf{X}] = E[Y_t(0) - Y_1(0)|\mathbf{X}],\ t = 2, \ldots, T.\ \square \tag{2.10}$$

Simple algebra shows that (2.10) is the same as assuming $E[Y_t(0) - Y_s(0)|D, \mathbf{X}] = E[Y_t(0) - Y_s(0)|\mathbf{X}]$ for all $t \neq s$. Wooldridge (2021) used very similar assumptions, along with linearity of conditional means, to derive identification of the τ_r. Here we are interested in applying methods other than regression adjustment to the transformed outcomes in (2.8). Assumption CPTC allows us to identify $E[\dot{Y}_r(0)|D = 1]$. To see how, first note that, by iterated expectations,

$$E[\dot{Y}_r(0)|D = 1] = E\{E[\dot{Y}_r(0)|D = 1, \mathbf{X}]|D = 1\} \tag{2.11}$$

Next, write

$$\dot{Y}_r(0) = (S-1)^{-1} \sum_{q=1}^{S-1} [Y_r(0) - Y_q(0)]$$

Then, by CPTC,

$$\begin{aligned} E[\dot{Y}_r(0)|D = 1, \mathbf{X}] &= (S-1)^{-1} \sum_{q=1}^{S-1} E[Y_r(0) - Y_q(0)|D = 1, \mathbf{X}] \\ &= (S-1)^{-1} \sum_{q=1}^{S-1} E[Y_r(0) - Y_q(0)|D = 0, \mathbf{X}] \\ &= E\left[Y_r(0) - (S-1)^{-1} \sum_{q=1}^{S-1} Y_q(0) \middle| D = 0, \mathbf{X} \right] \\ &= E[\dot{Y}_r(0)|D = 0, \mathbf{X}] \end{aligned} \tag{2.12}$$

The conclusion in equation (2.12) is simple but important. It says that, in terms of the potential outcome $\dot{Y}_r(0)$, treatment D is unconfounded conditional on $\mathbf{X}$. Assumption NA ensures that the ATTs can be expressed as in (2.8). This means that, for a post-intervention period r, we have turned the difference-in-differences problem into a standard problem of estimating an ATT in a cross-sectional population.

Using the fact that $Y_q = Y_q(0)$ when $D = 0$, (2.12) implies that,

$$E[\dot{Y}_r(0)|D = 1, \mathbf{X}] = E(\dot{Y}_r|D = 0, \mathbf{X})$$

Now the argument is the same as in the typical cross section setting: By iterated expectations,

$$\begin{aligned} E[\dot{Y}_r(0)|D = 1] &= E[E(\dot{Y}_r|D = 0, \mathbf{X})|D = 1] \\ &\equiv E[\dot{m}_{0r}(\mathbf{X})|D = 1], \end{aligned} \tag{2.13}$$

where $\dot{m}_{0r}(\mathbf{X}) \equiv E(\dot{Y}_r|D = 0, \mathbf{X} = \mathbf{x})$ is the conditional mean of the observed variable $\dot{Y}_r$ for the control group. This function is nonparametrically identified on Supp($\mathbf{X}|D = 0$), the support of the covariates for the control group. To ensure we can compute $E[\dot{m}_{0r}(\mathbf{X})|D = 1]$ without extrapolation to covariate values outside Supp($\mathbf{X}|D = 0$), we impose a standard overlap assumption.

Assumption OVLC (**Overlap**, **Common Timing**): Define the propensity score

$$p(\mathbf{x}) = P(D = 1|\mathbf{X} = \mathbf{x}),\ \mathbf{x} \in \text{Supp}(\mathbf{X}). \tag{2.14}$$

Then

$$p(\mathbf{x}) < 1,\ \mathbf{x} \in \text{Supp}(\mathbf{X}).\ \square \tag{2.15}$$

The previous derivations and discussion prove the following.

Theorem 2.1: Under Assumption NAC, τ_r can be expressed as in (2.8) for $r = S, \ldots, T$. Under Assumption CPTC, D is unconfounded (in the conditional mean sense) with respect to $\dot{Y}_r(0)$ conditional on $\mathbf{X}$. When we add Assumption OVLC, the parameters τ_r, $r = S, \ldots, T$, are identified. □

3. Estimation in the Common Timing Case

Given the identification result stated in Theorem 2.1, the estimation of the τ_r is straightforward. We can apply any estimation method once the outcome variable has been transformed as in equation (2.9). Essentially, this is the conclusion reached in Sant'Anna and Zhou (2020) in the $T = 2$ case. Earlier, Abadie (2005) proposed inverse probability weighting when $T = 2$.

For simplicity, assume in this section we observe a random sample of size N from the cross section. The observed outcome can be expressed as

$$Y_{it} = (1 - D_i) \cdot Y_{it}(0) + D_i \cdot Y_{it}(1) \tag{3.1}$$

where we use an i subscript to denote unit i. Under the strong form of no anticipation, $Y_{it}(1) = Y_{it}(0)$ for $t < S$ and all i. Given the derivations in the previous section, we only need Assumption NAC. For each i, we observe the time series $\{(Y_{it}, D_i, \mathbf{X}_i) : i = 1, 2, \ldots N\}$. To exploit the unconfoundedness and identification in Theorem 2.1, we simply need to obtain the transformed data. For each unit i, define

$$\dot{Y}_{ir} = Y_{ir} - \frac{1}{(S-1)} \sum_{q=1}^{S-1} Y_{iq} \equiv Y_{ir} - \bar{Y}_{i,pre} \tag{3.2}$$

Then, for any $r \in \{S, S+1, \ldots, T\}$, we can apply any standard treatment effect (TE) estimator to the data $\{(\dot{Y}_{ir}, D_i, \mathbf{X}_i) : i = 1, 2, \ldots N\}$.

A common TE estimator is called "regression adjustment," which means estimating separate regression functions for the control and treated units. Because $\dot{Y}_{ir}$ can take on negative and positive values, linear regression adjustment (RA) makes the most sense. Linear RA is based on the conditional mean, stated in terms of population random variables,

$$E(\dot{Y}_r | D = 0, \mathbf{X}) = \dot{\alpha}_r + \mathbf{X}\dot{\boldsymbol{\beta}}_r \tag{3.3}$$

The parameters $\dot{\alpha}_r$ and $\dot{\boldsymbol{\beta}}_r$ are estimated from the cross-sectional regression

$$\dot{Y}_{ir} \text{ on } 1, \mathbf{X}_i \text{ if } D_i = 0 \tag{3.4}$$

Then, τ_r can be estimated using imputation:

$$\hat{\tau}_r = \bar{\dot{Y}}_{r1} - N_1^{-1} \sum_{i=1}^{N} D_i\left(\hat{\dot{\alpha}}_r + \mathbf{X}_i\hat{\dot{\boldsymbol{\beta}}}_r\right) = \bar{\dot{Y}}_{r1} - \left(\hat{\dot{\alpha}}_r + \bar{\mathbf{X}}_1\hat{\dot{\boldsymbol{\beta}}}_r\right) \tag{3.5}$$

where $\bar{\dot{Y}}_{r1} = N_1^{-1} \sum_{i=1}^{N} D_i\dot{Y}_{ir}$ and $\bar{\mathbf{X}}_1 = N_1^{-1} \sum_{i=1}^{N} D_i\mathbf{X}_i$ are the averages over the treated units. From a practical perspective, the important thing to remember is that $\hat{\tau}_r$ can be obtained from standard software that does basic regression adjustment once the $\dot{Y}_{ir}$ have been obtained.

As discussed in Wooldridge (2021), $\hat{\tau}_r$ also can be obtained as the coefficient on D_i in the pooled OLS regression

$$\dot{Y}_{ir} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1), i = 1,2,\ldots,N, \tag{3.6}$$

which uses all observations in time period r. This formulation is convenient because it leads to simple inference for $\hat{\tau}_r$, allowing easy computation of standard errors robust to any kind of heteroskedasticity. Also, it is often easy to account for the sampling variation in $\bar{\mathbf{X}}_1$ as an estimator of $\boldsymbol{\mu}_1 \equiv E(\mathbf{X}|D = 1)$. Issues of clustering standard errors are relatively easy to deal with given we have a standard cross-sectional regression.

Because of the representation of $\dot{Y}_{ir}$ in (3.2), there is a simple characterization of $\hat{\tau}_r$. All of the coefficients in (3.6) are obtained by differencing the coefficients from two separate regressions. In the first, Y_{ir} is regressed on all variables in (3.6). Then $\bar{Y}_{i,pre}$ is regressed on the same set of variables, and these coefficients are subtracted from the first. In particular, $\hat{\tau}_r$ is obtained as the difference between two standard ATT estimators using regression adjustment. The first uses observations in period r only, and the second uses the average of Y_{iq} over the pre-treatment periods. Without covariates, the estimator would be

$$\hat{\tau}_r = (\bar{Y}_{1r} - \bar{Y}_{0r}) - (\bar{Y}_{1,pre} - \bar{Y}_{0,pre}) \tag{3.7}$$

where the first subscript indicates treatment or control units. This has a clear interpretation as a DiD estimator.

Recognizing that an estimator can be obtained from (3.6) has additional benefits. For example, if D_i is independent of $\dot{Y}_{ir}(0)$, then the covariates $\mathbf{X}_i$ need not be included in (3.6) in order to consistently estimate τ_r as the coefficient on D_i. Remember, this allows D_i to be correlated with, the level, say, $Y_1(0)$, the potential outcome in the first time period. If, in addition, D_i is independent of $\mathbf{X}_i$, the regression in (3.6) still can be used to improve efficiency over the simple estimator without the covariates. As discussed in Negi and Wooldridge (2021), such improvements are possible if $\mathbf{X}_i$ helps predict $\dot{Y}_{ir}$. In many cases, $\mathbf{X}_i$ may not have much predictive power for $\dot{Y}_{ir}$ even though it might predict the level, Y_{ir}, well. A special case is $T = 2$, in which case (3.6) is simply ΔY_i on 1, D_i, $\mathbf{X}_i$, $D_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1)$, $i = 1, 2, \ldots, N$ where $\Delta Y_i = Y_{i2} - Y_{i1}$. In the $T = 2$ case, whether including $\mathbf{X}_i$ substantively helps precision when D_i is independent of $\mathbf{X}_i$ hinges on how well $\mathbf{X}_i$ predicts the difference, ΔY_i.

It turns out there is another useful algebraic equivalence. Suppose we act *as if* the following conditional expectation holds for all population units and time periods:

$$\begin{aligned} E(Y_t|D,\mathbf{X}) &= \alpha + \mathbf{X}\boldsymbol{\beta} + \gamma D + (D \cdot \mathbf{X})\boldsymbol{\delta} + \sum_{r=2}^{T} \theta_r fr_t + \sum_{r=2}^{T} (fr_t \cdot \mathbf{X})\boldsymbol{\pi}_r \\ &+ \sum_{r=S}^{T} \tau_r (D \cdot fr_t) + \sum_{r=S}^{T} (D \cdot fr_t)(\mathbf{X} - \boldsymbol{\mu}_1)\boldsymbol{\eta}_r,\ t = 1, \ldots, T, \end{aligned} \quad (3.8)$$

where fr_t is a time period dummy equal to one if $r = t$ and zero otherwise. The interaction $D \cdot fr_t$ is the treatment indicator for time period r. Equation (3.6) suggests a pooled OLS regression across all i and t:

$$\begin{aligned} &Y_{it} \text{ on } 1, \mathbf{X}_i, D_i, D_i \cdot \mathbf{X}_i, f2_t, \ldots, fT_t, f2_t \cdot \mathbf{X}_i, \ldots, fT_t \cdot \mathbf{X}_i \\ &\quad D_i \cdot fS_t, \ldots, D_i \cdot fT_t, D_i \cdot fS_t \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1), \ldots, D_i \cdot fT_t \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1) \end{aligned} \quad (3.9)$$

The estimated treatment effects, say $\tilde{\tau}_r$, are the coefficients on the treatment dummies $D_i \cdot fS_t$, ..., $D_i \cdot fT_t$. Wooldridge (2021) shows that the $\tilde{\tau}_r$ are numerically identical to a two-stage imputation approach based on the levels, Y_{it}. It turns out that the $\tilde{\tau}_r$ are also equivalent to the $\hat{\tau}_r$ obtained by using the transformed outcome variable, $\dot{Y}_{ir}$, one period at a time.

Theorem 3.1: Let $\hat{\tau}_r$, $r = S, S+1, \ldots, T$ be the coefficients on D_i in the separate regressions (3.6) – equivalently, from equation (3.5) – and let $\tilde{\tau}_r$ be the coefficients on $D_i \cdot fS_t, \ldots, D_i \cdot fT_t$ from (3.9). Then $\tilde{\tau}_r = \hat{\tau}_r$, $r = S, \ldots, T$. Moreover, the coefficient vector on $D_i \cdot fr_t \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1)$ in (3.9) is identical to that on $D_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1)$ in (3.6). □

The proof of Theorem 3.1 is given in the appendix. The equivalence is valuable for a couple of reasons. First, it shows that two different ways to approach identification under the same set of assumptions – that in Wooldridge (2021) and the approach we use here – leads to the same estimation methods. Second, Wooldridge (2021, Theorem 6.2) shows that, under standard assumptions on the implied error term (which includes a unit-specific unobserved effect and a time-varying component), the estimators from (3.9) are both best linear unbiased and asymptotically efficient (with T fixed, $N \rightarrow \infty$) under random sampling across i. This establishes that the transformation used in (3.6) does not discard useful information.

Given the equivalence of our transformation approach and the OLS estimator pooled across i and t, what use is the former? Importantly, it allows us to use other treatment effects estimators beyond regression adjustment. For example, we can apply IPW or, even better, IPWRA, using the cross-sectional data $\{(\dot{Y}_{ir}, D_i, \mathbf{X}_i) : i = 1, 2, \ldots N\}$. We can also apply propensity score matching or nearest neighbor matching.

Procedure 3.1 (Rolling Methods, Common Timing):

1. For a given time period $r \in \{S, S+1, \ldots, T\}$ and each unit i, compute $\dot{Y}_{ir}$ as in (3.2).

2. Using all of the units, apply standard TE methods – such as linear RA, IPW, IPWRA, matching – to the cross section

$$\{(\dot{Y}_{ir}, D_i, \mathbf{X}_i) : i = 1, \ldots, N\}. \ \square$$

Inference on a single τ_r is simple when one uses built-in commands in step (2) of

Procedure 3.1. Joint inference on multiple τ_r is trickier because the estimators are not independent. For estimators such as IPW and IPWRA using parametric models, a general approach is to stack all moment conditions used in estimation and use the formulas from just-identified generalized method of moments estimation. Applying the panel bootstrap – resampling all time periods from the cross-sectional units – is valid for IPW and IPWRA, and should be computationally feasible in most cases.

It is instructive to compare the transformation in equation (3.2) to that in Callaway and Sant'Anna (2021). In the common timing case, the CS transformation, for $r \geq S$, is

$$\mathring{Y}_{ir} = Y_{ir} - Y_{i,S-1}, \tag{3.10}$$

so that $\mathring{Y}_{ir}$ is a "long" difference. (If $r = S$ then $\mathring{Y}_{iS} = Y_{iS} - Y_{i,S-1}$, which is differencing adjacent periods.) The CS transformation is generally inefficient compared with (3.2) because the CS (2021) differencing ignores time periods other than the one just prior to the intervention. When $T = 2$, the transformation is $\dot{Y}_2 = \mathring{Y}_2 = \Delta Y_2 \equiv Y_2 - Y_1$. Thus, our approach encompasses and extends Abadie (2005) and Sant'Anna and Zhou (2020) in the panel data case.

4. Staggered Interventions

4.1. Some Units Never Treated

We now turn to the staggered intervention case. As in Athey and Imbens (2022) and Wooldridge (2021, 2023), the potential outcomes are denoted

$$Y_t(g),\ g \in \{S,\ldots,T,\infty\},\ t \in \{1,2,\ldots,T\}, \tag{4.1}$$

where g indicates the first time subjected to the intervention – defining the treatment group cohorts – and t is calendar time. The case $g = \infty$ indicates the potential outcome in the never treated state. In other words, $Y_t(\infty)$ is the potential outcome at time t when a unit is not subjected to the intervention over the observed stretch of time. Listing potential outcomes that vary only by cohort and calendar time reflects the assumption of no reversibility with staggered entry.

We denote the group or cohort indicators by $\{D_S, D_{S+1}, \ldots, D_T, D_\infty\}$, where $D_\infty = 1$ indicates never treated. These dummy variables are mutually exclusive and exhaustive: $D_S + \cdots + D_T + D_\infty = 1$.

The ATTs of interest are now written as

$$\tau_{gr} = E[Y_r(g) - Y_r(\infty)|D_g = 1],\, r = g, \ldots, T;\; g = S, \ldots, T \tag{4.2}$$

For each (eventually) treated cohort g, τ_{gr}, $r = g, \ldots, T$ are the ATTs in all subsequent time periods.

To identify the τ_{gr}, we extend the trick for the common timing case by writing

$$\begin{aligned} Y_t(g) - Y_t(\infty) = & \left[Y_t(g) - \frac{1}{(g-1)} \sum_{s=1}^{g-1} Y_s(g) \right] \\ & - \left[Y_t(\infty) - \frac{1}{(g-1)} \sum_{s=1}^{g-1} Y_s(\infty) \right] \\ & + \frac{1}{(g-1)} \sum_{s=1}^{g-1} [Y_s(g) - Y_s(\infty)] \end{aligned} \tag{4.3}$$

As in the common timing case, we make a no anticipation assumption so that the third term can be dropped and that effectively allows using all available control units in each treated period. Here we condition on the covariates so that we can use not-yet-treated units as part of the control group.

Assumption CNAS (**Conditional No Anticipation**, **Staggered**): For $g \in \{S, \ldots, T\}$, $t \in \{1, \ldots, g-1\}$ and covariates $\mathbf{X}$,

$$E[Y_t(g)|D_g = 1, \mathbf{X}] = E[Y_t(\infty)|D_g = 1, \mathbf{X}]. \;\square \tag{4.4}$$

As in the common timing case, this assumption means that the "treatment" effects prior to the intervention are all zero. Because $s < g$ in the third sum, it follows that the expected value of the last term conditional on $D_g = 1$ is zero. Therefore,

$$\tau_{gr} = E[\dot{Y}_{rg}(g)|D_g = 1] - E[\dot{Y}_{rg}(\infty)|D_g = 1] \tag{4.5}$$

where $\dot{Y}_{rg}(g)$ and $\dot{Y}_{rg}(\infty)$ are defined as the first and second terms in (4.3), respectively. Note that the first subscript on $\dot{Y}_{rg}(g)$ and $\dot{Y}_{rg}(\infty)$ means that we are averaging all periods just prior to g and subtracting from the outcome in the current current calendar time period r.

As before, the first term in equation (4.3) is easily estimated because we observe $\dot{Y}_{rg}(g)$ when $D_g = 1$. A parallel trends assumption, stated in terms of the never treated state, is sufficient to identify $E[\dot{Y}_{rg}(\infty)|D_g = 1]$. We state an assumption conditional on a set of covariates, $\mathbf{X}$, with no covariates as a special case.

Assumption CPTS (**Conditional PT**, **Staggered**): For $\mathbf{D} = (D_S,\ldots,D_T)$ and $t = 1,2,\ldots,T$,

$$E[Y_t(\infty) - Y_1(\infty)|\mathbf{D},\mathbf{X}] = E[Y_t(\infty) - Y_1(\infty)|\mathbf{X}],\ t = 2,\ldots,T.\ \square \tag{4.6}$$

This assumption is used in Wooldridge (2021). Again, it is unconfoundedness of the treatment level, as given by $\mathbf{D}$, with respect to the trend in the untreated state, $Y_t(\infty) - Y_1(\infty)$. Wooldridge (2021) used this assumption, along with linearity of conditional means, to derive an imputation estimator and showed it was the same as a pooled OLS and TWFE estimator. Here we show how it can be used to identify the τ_{gr} very generally.

With $\dot{Y}_{rg}(\infty)$ defined above,

$$\begin{aligned} E[\dot{Y}_{rg}(\infty)|D_g = 1,\mathbf{X}] &= \frac{1}{(g-1)}\sum_{s=1}^{g-1} E[Y_r(\infty) - Y_s(\infty)|D_g = 1,\mathbf{X}] \\ &= \frac{1}{(g-1)}\sum_{s=1}^{g-1} E[Y_r(\infty) - Y_s(\infty)|D_\infty = 1,\mathbf{X}] \\ &= E[\dot{Y}_{rg}(\infty)|D_\infty = 1,\mathbf{X}] \end{aligned} \tag{4.7}$$

where the second equality follows from CPTS and the third follows by taking the expectation outside the summation.

We have shown the following.

Theorem 4.1: Under Assumption CNAS, equation (4.5) holds. If we add Assumption CPTS, the cohort assignments, $\mathbf{D} = (D_S, D_{S+1},\ldots,D_T)$ are unconfounded with respect to

$\dot{Y}_{rg}(\infty)$ (in the conditional mean sense), $g \in \{S,\ldots,T\}$, $r \in \{g,\ldots,T\}$, conditional on $\mathbf{X}$. □

Because the vector of cohort indicators is unconfounded with respect to $\dot{Y}_{rg}(\infty)$, Theorem 4.1 implies

$$E[\dot{Y}_{rg}(\infty)|D_\infty = 1,\mathbf{X}] = E[\dot{Y}_{rg}(\infty)|D_h = 1,\mathbf{X}],\ h = S,\ldots,T \tag{4.8}$$

We can combine this implication of CPTS with Assumption CNAS because, at time r, cohorts $h = r+1,\ldots,T$ have yet to be treated. Therefore,

$$E[\dot{Y}_{rg}(\infty)|D_h = 1,\mathbf{X}] = E[\dot{Y}_{rg}(h)|D_h = 1,\mathbf{X}],\ h = r+1,\ldots,T \tag{4.9}$$

Combined, (4.8) and (4.9) mean that, in addition to the never treated (NT) group, we can use treatment cohorts $h \in \{r+1,\ldots,T\}$ in estimating $E[\dot{Y}_{rg}(\infty)|D_\infty = 1,\mathbf{X}]$. Incidentally, this derivation shows that if we only use the NT group as the control for each (g,r) pair then we can drop the conditioning on $\mathbf{X}$ in Assumption CNAS. Later we discuss what can be identified in period T without a NT group (under CNAS).

We have established the following. For estimating $E[\dot{Y}_{rg}(\infty)|D_g = 1,\mathbf{X}]$ for $r \in \{g,g+1,\ldots,T\}$ we can use cohorts $\{r+1,\ldots,T,\infty\}$ as the control group. Define the indicator for the control group as

$$A_{r+1} \equiv D_{r+1} + D_{r+2} + \cdots + D_T + D_\infty \tag{4.10}$$

Then, within the subpopulation $D_g + A_{r+1} = 1$, D_g is unconfounded with respect to $\dot{Y}_{rg}(\infty)$, conditional on $\mathbf{X}$. Therefore, we can apply standard treatment effect estimators after transforming the observed outcome and conditioning on the subpopulation.

Naturally, we will need an overlap assumption in order to ensure identification when using methods that condition on covariates. For τ_{gr} and using all legitimate control groups under CNAS and CPTS, the overlap assumption is

Assumption OVLS (**Overlap**, **Staggered Case**): For cohorts $g \in \{S,S+1,\ldots,T\}$ and time periods $r \in \{g,g+1,\ldots,T\}$,

$$P(D_g = 1 | D_g + A_{r+1} = 1, \mathbf{X} = \mathbf{x}) < 1 \text{ for all } \mathbf{x} \in \text{Supp}(\mathbf{X}). \square \tag{4.11}$$

This condition ensures that, within the subpopulation of cohort *g* plus the never treated and not-yet-treated units at time *r*, every treated unit has a comparable control unit.

Given data on units again indexed by *i*, the following simple steps lead to a general analysis. Assumptions CNAS, CPTS, and the overlap assumption are in force.

Procedure 4.1 (**Rolling Methods**, **Staggered Interventions**):

1. For a given $g \in \{S, \ldots, T\}$ and time period $r \in \{g, g+1, \ldots, T\}$, compute

$$\dot{Y}_{irg} \equiv Y_{ir} - \frac{1}{(g-1)} \sum_{s=1}^{g-1} Y_{is} \equiv Y_{ir} - \bar{Y}_{i,pre(g)} \tag{4.12}$$

2. Choose as the control group the units with $D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1$ (or, if desired, a subset, such as only the NT group).

3. Using the subset of data with

$$D_{ig} + D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1, \tag{4.13}$$

apply standard TE methods – such as linear RA, IPW, IPWRA, matching – to the cross section

$$\{(\dot{Y}_{irg}, D_{ig}, \mathbf{X}_i) : i = 1, \ldots, N\},$$

with D_{ig} acting as the treatment indicator. □

Interestingly, when $r = g$, so that τ_{gg} is the instantaneous effect of the intervention for treatment cohort *g*, applying linear RA to

$$\{(\dot{Y}_{igg}, D_{ig}, \mathbf{X}_i) : i = 1, \ldots, N\},$$

with all possible control units, reproduces the POLS estimator proposed by Wooldridge (2021); verification requires a modification of the proof of Theorem 3.1 in the appendix. When $r > g$ this is not the case, which means, under standard assumptions, the rolling approach we propose is inefficient for the dynamic effects. The tradeoff is that we are able to apply many different kind of estimators, including doubly robust and matching estimators.

Procedure 4.1 can be compared with the Callaway and Sant'Anna (2021) approach with staggered interventions. CS (2021) suggest using a long difference of the form

$$\mathring{Y}_{irg} \equiv Y_{ir} - Y_{i,g-1} \tag{4.14}$$

and then choosing control groups from cohorts $\{r+1,\ldots,T,\infty\}$. The transformation in (4.14) is inefficient because it ignores the control periods prior to $g-1$. Also, the default implementation in commonly used software (R and Stata) is to use only the never treated group as controls.

Implementing Procedure 4.1 is straightforward because it simply requires obtaining $\dot{Y}_{irg}$ and then applying standard treatment effect software. Standard errors are easily obtained.

4.2. All Units Eventually Treated

As in Wooldridge (2021, 2023), we can handle situations where all units are treated by $t = T$ by simply modifying the parallel trends assumptions and changing the specifics of the estimation. Rather than stating the CPT assumption in terms of the NT state, $Y_t(\infty)$, it is stated in terms of $Y_t(T)$, the state of not being treated until the final time period. Now all of the treatment effects are, initially, defined relative to $Y_t(T)$: $E[Y_r(g) - Y_r(T)|D_g = 1]$, $g \in \{S,\ldots,T-1\}$, $r \in \{g,\ldots,T\}$. We can no longer estimate a treatment effect for the final treated cohort because there are no control units. As discussed in Wooldridge (2021), by no anticipation it follows that for $r < T$, $E[Y_r(g) - Y_r(T)|D_g = 1] = E[Y_r(g) - Y_r(\infty)|D_g = 1]$ and so, except for the final time period, the ATTs are interpreted as when we have a never treated group.

In terms of estimation, the modifications to Procedure 4.1 are straightforward. In particular, $\dot{Y}_{irg}$ is computed as in (4.12) but only for $g \in \{S,\ldots,T-1\}$. In steps (2) and (3), we drop $D_{i\infty}$ everywhere – which means still choosing as the control group for cohort g in period r those units not yet treated. Then, for each $g \in \{S,\ldots,T-1\}$ and for each period

$r \in \{g, \ldots, T\}$, we apply standard treatment effect estimators, as before. When $r = T$, $D_T = 1$ acts as the only control group for all cohorts first treated prior to period T.

5. Heterogeneous Trends

One way to test for violation of the PT assumption is to estimate placebo treatment effects prior to the intervention. Callaway and Sant'Anna (2021) take this approach using their differencing method. Here, we can apply Procedure 3.1 or 4.1 to pre-treatment periods and test for effects prior to the intervention. For cohort g, it makes sense to split pre-treatment periods, $\{1, 2, \ldots, g-1\}$, into roughly equal sizes. Then, the never treated group, or any of the groups not yet treated in $\{1, 2, \ldots, g-1\}$ can be used as the controls. Under the null hypothesis of (conditional) PT, the tests should not find a "treatment" effect.

As motivation for adjusting Procedure 4.1 (with common timing being a special case) to allow for heterogeneous trends, express the conditional PT assumption as

$$E[Y_t(\infty)|\mathbf{D}, \mathbf{X}] = q_\infty(\mathbf{X}) + \sum_{g=S}^{T} D_g q_g(\mathbf{X}) + m_t(\mathbf{X}),\ t = 1, \ldots, T, \tag{5.1}$$

where $q_g(\cdot)$ and $m_t(\cdot)$ are functions of the covariates, with the first not changing across time and the second not depending on the treatment cohort. As a normalization, $m_1(\mathbf{x}) \equiv 0$ for all $\mathbf{x} \in \text{Supp}(\mathbf{X})$. It is easily seen that

$$E[Y_t(\infty) - Y_1(\infty)|\mathbf{D}, \mathbf{X}] = m_t(\mathbf{X}),\ t = 2, \ldots, T,$$

which does not depend on $\mathbf{D}$. Moreover, for $r \geq g$, it follows from the definition of $\dot{Y}_{rg}(\infty)$ that

$$E[\dot{Y}_{rg}(\infty)|D_\infty = 1, \mathbf{X}] = m_r(\mathbf{X}) - \frac{1}{(g-1)} \sum_{s=1}^{g-1} m_s(\mathbf{X}) \equiv \dot{m}_{rg}(\mathbf{X}) \tag{5.2}$$

and so $\dot{m}_{rg}(\cdot)$ is the conditional mean function implicit in the methods from Section 4 that use a conditional mean specification (RA IPW or IPWRA). Because $\dot{m}_{rg}(\cdot)$ can take on positive and negative values, we essentially assumed in regression-based methods that $\dot{m}_{rg}(\cdot)$ can be approximated by a function linear in parameters (allowing controls to appear flexibly, as usual).

The representation in (5.1) suggests a way to relax parallel trends for cohorts where we have at least two pre-treatment time periods. We replace (5.1) with the following assumption.

Assumption CHT (**Conditional Heterogeneous Trends**): For $\mathbf{D} = (D_S, \ldots, D_T)$ and $t = 1, 2, \ldots, T$,

$$E[Y_t(\infty)|\mathbf{D}, \mathbf{X}] = \eta_S(D_S \cdot t) + \cdots + \eta_T(D_T \cdot t) + q_\infty(\mathbf{X}) + \sum_{g=S}^{T} D_g q_g(\mathbf{X}) + m_t(\mathbf{X}). \quad \square \tag{5.3}$$

Assumption CHT allows a separate linear trend in the never treated state for each treatment cohort. It is easy to see that

$$E[Y_t(\infty) - Y_{t-1}(\infty)|\mathbf{D}, \mathbf{X}] = \eta_g D_g + \cdots + \eta_T D_T + [m_t(\mathbf{X}) - m_{t-1}(\mathbf{X})] \tag{5.4}$$

and so PT, even conditional on $\mathbf{X}$, fails. Because the trend in the never treated state is systematically related to cohort, the estimation approaches in Sections 3 and 4 are no longer valid.

Instead, we can use a linearly detrending, unit by unit, to remove the relationship between $Y_t(\infty)$ and cohort assignment. For any i, we can write

$$\begin{aligned} Y_{it}(\infty) &= \mathbf{h}(\mathbf{D}_i, \mathbf{X}_i) + \mathbf{D}_i \cdot t \cdot \boldsymbol{\eta} + m_t(\mathbf{X}_i) + U_{it}(\infty) \\ E[U_{it}(\infty)|\mathbf{D}_i, \mathbf{X}_i] &= 0, \end{aligned} \tag{5.5}$$

where $\mathbf{h}(\mathbf{D}_i, \mathbf{X}_i)$ does not vary across t. For any $t \geq 2$, we can eliminate both $\mathbf{h}(\mathbf{D}_i, \mathbf{X}_i) + \mathbf{D}_i \cdot t \cdot \boldsymbol{\eta}$ using unit-specific linear detrending.

Equation (5.5) is an example of a heterogenous (or random) trend model of the kind discussed in Wooldridge (2010, Section 11.7.2). Now, for a cohort g, where we require $g \geq 3$ so there are at least two pre-treatment periods. Define the $(g-1) \times 2$ matrix

$$\mathbf{J}_{g-1} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & g-1 \end{pmatrix} \tag{5.6}$$

and let $\mathbf{Y}_{i,g-1}(\infty)$ be the $(g-1) \times 1$ vector

$$\mathbf{Y}_{i,g-1}(\infty) = [Y_{i1}(\infty), \ldots, Y_{i,g-1}(\infty)]'; \tag{5.7}$$

A similar definition holds for $\mathbf{U}_{i,g-1}(\infty)$. Also, let $\mathbf{M}_{i,g-1}$ be the $(g-1) \times 1$ vector with elements $m_t(\mathbf{X}_i)$, $t = 1, \ldots, g-1$. Note that we can write

$$\begin{aligned} \mathbf{Y}_{i,g-1}(\infty) &= \mathbf{J}_{g-1}\begin{pmatrix} \mathbf{h}(\mathbf{D}_i, \mathbf{X}_i) \\ \mathbf{D}_i\boldsymbol{\eta} \end{pmatrix} + \mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty) \\ &\equiv \mathbf{J}_{g-1}\mathbf{Q}_i + \mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty) \end{aligned} \tag{5.8}$$

Now regress $\mathbf{Y}_{i,g-1}(\infty)$ on $\mathbf{J}_{g-1}$, and obtain the coefficients,

$$\hat{\mathbf{B}}_{i,g-1} = (\mathbf{J}_{g-1}'\mathbf{J}_{g-1})^{-1}\mathbf{J}_{g-1}'\mathbf{Y}_{i,g-1}(\infty) = \mathbf{Q}_i + (\mathbf{J}_{g-1}'\mathbf{J}_{g-1})^{-1}\mathbf{J}_{g-1}'[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)] \tag{5.9}$$

For $r \geq g$, the detrend $Y_{ir}(\infty)$ using the unit-specific linear trend up through period $g-1$ to predict $Y_{ir}(\infty)$:

$$\hat{Y}_{irg}(\infty) \equiv (1, r)\hat{\mathbf{B}}_{i,g-1} \tag{5.10}$$

Use these predicted values to detrend $Y_{ir}(\infty)$:

$$\begin{aligned} \dot{Y}_{irg}(\infty) \equiv Y_{ir}(\infty) - \hat{Y}_{irg}(\infty) &= Y_{ir}(\infty) - (1, r)\hat{\mathbf{B}}_{i,g-1} = (1, r)\mathbf{Q}_i + m_r(\mathbf{X}_i) + U_{ir}(\infty) \\ &\quad - (1, r)\left\{\mathbf{Q}_i + (\mathbf{J}_{g-1}'\mathbf{J}_{g-1})^{-1}\mathbf{J}_{g-1}'[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)]\right\} \\ &= m_r(\mathbf{X}_i) + U_{ir}(\infty) - (1, r)\left\{(\mathbf{J}_{g-1}'\mathbf{J}_{g-1})^{-1}\mathbf{J}_{g-1}'[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)]\right\} \end{aligned} \tag{5.11}$$

The expression in (5.9) shows that $\dot{Y}_{ir}(\infty)$ does not depend on $\mathbf{D}_i$. In particular,

$$E[\dot{Y}_{irg}(\infty)|\mathbf{D}_i, \mathbf{X}_i] = E[\dot{Y}_{irg}(\infty)|\mathbf{X}_i] = m_r(\mathbf{X}_i) - (1, r)(\mathbf{J}_{g-1}'\mathbf{J}_{g-1})^{-1}\mathbf{J}_{g-1}'\mathbf{M}_{i,g-1} \tag{5.12}$$

This conclusion is practically important because it shows that the treatment cohort indicators, $\mathbf{D}_i$, are unconfounded with respect to the detrended variable $\dot{Y}_{ir}(\infty)$, conditional on $\mathbf{X}_i$. Note how this extends the argument in Section 4: instead of $\mathbf{J}_{g-1}$ having rows $(1, t)$, its rows simply consisted of unity.

Now the modification to the arguments in Section 4 are now straightforward. In place of (4.3) we have

$$Y_{ir}(g) - Y_{ir}(\infty) = \dot{Y}_{irg}(g) - \dot{Y}_{irg}(\infty) + \left[\hat{Y}_{irg}(g) - \hat{Y}_{irg}(\infty)\right], \tag{5.13}$$

where $\dot{Y}_{irg}(g) \equiv Y_{ir}(g) - \hat{Y}_{irg}(g)$ and $\hat{Y}_{irg}(g)$ are the predicted values from (5.11) and (5.10) but with $\mathbf{Y}_{i,g-1}(g)$ and $Y_{ir}(g)$ in place of $\mathbf{Y}_{i,g-1}(\infty)$ and $Y_{ir}(\infty)$. Now take the expectation conditional on $D_g = 1$:

$$\begin{aligned} \tau_{gr} = E[Y_{ir}(g) - Y_{ir}(\infty)|D_{ig} = 1] &= E[\dot{Y}_{irg}(g) - \dot{Y}_{irg}(\infty)|D_{ig} = 1] \\ &+ E\left[\hat{Y}_{irg}(g) - \hat{Y}_{irg}(\infty)|D_{ig} = 1\right] \end{aligned} \tag{5.14}$$

The second term in (5.14) is zero by no anticipation because $\hat{Y}_{irg}(g)$ and $\hat{Y}_{irg}(\infty)$ are the same linear functions of the potential outcomes in periods $\{1, 2, \ldots, g-1\}$. Therefore,

$$\tau_{gr} = E[\dot{Y}_{irg}(g) - \dot{Y}_{irg}(\infty)|D_{ig} = 1] \tag{5.15}$$

Now the argument is as in Section 4: $\dot{Y}_{irg}(g)$ is observed when $D_{ig} = 1$ and $\mathbf{D}_i$ is unconfounded for $\dot{Y}_{irg}(\infty)$ given $\mathbf{X}_i$. Moreover, by Assumption CNAS, cohorts with $h > r$ (including $h = \infty$) can be used as part of the control group. We have justified the following procedure as producing consistent estimators of τ_{gr} under Assumptions CNAS, CHT, and OVLS.

Procedure 5.1 (Staggered Entry, Heterogeneous Linear Trends):

1. For a cohort $g \in \{S, \ldots, T\}$, run the unit-specific regressions

$$Y_{it} \text{ on } 1, t, \quad t = 1, \ldots, g-1 \tag{5.16}$$

For $r \in \{g, \ldots, T\}$, compute the out-of-sample predicted value $\hat{Y}_{irg}$ and the prediction error (detrended variable) $\dot{Y}_{irg} \equiv Y_{ir} - \hat{Y}_{irg}$.

2. Choose as the control group the units with $D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1$ (or, if desired, a subset, such as only the NT group).

3. Using the subset of data with

$$D_{ig} + D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1, \tag{5.17}$$

apply standard TE methods – such as linear RA, IPW, IPWRA, matching – to the cross section

$$\{(\dot{Y}_{irg}, D_{ig}, \mathbf{X}_i) : i = 1, \ldots, N\},$$

with D_{ig} acting as the treatment indicator. □

Procedure 5.1 is very easy to implement, requiring just many unit-specific simple regressions on a constant and linear time trend. The common timing case is especially easy because the regression in (5.5) is done with $g = S$ only and then the detrended outcomes $\dot{Y}_{ir}$ are used in standard treatment effect estimation for $r = S, \ldots, T$.

In the simplest case where Procedure 5.1 can be applied, with $T = 3$ and common intervention at $S = 3$, and without covariates, the resulting estimator of τ_3 is

$$N_1^{-1} \sum_{i=1}^{N} D_i \dot{Y}_{i3} - N_0^{-1} \sum_{i=1}^{N} (1 - D_i) \dot{Y}_{i3} \tag{5.18}$$

where $\dot{Y}_{i3}$ is obtained as the prediction error in period three after the regression Y_{it} on 1, t, $t = 1, 2$. After a little algebra, (5.18) can be shown to equal

$$[(\bar{Y}_{13} - \bar{Y}_{12}) - (\bar{Y}_{03} - \bar{Y}_{02})] - [(\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})], \tag{5.19}$$

where the first subscript on the average is one for treated and zero for control, and the second subscript indicates time period. The first term in brackets is the usual two-period DiD estimator if the first time period is ignored. The second term is an estimate of the difference in trends prior to the intervention – often interpreted as estimating a placebo effect. The estimator in (5.19) is an example of a difference-in-difference-in-differences estimator. Procedure 5.1 allows one to control for covariates in case removing an estimate of the pre-intervention difference in twins is still not deemed sufficient to uncover a causal effect.

Before ending this section, we head off a potential source of confusion. The fact that we are running unit-specific linear trend regressions in (5.16) does not mean there is an incidental parameters problem that can cause inconsistency in the $\hat{\tau}_{gr}$ when the number of time periods is small. In fact, we are just using these regressions to eliminate unit-specific heterogeneity that can be correlated with treatment cohorts. It is substantively the same as removing the unit-specific pre-treatment means in Procedure 4.1. In fact, this kind of unit-specific detrending is the same idea prevalent in the panel data literature with heterogeneous trends. See, for example, Wooldridge (2010, Section 11.7.2).

6. Violation of No Anticipation. Unbalanced Panels

The no anticipation assumption requires that, prior to the first intervention period for a given treatment cohort, the potential outcomes are the same (on average) as in the never treated state. This assumption can fail if units know that a program or policy change is approaching prior to its being actually implemented. If the NA assumption is in doubt, one can leave one or more periods prior to the intervention time, and redo the analysis as a robustness check.

As an example, suppose a cohort is first treated in $g = 5$. In Procedure 4.1, one would average over periods $\{1, 2, 3, 4\}$ in obtaining the average to remove for Y_{i5} (or, one would remove a unit-specific linear trend, as in Procedure 5.1). Instead, one might drop period four altogether, or maybe even periods three and four. Any precise recommendation is context specific. It is very easy to apply any of the procedures we have recommended to cases where time periods are skipped.

Another issue that often arises in practice is unbalanced panels. With time-constant controls, unbalancedness would typically arise because of missing data on Y_{it}, possibly due to attrition. If data are missing on Y_{it} for some time periods for unit i, the demeaning or detrending is simply applied to the observed data. The mechanics of the procedure are then exactly the same. For treatment cohort g in period r, $\dot{Y}_{ir}$ can only be used if there are enough observed data in the periods $t < g$ to compute an average (one period) or a linear trend (two periods). Of course, Y_{ir} must also be observed.

It is natural to wonder when ignoring the reason the panel is unbalanced does not cause systematic bias. Because our method removes unit-specific averages in Procedure 4.1, selection is allowed to depend on unobserved time-constant heterogeneity – just like with the usual fixed effects estimator. Selection cannot be systematically related to the shocks to $Y_{it}(\infty)$ – again, just as with the FE estimator. When we remove a unit-specific linear trend, now selection can be correlated with both a level heterogeneity term and a trend heterogeneity term, providing for somewhat more robustness to sample selection bias.

7. Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to study the exact properties of our proposed estimators and compare them with competing approaches. We evaluate the performance of five different estimators. The first is the POLS/ETWFE estimator in Wooldridge (2021), which is efficient under a commonly imposed set of assumptions (but is not doubly robust). Three of the estimators use our transformation approach: regression adjustment (RA), inverse-probability-weighted regression adjustment (IPWRA), and propensity score matching (PSM). The final estimator is Callaway Sant'Anna (2021), who apply the augmented IPW (AIPW) estimator – a different doubly robust estimator than IPWRA. We use the never treated group as the control in CS (2021).

7.1. Common Timing

We start with the common timing case. Recall that the POLS method in Wooldridge (2021) and regression adjustment using our rolling method, based on equation (3.8), are the same in the common timing case. Therefore, we have four estimators in the simulations. The CS (2021) transformation is in equation (4.14). We assume that Assumptions NA and CPTC hold, along with overlap. However, we consider scenarios where the functional form of the conditional means and the functional form of the propensity score can be misspecified. For each scenario, we use $T = 6$ with the first treatment in $S = 4$. Across Monte Carlo simulations we draw random samples of sizes 100, 500, and 1,000. For the three ATT parameters, we report Monte Carlos bias, Monte Carlo standard deviation, and the root mean squared error (RMSE). All simulations use 1,000 Monte Carlo replications.

7.1.1. Data Generation

We generate the data as follows. Two control variables are included: $\mathbf{X} = (X_1, X_2)$, where X_1 and X_2 are independent with $X_1 \sim Gamma(2,2)$ [and so $E(X_1) = 4$] and $X_2 \sim Bernoulli(0.6)$. The treatment indicator, D, has propensity score

$$p(\mathbf{x}) = P(D = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\mathbf{Z}_1\boldsymbol{\gamma}_1)}{1 + \exp(\mathbf{Z}_1\boldsymbol{\gamma}_1)} \tag{7.1}$$

where $\mathbf{Z}_1\boldsymbol{\gamma}_1$ (i,e, the propensity score index function) is

$$\mathbf{Z}_1\boldsymbol{\gamma}_1 = -1.2 + \frac{(X_1 - 4)}{2} - X_2 \tag{7.2}$$

Our second step is generating heterogeneous treatment effects as follows:

$$\tau_r(\mathbf{X}) = \theta \cdot \sum_{r=S}^{T}(r - S + 1)^{-1} + \lambda_r h(\mathbf{X}),\ r \in \{S, \ldots, T\}, \tag{7.3}$$

where $\theta = T - S + 1$ and λ_r is a time-varying parameter, set as $(\lambda_S, .., \lambda_T) = (0.5, 0.6, 1.0)$ in each simulation. This setup allows dynamic effects of being treated to vary across time and to increase as the length of exposure to the treatment increases. We consider two different functional forms of $h(\mathbf{X})$ in simulations. The first is

$$h(\mathbf{X}) = \frac{(X_1 - 4)}{2} + \frac{X_2}{3} \tag{7.4}$$

In the second, $h(\mathbf{X})$ includes a quadratic in X_1 and an interaction between X_1 and X_2:

$$h(\mathbf{X}) = \frac{(X_1 - 4)}{2} + \frac{X_2}{3} + \frac{(X_1 - 4)^2}{4} + (X_1 - 4) \cdot \frac{X_2}{2} \tag{7.5}$$

We generate the potential outcomes in the untreated state as

$$Y_t(0) = \delta_t + C + \beta_t \cdot f(\mathbf{X}) + U_t(0), \tag{7.6}$$

where $\delta_t = t$ is a time-specific component, $C|D,X \sim Normal(2,1)$ is an individual-specific component, $U_t(0)|D,\mathbf{X} \sim Normal(0,4)$ is the time-varying shock. The time-varying β_t allows the effect of the covariates on potential outcome paths to vary across time. For each simulation, the parameters are fixed as bellow:

$$\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_T) = (1.0, 1.5, 0.8, 1.5, 2, 2.5) \tag{7.7}$$

We consider two functional forms for $f(\mathbf{X})$:

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2} \tag{7.8}$$

and

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2} + \frac{(X_1 - 4)^2}{3} + (X_1 - 4) \cdot \frac{X_2}{4} \tag{7.9}$$

Finally, the post-treatment period outcome in the treated state is generated as

$$Y_t(1) = \begin{cases} Y_t(0) & ,t < S \\ Y_t(0) + \tau_t + U_t(1) - U_t(0) & ,t \geq S \end{cases}$$

where $U_t(1)|D,\mathbf{X} \sim Normal(0,4)$.

For each simulation, all estimators that involve estimating the conditional means of Y_t assume the correct model is linear in $\mathbf{X}$. Therefore, when (7.4) and (7.8) are used in simulations, the conditional mean is correctly specified, given as Scenarios 1C and 2C below. However, when we consider (7.5) and (7.9), each of which includes quadratic term X_1^2 and an interaction term $X_1 \cdot X_2$, the conditional mean is misspecified (Scenarios 3C and 4C).

We also use a simulation where $(X_1 - 4)^2/2$ is added to the index function in the propensity score, and so the estimated logit model, which assumes an index linear in X_1 and X_2, is misspecified as follows:

$$\mathbf{Z}_2\boldsymbol{\gamma}_2 = -1.2 + \frac{(X_1 - 4)}{2} - X_2 + \frac{(X_1 - 4)^2}{2} \tag{7.10}$$

which is used in Scenarios 2C and 4C.

Table 7.1 describes basic setup for each scenario.

Table 7.1. Scenarios with Common Timing

	Conditional Mean			Propensity Score	
	Correctly Specified?	$h(\mathbf{X})$	$f(\mathbf{X})$	Correctly Specified?	PS Index Function
Scenario 1C	Yes	(7.4)	(7.8)	Yes	(7.3)
Scenario 2C	Yes	(7.4)	(7.8)	No	(7.10)
Scenario 3C	No	(7.5)	(7.9)	Yes	(7.3)
Scenario 4C	No	(7.5)	(7.9)	No	(7.10)

7.1.2. Results for Common Timing Case

This section shows the simulation results under four different scenarios listed in Table 7.1. The results for Scenario 1C are shown in Table 7.2. Here the conditional means and the propensity score are correctly specified, so we expect all estimators to have little bias. This is indeed the case, with the biases being trivial as a percentage of the effect sizes (which are obtained as sample average treatment effects, averaged across the observations and simulations).

The biases are small even when $N = 100$. Because the POLS estimator, which is the same as RA on the transformed outcome, is best linear unbiased, it is not surprising that it produces notably smaller standard deviations compared with PSM and CS (2021). For example, with $N = 500$, and for τ_4, the PSM SD is about 37 percent higher than the POLS/RA SD and the CS SD is about 25 percent higher. Because POLS/RA averages all three pre-treatment periods, its main competitor is the doubly robust IPWRA estimator, which also averages the three pre-treatment periods. When $N = 500$, the rolling IPWRA estimator has SDs that are, at most, three percent higher than those for POLS.

In terms of RMSE, the POLS estimator is uniformly better in Table 7.2 – again, this is not surprising because POLS is the BLUE. When $N \geq 1,000$, the IPWRA estimator has RMSEs that are just slightly larger than POLS. For example, the RMSE of Rolling IPWRA for τ_6 is

0.399, which is slightly higer than that of POLS/RA estimator: 0.379.

Table 7.2. Scenario 1C: $E(Y_t|\mathbf{X} = \mathbf{x})$ and $p(\mathbf{x})$ are Correctly Specified

			τ_4			τ_5			τ_6	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT			3.326			4.800			5.858	
POLS/RA	100	–0.002	1.241	1.241	0.006	1.220	1.220	0.036	1.285	1.285
PSM	100	0.020	1.784	1.784	0.130	1.803	1.807	0.195	1.820	1.831
IPWRA	100	–0.014	1.318	1.318	0.018	1.352	1.352	0.046	1.380	1.381
CS(2021)	100	0.015	1.534	1.534	0.036	1.554	1.554	0.065	1.576	1.577
Sample ATT			3.218			4.809			5.992	
POLS/RA	500	0.008	0.541	0.541	–0.036	0.537	0.538	–0.010	0.552	0.552
PSM	500	0.002	0.893	0.893	0.001	0.931	0.931	0.084	0.939	0.943
IPWRA	500	0.009	0.566	0.566	–0.034	0.562	0.563	–0.009	0.579	0.579
CS(2021)	500	0.011	0.662	0.662	–0.033	0.659	0.660	–0.009	0.684	0.684
Sample ATT			3.220			4.802			5.959	
POLS/RA	1,000	0.006	0.375	0.375	0.009	0.382	0.382	0.020	0.378	0.379
PSM	1,000	0.023	0.710	0.710	0.055	0.673	0.676	0.101	0.679	0.686
IPWRA	1,000	0.007	0.395	0.395	0.007	0.411	0.411	0.021	0.398	0.399
CS(2021)	1,000	–0.008	0.474	0.474	–0.006	0.486	0.486	0.007	0.476	0.476

Notes: (i) The population R-squared values are about 0.39, 0.36, and 0.36, respectively.
(ii) The average propensity score is about 0.26.

In Table 7.3, the propensity score is misspecified but the conditional means are correctly specified. Of the four estimation methods, only PSM should show systematic bias, as the other estimators are consistent whenever the mean functions are correctly specified. In some scenarios and for some of the ATTs, the PSM estimator does not show much bias, especially as a percentage of the effect sizes. Nevertheless, PSM is substantially more biased than the other procedures in some of the runs. The clear winner is again POLS, which has essentially no bias even when $N = 100$ and easily has the smallest SDs. IPWRA (applied, as always, to the transformed outcome) still performs better than PSM and CS (2021) but, evidently, using a misspecified PS increases its imprecision relative to just using POLS.

Table 7.3. **Scenario 2C**: $E(Y_t|\mathbf{X} = \mathbf{x})$ Correctly Specified, $p(\mathbf{x})$ Misspecified.

			τ_4			τ_5			τ_6	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT			3.326			4.800			5.858	
POLS/RA	100	−0.002	1.241	1.241	0.006	1.220	1.220	0.036	1.285	1.285
PSM	100	0.020	1.784	1.784	0.130	1.803	1.807	0.195	1.820	1.831
IPWRA	100	−0.014	1.318	1.318	0.018	1.352	1.352	0.046	1.380	1.381
CS(2021)	100	0.015	1.534	1.534	0.036	1.554	1.554	0.065	1.576	1.577
Sample ATT			3.218			4.809			5.992	
POLS/RA	500	0.008	0.541	0.541	−0.036	0.537	0.538	−0.010	0.552	0.552
PSM	500	0.002	0.893	0.893	0.001	0.931	0.931	0.084	0.939	0.943
IPWRA	500	0.009	0.566	0.566	−0.034	0.562	0.563	−0.009	0.579	0.579
CS(2021)	500	0.011	0.662	0.662	−0.033	0.659	0.660	−0.009	0.684	0.684
Sample ATT			3.220			4.802			5.959	
POLS/RA	1,000	0.006	0.375	0.375	0.009	0.382	0.382	0.020	0.378	0.379
PSM	1,000	0.023	0.710	0.710	0.055	0.673	0.676	0.101	0.679	0.686
IPWRA	1,000	0.007	0.395	0.395	0.007	0.411	0.411	0.021	0.398	0.399
CS(2021)	1,000	−0.008	0.474	0.474	−0.006	0.486	0.486	0.007	0.476	0.476

Notes: (i) The population R-squared values are about 0.39, 0.36, and 0.36, respectively.
(ii) The average propensity score is about 0.26.

Table 7.4 reports the findings for Scenario 3C, where now the conditional means are misspecified because the linear regressions omits the terms X_1^2 and $X_1 \cdot X_2$. Because the propensity score is correctly specified in this scenario, POLS is the only estimator that, theoretically, will exhibit systematic bias. The doubly robust IPWRA and CS (2021) approaches are still consistent, as is PSM. In these simulations, the POLS estimator does have the most bias, although it is fairly small as a fraction of the size effects. For instance, for $N = 1,000$, the bias of POLS/RA estimator for τ_6 is 0.154, which is at least five times larger (in absolute value) than those of the PSM, IPWRA, and CS(2021) estimators: 0.032, -0.002, and -0.008, respectively. Nevertheless, 0.154 is still small as a percentage of the effect size, 6.361.

In some cases, the smaller SD of POLS gives it the smallest RMSE even when it is biased. Among the consistent estimators, IPWRA is the most efficient. And, in several cases, the IPWRA estimator has the smallest RMSE. For example, the RMSEs for $\hat{\tau}_6$ with

$N = 1,000$ are 0.477, 0.617, 0.448, and 0.559 for POLS, PSM, IPWRA, and CS(2021), respectively. Our application of IPWRA to the transformed outcome that uses all pre-treatment periods not only reduces bias compared with POLS, but it largely preserves the efficiency of the POLS estimator.

Table 7.4. Scenario 3C: $E(Y_t|\mathbf{X} = \mathbf{x})$ Misspecified, $p(\mathbf{x})$ Correctly Specified.

			τ_4			τ_5			τ_6	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT			3.550			4.975			6.295	
POLS/RA	100	−0.034	1.412	1.413	0.104	1.406	1.410	0.222	1.568	1.583
PSM	100	−0.060	1.797	1.798	0.071	1.867	1.868	0.054	1.966	1.967
IPWRA	100	−0.099	1.431	1.434	0.025	1.433	1.433	0.071	1.561	1.563
CS(2021)	100	−0.100	1.751	1.753	0.009	1.734	1.734	0.053	1.863	1.864
Sample ATT			3.418			5.053			6.356	
POLS/RA	500	0.079	0.608	0.614	0.085	0.613	0.619	0.197	0.670	0.699
PSM	500	0.032	0.827	0.828	−0.015	0.863	0.863	0.081	0.878	0.882
IPWRA	500	0.034	0.618	0.619	−0.013	0.619	0.619	0.047	0.665	0.666
CS(2021)	500	0.055	0.764	0.766	0.006	0.763	0.763	0.062	0.830	0.833
Sample ATT			3.440			5.017			6.361	
POLS/RA	1,000	0.044	0.402	0.404	0.091	0.426	0.436	0.154	0.452	0.477
PSM	1,000	0.031	0.569	0.570	0.017	0.576	0.577	0.032	0.616	0.617
IPWRA	1,000	0.000	0.408	0.408	−0.011	0.423	0.423	−0.002	0.448	0.448
CS(2021)	1,000	−0.003	0.513	0.513	−0.015	0.523	0.523	−0.008	0.559	0.559

Notes: (i) The population R-squared values are about 0.41, 0.38, and 0.38, respectively.
(ii) The average propensity score is about 0.17.

The final simulation, based on Scenario 4C, has both the means and propensity scores misspecified. The findings are in Table 7.5. The IPWRA estimator is always less biased than POLS and has a lower RMSE. Compared with rolling IPWRA, neither PSM nor CS (2021) is competitive, as they either have more bias, larger SDs, or both. (In some cases PSM has slightly less bias.)

Table 7.5. **Scenario 4C**: $E(Y_t|\mathbf{X} = \mathbf{x})$ and $p(\mathbf{x})$ are Misspecified.

			τ_4			τ_5			τ_6	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT			3.655			5.194			6.516	
POLS/RA	100	0.258	1.269	1.295	0.593	1.279	1.410	0.949	1.421	1.709
PSM	100	0.210	1.797	1.809	0.560	1.856	1.939	0.863	1.928	2.113
IPWRA	100	0.193	1.334	1.348	0.485	1.394	1.476	0.774	1.476	1.667
CS(2021)	100	0.428	1.562	1.620	0.745	1.613	1.777	1.068	1.705	2.012
Sample ATT			3.546			5.203			6.649	
POLS/RA	500	0.266	0.555	0.615	0.546	0.573	0.791	0.895	0.617	1.087
PSM	500	0.177	0.901	0.918	0.395	0.943	1.022	0.696	0.980	1.202
IPWRA	500	0.211	0.575	0.613	0.422	0.585	0.721	0.699	0.621	0.935
CS(2021)	500	0.420	0.679	0.798	0.669	0.694	0.964	0.986	0.741	1.234
Sample ATT			3.549			5.197			6.617	
POLS/RA	1,000	0.264	0.385	0.467	0.592	0.399	0.714	0.926	0.433	1.022
PSM	1,000	0.194	0.716	0.742	0.444	0.687	0.818	0.706	0.714	1.004
IPWRA	1,000	0.210	0.404	0.455	0.466	0.421	0.628	0.735	0.438	0.856
CS(2021)	1,000	0.404	0.488	0.633	0.701	0.505	0.864	1.007	0.528	1.137

Notes: (i) The population R-squared values are about 0.48, 0.46, and 0.46, respectively.
(ii) The average propensity score is about 0.26.

7.2. Staggered Intervention

In this subsection, we consider the staggered intervention case. For simulations, we follow Procedure 4.1 by using the transformation in (4.12). In applying RA, PSM, and IPWRA, we use the never treated and not-yet treated units as the control group. Similar to common timing cases, we consider four different scenarios where the conditional mean and propensity score are either correctly specified or misspecified given Assumptions CNAS, CPTS, and OVLS hold. For each scenario, we use $T = 6$ and $g \in G = \{4, 5, 6, \infty\}$ and 1,000 replications. For treated cohort g, initial treatment occurs at time $t = g$; $g = \infty$ indicates the never treated cohort.

7.2.1. Data Generation

We generate the data for each simulation following similar steps in the common timing case. In particular, (X_1, X_2) are generated in the same way, with X_1 continuous and nonnegative and X_2 binary. The cohort probabilities, $P(D_g = 1|\mathbf{X})$, are generated using an ordered logit specification. This means that the logit propensity score is never correctly specified because, conditional on using a subset of the units as controls, the response probability is different from logit. Still, in one case we use an index linear in **X** as in (7.3) and in another the index is nonlinear, as in (7.10); both are as described earlier. Although the language is not precise, we refer to the first case as "correct specification" of the PS because the index is correctly specified and the second as "misspecification" because the nonlinear terms in the index are are ignored.

The treatment effects for treated cohort g in period $r \geq g$, conditional on **X**, are given by

$$\tau_{gr}(\mathbf{X}) = \theta_g \cdot \sum_{r=S}^{T}\left(\frac{1}{2}(r-S)+1\right)^{-1} + \lambda_{gr}h_g(\mathbf{X});\ g \in \{4,5,6\},\ r \in \{g, g+1, \ldots, T\}, \tag{7.11}$$

where $\{\theta_4, \theta_5, \theta_6\} = \{4,3,2\}$ and λ_{gr} is time-varying parameter, each of which varies across treated cohort g at $r \in \{4,5,6\}$ as follows:

$$\begin{aligned}(\lambda_{44}, \lambda_{54}, \lambda_{64}) &= (0.5, 0.6, 1)\\ (\lambda_{45}, \lambda_{55}, \lambda_{65}) &= (0, 1, 0.5)\\ (\lambda_{46}, \lambda_{56}, \lambda_{66}) &= (0, 0, 0.5)\end{aligned}$$

When the mean is correctly specified (and linear in X_1 and X_2), the functions $h_g(\cdot)$ are just as in (7.4) for all $g \in \{4,5,6\}$. When we add neglected nonlinearity,

$$h_g(\mathbf{X}) = \frac{g \cdot (X_1 - 4)^2}{3} + \frac{X_2(X_1 - 4)}{3},\ g \in \{4,5,6\} \tag{7.12}$$

We generate the outcome in the never treated state as

$$Y_{it}(\infty) = \delta_t + C_i + \beta_t \cdot f(\mathbf{X}_i) + U_{it}(\infty), \tag{7.13}$$

where $\delta_t = t$ is a time-specific component, $C_i|\mathbf{D}_i, \mathbf{X}_i \sim Normal(2,1)$ is an individual-specific component, $U_{it}(\infty)|\mathbf{D}_i, \mathbf{X}_i \sim Normal(0,4)$ is the time-varying shock, and the β_t are fixed at the same values in (7.7). The function $f(\mathbf{X})$ is initially taken to be linear, as in (7.8). We introduce neglected nonlinearity as in (7.9).

Finally, the post-treatment period outcome in the treated state is generated as

$$Y_{it}(g) = \begin{cases} Y_t(\infty) & ,t < S \\ Y_t(\infty) + \tau_{gt}(\mathbf{X}) + U_{it}(g) - U_t(\infty) & ,t \geq S \end{cases} \tag{7.14}$$

where $U_{it}(g)|\mathbf{D}_i, \mathbf{X}_i \sim Normal(0,4)$ and independent across g and t.

The different scenarios are summarized in Table 7.6.

Table 7.6. Scenarios with Staggered Intervention

	Conditional Mean			Propensity Score	
	Correctly Specified?	$h_g(\mathbf{X})$	$f(\mathbf{X})$	Correctly Specified?	PS Index Function
Scenario 1S	Yes	(7.4)	(7.8)	Yes	(7.3)
Scenario 2S	Yes	(7.4)	(7.8)	No	(7.10)
Scenario 3S	No	(7.12)	(7.9)	Yes	(7.3)
Scenario 4S	No	(7.12)	(7.9)	No	(7.10)

7.2.2. Results for Staggered Case

For each simulation, we use 1,000 Monte Carlo replications. We obtain the SATTs for cohorts $g \in \{4,5,6\}$ at post-treatment periods $r \in \{g, g+1, \ldots, 6\}$. As before we report the bias, standard deviations, and RMSEs of each estimator. Recall that the POLS method in Wooldridge (2021) and the RA method applied to $\dot{Y}_{irg}$ are the same only when $r = g$; therefore, these estimators are reported separately.

The findings for Scenario 1S are given in Table 7.7. Because all estimators are consistent, it is not surprising that each is essentially unbiased. The same general pattern for precision that we found in the common timing case holds in the staggered case: POLS and RA are the most efficient, IPWRA is a bit less precise, and PSM and CS (2021) have notably larger SDs.

Table 7.8 contains the simulation results for Scenario 2S, where the propensity score neglects the nonlinear terms. Even PSM shows little relative bias, although it is more than the other estimators. POLS, RA, and IPWRA again perform best in terms of precision, with IPWRA doing a bit worse for some of the cohort/year combinations.

Table 7.7. Scenario 1S: $E(Y_t|\mathbf{X} = \mathbf{x})$ and $p_g(\mathbf{x})$ are Correctly Specified.

			τ_{44}			τ_{45}			τ_{46}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		4.168			6.207			8.338	
POLS	1,000	0.006	0.466	0.466	–0.007	0.472	0.472	–0.012	0.488	0.488
RA	1,000	0.006	0.466	0.466	–0.007	0.472	0.472	–0.012	0.491	0.491
PSM	1,000	0.008	0.640	0.640	0.005	0.663	0.663	–0.005	0.676	0.676
IPWRA	1,000	0.006	0.468	0.468	–0.007	0.473	0.473	–0.011	0.495	0.496
CS (2021)	1,000	0.001	0.595	0.595	–0.020	0.589	0.589	–0.020	0.676	0.676

			τ_{55}			τ_{56}			τ_{66}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		3.390			4.689			2.172	
POLS	1,000	–0.031	0.452	0.453	–0.033	0.472	0.473	–0.019	0.461	0.461
RA	1,000	–0.031	0.452	0.453	–0.034	0.473	0.474	–0.019	0.461	0.461
PSM	1,000	–0.021	0.641	0.641	–0.022	0.645	0.646	–0.020	0.648	0.648
IPWRA	1,000	–0.032	0.455	0.456	–0.034	0.482	0.483	–0.019	0.468	0.468
CS (2021)	1,000	–0.037	0.600	0.601	–0.036	0.620	0.622	–0.009	0.615	0.615

Notes: (i) The population *R*-squared value from the POLS regression is 0.36.

(ii) The cohort shares are 0.66 , 0.12, 0.11, and 0.11.

Table 7.8. Scenario 2S:$E(Y_t|\mathbf{X} = \mathbf{x})$ CorrectlySpecified, $p_g(\mathbf{x})$ Misspecified

			τ_{44}			τ_{45}			τ_{46}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		4.181			6.202			8.336	
POLS	1,000	–0.002	0.401	0.401	0.004	0.422	0.422	–0.001	0.461	0.461
RA	1,000	–0.002	0.401	0.401	0.003	0.424	0.424	–0.003	0.470	0.470
PSM	1,000	0.010	0.566	0.566	0.025	0.588	0.589	0.062	0.718	0.721
IPWRA	1,000	–0.002	0.402	0.402	0.006	0.427	0.427	–0.002	0.480	0.480
CS (2021)	1,000	0.016	0.565	0.565	0.009	0.565	0.565	0.013	0.718	0.718

			τ_{55}			τ_{56}			τ_{66}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		3.334			4.685			2.168	
POLS	1,000	–0.023	0.408	0.409	–0.029	0.455	0.456	–0.015	0.437	0.437
RA	1,000	–0.023	0.408	0.409	–0.032	0.458	0.459	–0.015	0.437	0.437
PSM	1,000	–0.008	0.580	0.580	0.057	0.685	0.687	0.036	0.680	0.681
IPWRA	1,000	–0.021	0.409	0.409	–0.030	0.463	0.464	–0.014	0.447	0.448
CS (2021)	1,000	–0.029	0.556	0.556	–0.026	0.562	0.563	0.003	0.570	0.570

Notes: (i) The population *R*-squared from the POLS regression is about 0.40.

(ii) The cohort shares are about 0.55, 0.15, 0.15, and 0.15.

The results for Scenario 3S are given in Table 7.9. When the coditional mean is misspecified whereas the propensity is correctly specified, our rolling method with IPWRA has the lowest RMSE among different estimators.

Table 7.9. **Scenario 3S**: $E(Y_t|\mathbf{X} = \mathbf{x})$ Misspecified, $p_g(\mathbf{x})$ Correctly Specified

			τ_{44}			τ_{45}			τ_{46}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		5.113			7.341			10.227	
POLS	1,000	0.014	0.522	0.523	0.030	0.544	0.545	0.103	0.680	0.688
RA	1,000	0.014	0.522	0.523	0.033	0.545	0.546	0.118	0.682	0.692
PSM	1,000	0.011	0.679	0.679	0.015	0.709	0.710	0.019	0.807	0.807
IPWRA	1,000	−0.009	0.521	0.521	−0.036	0.536	0.537	−0.012	0.667	0.667
CS (2021)	1,000	−0.003	0.662	0.662	−0.026	0.671	0.671	−0.029	0.807	0.808

			τ_{55}			τ_{56}			τ_{66}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		5.869			5.928			3.678	
POLS	1,000	0.010	0.666	0.667	0.090	0.577	0.584	0.093	0.591	0.598
RA	1,000	0.010	0.666	0.667	0.097	0.579	0.587	0.093	0.591	0.598
PSM	1,000	−0.008	0.783	0.783	0.006	0.725	0.725	0.004	0.747	0.747
IPWRA	1,000	−0.052	0.655	0.657	−0.028	0.574	0.574	−0.013	0.584	0.585
CS (2021)	1,000	−0.037	0.764	0.764	−0.035	0.696	0.697	−0.008	0.674	0.674

Notes: (i) The population R-squared value from the POLS regression is about 0.47.

(ii) The cohort probabilities are about 0.66, 0.12, 0.11 and 0.11.

Finally, Table 7.10 includes the results when the conditional mean and propensity scores have neglected nonlinearties (Scenario 4S). The IPWRA estimator has the smallest RMSE for all ATTs, τ_{gr}, $g = S, \ldots, T; r = g, \ldots, T.$ in these particular simulations.

Table 7.10. **Scenario 4S**: $E(Y_t|\mathbf{X} = \mathbf{x})$ and $p_g(\mathbf{x})$ are Misspecified

			τ_{44}			τ_{45}			τ_{46}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		5.376			7.636			10.725	
POLS	1,000	0.074	0.442	0.448	0.234	0.490	0.543	0.596	0.641	0.875
RA	1,000	0.074	0.442	0.448	0.241	0.492	0.548	0.655	0.650	0.923
PSM	1,000	0.054	0.591	0.594	0.160	0.638	0.658	0.535	0.860	1.013
IPWRA	1,000	0.053	0.440	0.443	0.170	0.492	0.521	0.560	0.649	0.857
CS (2021)	1,000	0.323	0.607	0.687	0.535	0.627	0.824	0.759	0.860	1.147
			τ_{55}			τ_{56}			τ_{66}	
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample ATT	1,000		6.390			6.213			4.029	
POLS	1,000	0.188	0.584	0.613	0.547	0.538	0.767	0.520	0.547	0.755
RA	1,000	0.188	0.584	0.613	0.578	0.540	0.791	0.520	0.547	0.755
PSM	1,000	0.113	0.695	0.704	0.498	0.742	0.893	0.425	0.751	0.863
IPWRA	1,000	0.126	0.573	0.587	0.492	0.534	0.726	0.445	0.544	0.703
CS (2021)	1,000	0.190	0.675	0.701	0.412	0.615	0.740	0.223	0.620	0.659

Notes: (i) The population R-squared value from the POLS regression is about 0.51.

(ii) The cohort probabilities are about 0.55, 0.15, 0.15 and 0.15.

8. Concluding Remarks

In this paper we propose an alternative estimation approach in order to obtain consistent estimates of the average treatment effect on the treated in difference-in-differences setting where there are multiple periods and possible staggered adoption. The key advantage of this approach is that once the transformed dependent variable is defined – whether in the common timing case, the staggered case, or when unit-specific trends have been removed in either case – one can apply standard TE estimators to the cross sectional data for a given time period and each cohort treated in that time period. One need only be careful about choosing units not yet treated as the control group (including possibly limiting the controls the the never-treated group). Regression adjustment (RA) using the transformed outcome is one of many methods that one can apply. In the common timing case, we show that the RA estimator is algebraically equivalent to the POLS/ETWFE/RE estimators in Wooldridge (2021). This equivalence implies RA estimator with our proposed transformed data yields consistent estimators of ATT when the outcome model is linear in the chosen covariates and the no anticipation and parallel trends assumptions hold only after conditioning on covariates. If the idiosyncratic errors are serially uncorrelated and the composite error is homoskedasticity, the RA estimator is also efficient.

The Monte Carlo simulation results provide evidence that the doubly robust IPWRA estimator achieves close to the efficiency of RA when the conditional means are correctly specified and often has less bias when the mean is misspecified and the propensity score model is correctly specified or even misspecified. The IPWRA estimator applied to our transformed variable has better precision compared with that of the CS(2021), something that makes intuitive sense because our approach uses all suitable time periods and cohorts in the control group.

References

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

Athey, S., and Imbens, G. W. (2022), “Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption,” *Journal of Econometrics* 226, 62-79.

Borusyak, K., and Jaravel, X. (2018), “Revisiting Event Study Designs,” Availabile at SSRN 2826228.

Borusyak, K., Jaravel, X., and Spiess, J. (2022), “Revisiting Event Study Designs: Robust and Efficient Estimation,” arXiv:2108.12419.

Callaway, B., and P.H. Sant’Anna (2021), “Difference-in-Differences with MultipleTime Periods,” *Journal of Econometrics* 225, 200-230.

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

de Chaisemartin, C., and D’Haultfoeuille, X. (2023), “Two-way Fixed Effects and Difference-in-Differences with HeterogeneousTreatment Effects: A Survey,” forthcoming, *The Econometrics Journal*.

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

Negi, A., and J.M. Wooldridge (2021), “Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects,” *Econometric Reviews* 40, 504-534.

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

Sun, L., and S. Abraham (2021), “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects,” *Journal of Econometrics* 225, 175–199.

Goodman-Bacon, A. (2021), “Difference-in-Differences with Variation in Treatment Timing,” *Journal of Econometrics* 225, 254-277.

Wooldridge, J. M. (2007), “Inverse Probability Weighted Estimation for General Missing Data Problems,” *Journal of Econometrics* 141, 1281-1301.

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

Wooldridge, J. M. (2021), “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” Available at SSRN 3906345.

Wooldridge, J.M. (2023), “Simple Approaches to Nonlinear Difference-in-Differences with Panel Data,” forthcoming, *The Econometrics Journal*.

Proof of Theorem 3.1

We modify the argument in Wooldridge (2021, Theorem 8.1). The $\hat{\tau}_r$ are obtained from regression (3.6). Because $\dot{Y}_{ir} = Y_{ir} - \bar{Y}_{i,pre}$, basic OLS algebra shows that all coefficients from (3.6) are obtained by differencing the coefficients from the two regressions

$$\dot{Y}_{ir} \text{ on } 1,\ D_i,\ \mathbf{X}_i,\ D_i \cdot \dot{\mathbf{X}}_i,\ i = 1, 2, \ldots, N \tag{A.1}$$

$$\bar{Y}_{i,pre} \text{ on } 1,\ D_i,\ \mathbf{X}_i,\ D_i \cdot \dot{\mathbf{X}}_i,\ i = 1, 2, \ldots, N \tag{A.2}$$

where $\dot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}_1$ are the covariates demeaned using the treated units. In particular, letting $\hat{\rho}_r$ be the coefficient on D_i from (A.1) and $\hat{\rho}_{pre}$ the coefficient on D_i from (A.2),

$$\hat{\tau}_r = \hat{\rho}_r - \hat{\rho}_{pre} \tag{A.3}$$

Note also that the coefficients on the "moderating" terms, $D_i \cdot \dot{\mathbf{X}}_i$, are also obtained by differencing across the two regressions.

To show (A.3) is the same as the coefficient on $D_i \cdot fr_t$ in (3.9), first note that, by Wooldridge (2021, Theorem 3.2), we can drop $(fq_t, fq_t \cdot \mathbf{X}_i)$ for $q < S$ without affecting the estimates. In other words, the $\tilde{\tau}_r$ are the coefficients on $D_i \cdot fr_t$ in the regression

$$\begin{gathered} Y_{it} \text{ on } 1,\ \mathbf{X}_i,\ D_i,\ D_i \cdot \mathbf{X}_i, fS_t,\ \ldots, fT_t, fS_t \cdot \mathbf{X}_i,\ \ldots, fT_t \cdot \mathbf{X}_i \\ D_i \cdot fS_t,\ \ldots,\ D_i \cdot fT_t,\ D_i \cdot fS_t \cdot \dot{\mathbf{X}}_i,\ \ldots,\ D_i \cdot fT_t \cdot \dot{\mathbf{X}}_i \end{gathered} \tag{A.4}$$

Now, define $\mathbf{H}_i \equiv (1, \mathbf{X}_i, D_i, D_i \cdot \mathbf{X}_i)$, a $1 \times 2(K+1)$ vector, and

$$\mathbf{L}_{it} \equiv (fS_t, fS_t \cdot \mathbf{X}_i, D_i \cdot fS_t, D_i \cdot fS_t \cdot \dot{\mathbf{X}}_i, \ldots, fT_t, , fT_t \cdot \mathbf{X}_i, D_i \cdot fT_t \cdot \dot{\mathbf{X}}_i), \tag{A.5}$$

a row vector with $2(T - S + 1)(K + 1)$ elements. Note that for $q \neq r$, $(fq_t, fq_t \cdot \mathbf{X}_i, D_i \cdot fq_t, D_i \cdot fq_t \cdot \dot{\mathbf{X}}_i)$ and $(fr_t, fr_t \cdot \mathbf{X}_i, D_i \cdot fr_t, D_i \cdot fr_t \cdot \dot{\mathbf{X}}_i)$ are orthogonal in sample because $fq_t \cdot fr_t = 0$. The full set of regressors in (A.4) is simply $(\mathbf{H}_i, \mathbf{L}_{it})$. With $p_t = fS_t + \cdots + fT_t$, the post-treatment period indicator, $(1 - p_t)fr_t = 0$, $r = S,\ S+1,\ \ldots,\ T$, which means $(1 - p_t)\mathbf{L}_{it} = \mathbf{0}$. Therefore, the objective function underlying the regression in (A.4) can be written as

$$\min_{\boldsymbol{\theta},\boldsymbol{\delta}} \sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)(Y_{it}-\mathbf{H}_i\boldsymbol{\theta})^2+\sum_{i=1}^{N}\sum_{t=1}^{T}p_t(Y_{it}-\mathbf{H}_i\boldsymbol{\theta}-\mathbf{L}_{it}\boldsymbol{\delta})^2 \tag{A.6}$$

Letting $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\delta}}$ denote the POLS estimators, the first order conditions are

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}})+\sum_{i=1}^{N}\sum_{t=1}^{T}p_t\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}}-\mathbf{L}_{it}\tilde{\boldsymbol{\delta}}) = \mathbf{0} \tag{A.7}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T}p_t\mathbf{L}_{it}'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}}-\mathbf{L}_{it}\tilde{\boldsymbol{\delta}}) = \mathbf{0} \tag{A.8}$$

Next, note that because $p_t = fS_t + \cdots + fT_t$, we can write

$$p_t\mathbf{H}_i = [p_t, p_t \cdot D_i, p_t \cdot \mathbf{X}_i, p_t \cdot D_i \cdot \dot{\mathbf{X}}_i] = \sum_{q=S}^{T}[fq_t, fq_t \cdot D_i, fq_t \cdot \mathbf{X}_i, fq_t \cdot D_i \cdot \dot{\mathbf{X}}_i], \tag{A.9}$$

which is simply the sum the subvectors in $\mathbf{L}_{it}$ consisting of different time periods. It follows that $p_t\mathbf{H}_i = p_t\mathbf{L}_{it}\mathbf{A}$ for a $2(T-S+1)(K+1) \times 2(K+1)$ matrix $\mathbf{A}$. Plugging into (A.7) gives

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}})+\mathbf{A}'\sum_{i=1}^{N}\sum_{t=1}^{T}p_t\mathbf{L}_{it}'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}}-\mathbf{L}_{it}\tilde{\boldsymbol{\delta}}) = \mathbf{0} \tag{A.10}$$

Along with (A.8), (A.10) implies that the FOCs for $\left(\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\delta}}\right)$ are

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\tilde{\boldsymbol{\theta}}) = \mathbf{0} \tag{A.11}$$

But (A.11) means that $\tilde{\boldsymbol{\theta}}$ is the OLS estimator from the regression Y_{it} on $\mathbf{H}_i$ using the pre-treatment period observations. With $\mathbf{H}_i$ not varying over time, $\tilde{\boldsymbol{\theta}}$ is the same as the cross-sectional regression

$$\bar{Y}_{i,pre} \text{ on } 1,\, D_i,\, \mathbf{X}_i,\, D_i \cdot \dot{\mathbf{X}}_i \tag{A.12}$$

In particular, the coefficient on D_i is precisely $\hat{\rho}_{pre}$ in (A.3).

Next, the FOC in (A.8) shows that $\tilde{\boldsymbol{\delta}}$ is from a POLS regression using the post-treatment periods:

$$Y_{it} - \mathbf{H}_i\tilde{\boldsymbol{\theta}} \text{ on } \mathbf{L}_{it},\; t = S,\ldots,T;\; i = 1,\ldots,N$$

By definition of $\mathbf{L}_{it}$ and the orthogonality of the elements of $\mathbf{L}_{it}$ across the subvectors representing the different time periods, each $2(K+1)$ subvector, $\tilde{\boldsymbol{\delta}}_r$, $r = S,\ldots,T$, is obtained from a separate cross-sectional regression for each post-treatment period. Namely, because $fr_r = 1$, the regression for period r is

$$Y_{ir} - \mathbf{H}_i\tilde{\boldsymbol{\theta}} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \dot{\mathbf{X}}_i, i = 1,\ldots,N \tag{A.13}$$

The vector on right is simply $\mathbf{H}_i$, and so

$$\tilde{\boldsymbol{\delta}}_r = \left(\sum_{i=1}^{N}\mathbf{H}_i'\mathbf{H}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{H}_i'Y_{ir}\right) - \tilde{\boldsymbol{\theta}} \tag{A.14}$$

The first term is the regression coefficients from

$$Y_{ir} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \dot{\mathbf{X}}_i, i = 1,\ldots,N,$$

which is $\hat{\rho}_r$ from (A.3). We have shown that the coefficient corresponding to $D_i \cdot fr_t$ in the regression (A.4) is $\hat{\rho}_r - \hat{\rho}_{pre}$, which establishes the equivalence of the pooled OLS estimator across all time periods, (3.9), and the cross-sectional OLS estimators using the transformed variable $\dot{Y}_{ir}$ for each $r \in \{S,\ldots,T\}$ separately. Essentially the same argument shows that the coefficients on the interaction terms $D_i \cdot fr_t \cdot \dot{\mathbf{X}}_i$ in (3.9) are the same as the coefficients on $D_i \cdot \dot{\mathbf{X}}_i$ in (3.6). □