A ROLLING METHOD FOR CAUSAL INFERENCE: A FLEXIBLE ESTIMATION METHOD FOR DIVERSE PANEL DATA SETUPS

By

Soo Jeong Lee

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics—Doctor of Philosophy

2025

ABSTRACT

This dissertation develops new methods for estimating treatment effects using panel data, with a focus on flexible and robust identification under treatment effect heterogeneity across units and over time. Chapter 1 provides an overview of the subsequent chapters. Chapter 2 introduces a simple time-series transformation—termed the Rolling Method—for Difference-in-Differences estimation. This method converts panel data into a sequence of cross-sectional datasets through unit-specific outcome transformations, enabling consistent estimation of group-time-specific treatment effects even in the presence of treatment heterogeneity. Chapter 3 extends the Rolling Method to small-sample settings, particularly when the number of treated or control units is limited, and demonstrates improved finite-sample inference properties. Chapter 4 further generalizes the framework to accommodate dynamic treatment paths, allowing for treatment reversals and subgroup-specific moderating effects. This extension is applied in an empirical case study examining the effects of the entry and exit of chain pharmacies in rural areas.

ACKNOWLEDGEMENTS

Above all, I extend my deepest gratitude to my advisor and mentor, Dr. Jeffrey M. Wooldridge. The excitement in his eyes during our research discussions has been a constant source of inspiration. I aspire to follow his example–to be a scholar who speaks with passion, wisdom, and boundless curiosity. Under his guidance, I have learned invaluable lessons that I will carry with me as I mentor future students, sharing the joy of economics and the fulfillment that comes from contributing to the world through research.

I am sincerely grateful to my committee members–Dr. Antonio Galvao, Dr. Kyoo il Kim, and Dr. Haolei Weng– for their invaluable feedback and encouragement, which have been instrumental to my growth throughout this journey. I would also like to extend special thanks to Dr. Timothy Vogelsang and Dr. Hugo Freeman. Being part of the MSU econometrics group has been an extraordinary privilege, and I could not have been happier to be a part of it.

I am deeply grateful to the faculty and friends in the MSU Economics Department. Your presence made the joyful moments even brighter and the challenging times easier to navigate. I also extend my heartfelt thanks to my EPIC Grad Lab friends and my Project Investigator, Dr. Tara Kilbride. A special thank you to Dr. Todd Elder, whose support as DGS made it possible for me to join the program and become part of such an inspiring community.

Finally, I am thankful to my family. To my parents, Sunduk Jeong and Jinrak Lee–the most loving couple I know–thank you for your unwavering support and for always reminding me, "*We believe in you. You always do well.*" To my older sister, Eunjeong, who gifted me my very first flight ticket to the U.S. and took on all family responsibilities so I could devote myself fully to my studies–thank you. And to my younger sister, Minjeong–I could not love you more.

To all who have stood by me through moments of struggle, happiness, and solitude– thank you! My journey as an economist will never cease, thanks to your unwavering support.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This dissertation advances estimation and inference techniques for panel data by developing novel methods for treatment effect estimation. By introducing a time-series outcome transformation, addressing heterogeneous treatment effects, and enhancing inference in small-sample settings, it offers more robust and flexible tools for policy evaluation and empirical research.

Chapter 2 introduces a time-series transformation technique–termed the Rolling Method–that can be combined with various treatment effect estimators, including regression adjustment, matching methods, and doubly robust estimators. Unlike conventional approaches that assume constant treatment effects across units and over time, our method accounts for treatment effect heterogeneity. In the common timing case, we show that applying the transformation with linear regression adjustment numerically reproduces the pooled OLS estimator in Wooldridge (2021), which is the Best Linear Unbiased Estimator (BLUE) under classical assumptions.

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 for each time period. We show 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. The doubly robust method of combining inverse probability weighting with linear regression (IPWRA) works particularly well in terms of bias and efficiency. Importantly, our transformation is easily modified to account for unit-specific trends.

We illustrate the method using empirical data on Walmart's entry into local labor markets. The application demonstrates how the transformation approach yields more reliable estimates of employment effects, particularly in cases where traditional methods may suffer from bias and inefficiency due to pre-existing trends.

Chapter 3 proposes simple methods for valid inference in panel data difference-in-differences (DiD) settings with a small number of treated units or a small number of control units (including a small number of both). The approach uses a suitable transformation to collapse the panel data set into a cross-sectional data set. If the classical linear model assumptions hold in the cross-sectional regression, exact inference is available – even in cases with only two control units and one treated unit. The approach likely works best with a large number of time periods before and after the intervention so that the central limit theorem across time makes normality a better approximation. Nevertheless, under joint normality and homoskedasticity of the time-varying errors, the exact approach can be applied with few time periods – both before and after the intervention – as well as few cross-sectional units. In addition to standard DiD estimation, the approach permits removal of unit-specific trends. With large enough sample sizes, control variables may be included. If the cross-sectional sample size is not too small, one can use a particular heteroskedasticity-robust standard error. We illustrate the approach using increased smoking restrictions in California, where there is only a single treated unit, as an alternative to the synthetic control approach. In the staggered intervention case, we reexamine the expansion of so-called "castle" laws in the United States.

Chapter 4 extends the Rolling Method framework to accommodate complex treatment patterns and moderating effects. I propose a novel aggregation strategy suitable for measuring treatment effects with staggered entry and exit of treatment. Furthermore, I develop a two-stage IPWRA estimator that incorporates moderating effects, allowing for heterogeneous treatment effects across subgroups defined by observed characteristics. This extension enables a more nuanced understanding of how treatment effects evolve over time and vary across populations. To demonstrate its practical utility, I apply the method to evaluate the impact of chain pharmacy entry and exit on pharmacy access in rural areas. The results reveal substantial heterogeneity in treatment effects across three distinct trajectories—initial treatment, exit, and re-entry—that conventional approaches may overlook. This framework offers applied researchers a robust and flexible tool for analyzing dynamic and heterogeneous treatment paths.

CHAPTER 2

A SIMPLE TRANSFORMATION APPROACH TO DIFFERENCE-IN-DIFFERENCES ESTIMATION FOR PANEL DATA

(CO-AUTHORED WITH JEFFREY M. WOOLDRIDGE)

2.1 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 under staggered interventions, including Borusyak and Jaravel (2018), Goodman-Bacon (2021), and De Chaisemartin and d'Haultfoeuille (2020), who propose alternative representations of the simple TWFE estimator when the treatment effects (TEs) are heterogeneous across treatment cohort or calendar time.

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 CS (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.

The co-author has approved that the co-authored chapter is included. The co-author's contact: Jeffrey M. Wooldridge, Department of Economics, 486 W. Circle Drive, 110 Marshall-Adams Hall, Michigan State University, East Lansing, MI 48824-1038. Email: wooldri1@msu.edu

Borusyak et al. (2024) 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 the 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. Under strong forms of positive serial correlation, CS (2021) can be more efficient because it is similar to a first-differencing estimator. Balanced against the loss in precision is that the CS approach can be less biased when parallel trends are violated, although there is no guarantee. See a 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 2.2 begins with the common timing case, defining the potential outcomes and parameters of interest, and establishing identification assumptions. In Section 2.3 we propose a general approach to estimation using a transformed outcome variable. In Section 2.4 we extend the framework and identification argument to the staggered case. In Section 2.5, 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 2.6 discusses how one might accommodate suspected failures of no anticipation, and how one modifies the procedure for unbalanced panels. In Section 2.7, we revisit the effect of Walmart's opening on the local labor markets. Section 2.8 contains some concluding remarks. Supplementary Appendix 2B.1 presents Monte Carlo simulations demonstrating that the rolling methods perform well in terms of both bias and precision.

2.2 Setup and Identification with Common Timing

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 indicator equal to 1 if $t \geq S$ and otherwise zero.

Without treated units prior to time S we can, at most, hope to identify average treatment effects in periods $S, S+1, \ldots, T$. 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:

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

The methods we propose can, under stronger assumptions than we propose, recover the overall average treatment effects (ATEs), $E\left[Y_r(1) - Y_r(0)\right]$, 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

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

Importantly, when $D = 1$, $Y_r = Y_r(1)$, which means

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

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

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

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

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

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

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

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

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,

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

The name of this assumption derives from the fact that $E\left[Y_t(1) - Y_t(0) | D=1 \right]$ 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 et al. (1997) and made explicit in Abadie (2005) and elsewhere. Because the variable indexing $Y_t(\cdot)$ is treatment status not yet assigned, the assumption rules out anticipatory changes in the potential outcomes, on average. If one is concerned that the announcement of a policy prior to its implementation may

result in some units strategically manipulating their pre-intervention outcomes, one might drop a period or two just prior to the intervention – as a minimum, as a robustness check. Naturally, this can result in less precise estimators.

Given Assumption NAC, we can express τ_r as

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

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 $E\left[\dot{Y}_r(1)\,|D=1\right] = E(\dot{Y}_r|D=1)$ and the latter is trivially identified (as usual, under a suitable sampling scheme). Notice in the simple $T = 2$ case with $S = 2$, $\dot{Y}_2 = Y_2 - Y_1$, the difference from period one to two.

The difficult term in identifying τ_r is $E\left[\dot{Y}_r(0)\,|D=1\right]$. The unconditional parallel trends assumption implies that $E\left[\dot{Y}_r(0)\,|D=1\right] = E\left[\dot{Y}_r(0)\,|D=0\right]$. 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 $\mathbf{X}$,

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

Simple algebra shows that (2.10) is the same as assuming $E\left[Y_t(0) - Y_s(0)|D,\mathbf{X}\right] = E[Y_t(0) - Y_s(0)|\mathbf{X}]$ 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 (**??**). Assumption CPTC allows us to identify $E\left[\dot{Y}_r(0)\,|D=1\right]$. To see how, first note that, by iterated expectations,

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

Next, write

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

Then, by CPTC,

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

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,

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

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

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

where $\dot{m}_{0r}(\mathbf{X}) \equiv E\left(\dot{Y}_r|D=0,\mathbf{X}=\mathbf{x}\right)$ 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 $E\left[\dot{m}_{0r}(\mathbf{X})\,|D=1\right]$ 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.*

□

2.3 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 Zhao (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

$$Y_{it} = (1 - D_i) \cdot Y_{it}(0) + D_i \cdot Y_{it}(1) \tag{2.16}$$

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

$$\dot{Y}_{ir} = Y_{ir} - \frac{1}{(S-1)} \sum_{q=1}^{S-1} Y_{iq} \equiv Y_{ir} - \bar{Y}_{i,pre} \tag{2.17}$$

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

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,

$$E\left(\dot{Y}_r | D = 0, \mathbf{X}\right) = \dot{\alpha}_r + \mathbf{X}\beta_r \tag{2.18}$$

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

$$\dot{Y}_{ir} \text{ on } 1, \mathbf{X}_i \text{ if } D_i = 0 \tag{2.19}$$

Then, τ_r can be estimated using imputation:

$$\widehat{\tau}_r = \overline{\dot{Y}}_{r1} - N_1^{-1} \sum_{i=1}^{N} D_i \left(\widehat{\dot{\alpha}}_r + \mathbf{X}_i \widehat{\beta}_r \right) = \overline{\dot{Y}}_{r1} - \left(\widehat{\dot{\alpha}}_r + \bar{\mathbf{X}}_1 \widehat{\beta}_r \right) \tag{2.20}$$

where $\overline{\dot{Y}}_{r1} = N_1^{-1} \sum_{i=1}^{N} D_i \dot{Y}_{ir}$ 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 $\widehat{\tau}_r$ can be obtained from standard software that does basic regression adjustment once the $\dot{Y}_{ir}$ have been obtained.

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

$$\dot{Y}_{ir} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1 \right), i = 1, 2, \ldots, N, \tag{2.21}$$

which uses all observations in time period r. This formulation is convenient because it leads to simple inference for $\widehat{\tau}_r$, 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 $\mu_1 \equiv E(\mathbf{X}|D = 1)$. 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 (2.17), there is a simple characterization of $\widehat{\tau}_r$. All of the coefficients in (2.21) are obtained by differencing the coefficients from two separate regressions. In the first, Y_{ir} is regressed on all variables in (2.21). Then $\bar{Y}_{i,pre}$ is regressed on the same set of variables, and these coefficients are subtracted from the first. In particular, $\widehat{\tau}_r$ 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

$$\widehat{\tau}_r = \left(\bar{Y}_{1r} - \bar{Y}_{0r} \right) - \left(\bar{Y}_{1,pre} - \bar{Y}_{0,pre} \right), \tag{2.22}$$

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 (2.21) has additional benefits. For example, if D_i is independent of $\dot{Y}_{ir}(0)$–so that the unconditional PT assumption holds–, then the covariates $\mathbf{X}_i$ need not be included in (2.21) 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 (2.21) 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 (2.21) is simply ΔY_i on $1, D_i, \mathbf{X}_i, D_i \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right), i = 1, 2, \ldots, N$ 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:

$$E(Y_t|D, \mathbf{X}) = \alpha + \mathbf{X}\beta + \gamma D + (D \cdot \mathbf{X})\,\delta + \sum_{r=2}^{T} \theta_r f r_t + \sum_{r=2}^{T} (f r_t \cdot \mathbf{X})\,\pi_r + \sum_{r=S}^{T} \tau_r (D \cdot f r_t) + \sum_{r=S}^{T} (D \cdot f r_t)(\mathbf{X} - \mu_1)\,\eta_r, \; t = 1, \ldots, T, \tag{2.23}$$

where $f r_t$ is a time period dummy equal to one if $r = t$ and zero otherwise. The interaction $D \cdot f r_t$ is the treatment indicator for time period r. Equation (2.21) suggests a pooled OLS regression across all i and t:

$$\begin{aligned} &Y_{it} \text{ on } 1, \mathbf{X}_i, D_i, D_i \cdot \mathbf{X}_i, f2_t, \ldots, fT_t, f2_t \cdot \mathbf{X}_i, \ldots, fT_t \cdot \mathbf{X}_i \\ &\quad D_i \cdot fS_t, \ldots, D_i \cdot fT_t, D_i \cdot fS_t \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right), \ldots, D_i \cdot fT_t \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right) \end{aligned} \tag{2.24}$$

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

Theorem 2.2 *Let $\widehat{\tau}_r$, $r = S, S+1, \ldots, T$ be the coefficients on D_i in the separate regressions (2.21) – equivalently, from equation (2.20) – and let $\widetilde{\tau}_r$ be the coefficients on $D_i \cdot fS_t, \ldots, D_i \cdot fT_t$ from (2.24). Then $\widetilde{\tau}_r = \widehat{\tau}_r$, $r = S, \ldots, T$. Moreover, the coefficient vector on $D_i \cdot fr_t \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)$ in (2.24) is identical to that on $D_i \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right)$ in (2.21).* □

The proof of Theorem 2.2 can be found in Appendix 2A. 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 (2.24) are both best linear unbiased and asymptotically efficient (with T fixed, $N \to \infty$) under random sampling across i. This establishes that the transformation used in (2.21) 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 $\left\{\left(\dot{Y}_{ir}, D_i, \mathbf{X}_i\right) : i = 1, 2, \ldots, N\right\}$. We can also apply propensity score matching, covariate matching or nearest neighbor matching.

Procedure 3.1 (Rolling Methods, Common Timing):

Step 1. For a given time period $r \in \{S, \ldots, T\}$ and each unit i, compute $\dot{Y}_{ir}$ as in (2.17).

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

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

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 (2.17) to that in CS (2021). In the common timing case, the CS transformation, for $r \geq S$, is

$$\mathring{Y}_{ir} = Y_{ir} - Y_{i,S-1}, \tag{2.25}$$

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. In many cases, the CS transformation will be inefficient compared with (2.17) because the CS differencing ignores time periods other than $S - 1$. However, there are cases where CS (2021) can be more efficient. 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 Zhao (2020) in the panel data case.

2.4 Staggered Interventions

2.4.1 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

$$Y_t(g), g \in \{S, \ldots, T, \infty\}, t \in \{1, 2, \ldots, T\}, \tag{2.26}$$

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 the 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

$$\tau_{gr} = E\left[Y_r(g) - Y_r(\infty) \,|D_g = 1\right], r = g, \ldots, T;\ g = S, \ldots, T \tag{2.27}$$

For each treated cohort g, $\tau_{gr}(r = g, \ldots, T)$ denotes the ATT in all subsequent time period.

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

$$\begin{aligned} Y_t(g) - Y_t(\infty) = &\left[Y_t(g) - \frac{1}{(g-1)}\sum_{s=1}^{g-1} Y_s(g)\right] - \left[Y_t(\infty) - \frac{1}{(g-1)}\sum_{s=1}^{g-1} Y_s(\infty)\right] \\ &+ \frac{1}{(g-1)}\sum_{s=1}^{g-1}\left[Y_s(g) - Y_s(\infty)\right] \end{aligned} \tag{2.28}$$

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 $\mathbf{X}$,

$$E\left[Y_t(g) \,|D_g = 1, \mathbf{X}\right] = E\left[Y_t(\infty) \,|D_g = 1, \mathbf{X}\right]. \square \tag{2.29}$$

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,

$$\tau_{gr} = E\left[\dot{Y}_{rg}(g) \,|D_g = 1\right] - E\left[\dot{Y}_{rg}(\infty) \,|D_g = 1\right], \tag{2.30}$$

where $\dot{Y}_{rg}(g)$ and $\dot{Y}_{rg}(\infty)$ are defined as the first and second terms in (2.28), 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 (2.28) 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 $E\left[\dot{Y}_{rg}(\infty) \,|D_g = 1\right]$. 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 $t = 1, 2, \ldots, T$,

$$E\left[Y_t(\infty) - Y_1(\infty)|\mathbf{D}, \mathbf{X}\right] = E\left[Y_t(\infty) - Y_1(\infty)|\mathbf{X}\right],\ t = 2, \ldots, T. \square \tag{2.31}$$

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,

$$\begin{aligned} E\left[\dot{Y}_{rg}(\infty)|D_g = 1, \mathbf{X}\right] =& \frac{1}{(g-1)} \sum_{s=1}^{g-1} E\left[Y_r(\infty) - Y_s(\infty)|D_g = 1, \mathbf{X}\right] \\ =& \frac{1}{(g-1)} \sum_{s=1}^{g-1} E\left[Y_r(\infty) - Y_s(\infty)|D_\infty = 1, \mathbf{X}\right] \\ =& E\left[\dot{Y}_{rg}(\infty)|D_\infty = 1, \mathbf{X}\right] \end{aligned} \tag{2.32}$$

where the second equality follows from CPTS and the third follows by taking the expectation outside the summation. We have shown the following.

Theorem 2.3 *Under Assumption CNAS, equation (2.30) holds. If we add Assumption CPTS, the cohort assignments,* $\mathbf{D} = (D_S, \ldots, D_T)$ *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* $\mathbf{X}$. $\square$

Because the vector of cohort indicators is unconfounded with respect to $\dot{Y}_{rg}(\infty)$, Theorem 2.3 implies

$$E\left[\dot{Y}_{rg}(\infty)|D_\infty = 1, \mathbf{X}\right] = E\left[\dot{Y}_{rg}(\infty)|D_h = 1, \mathbf{X}\right],\ h = S, \ldots, T \tag{2.33}$$

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,

$$E\left[\dot{Y}_{rg}(\infty)|D_h = 1, \mathbf{X}\right] = E\left[\dot{Y}_{rg}(h)|D_h = 1, \mathbf{X}\right],\ h = r + 1, \ldots, T \tag{2.34}$$

Combined, (2.33) and (2.34) mean that, in addition to the never treated (NT) group, we can use treatment cohorts $h \in \{r+1, \ldots, T\}$ in estimating $E\left[\dot{Y}_{rg}(\infty) | D_\infty = 1, \mathbf{X}\right]$. 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 $E\left[\dot{Y}_{rg}(\infty) | D_g = 1, \mathbf{X}\right]$ 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 $A_{r+1} \equiv D_{r+1} + D_{r+2} + \cdots + D_T + D_\infty$. 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 $r \in \{g, g+1, \ldots, T\}$,

$$P\left(D_g = 1 | D_g + A_{r+1} = 1, \mathbf{X} = \mathbf{x}\right) < 1 \text{ for all } \mathbf{x} \in \text{Supp}(\mathbf{X}) . \square \tag{2.35}$$

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):

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

$$\dot{Y}_{irg} \equiv Y_{ir} - \frac{1}{(g-1)} \sum_{s=1}^{g-1} Y_{is} \equiv Y_{ir} - \bar{Y}_{i,pre(g)} \tag{2.36}$$

Step 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 the NT group).

Step 3. Using the subset of data with

$$D_{ig} + D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1, \tag{2.37}$$

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

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

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

When $\mathbf{X}_i$ has high dimension, Procedure 4.1 implies that machine learning methods can be applied after obtaining the transformation in (2.36). See, for example, Belloni et al. (2014) and Chernozhukov et al. (2018).

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

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

with all possible control units, reproduces the POLS estimator proposed by Wooldridge (2021). 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 trade off is that we are able to apply many different kind of estimators, including doubly robust and matching estimators.

2.4.2 Comparison to Alternative Methods

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

$$\mathring{Y}_{irg} \equiv Y_{ir} - Y_{i,g-1} \tag{2.38}$$

and then choosing control groups from cohorts $\{r + 1, \ldots, T, \infty\}$. The transformation in (2.38) ignores the control periods prior to $g - 1$ and is generally inefficient under classical assumptions. Also, the default implementation in commonly used software (R and Stata) is to use only the never-treated group as controls. Several authors, including Marcus and Sant'Anna (2021) and Callaway (2023), recognize that the version of parallel trends in Assumption CPT implies that

more information is available for estimating the ATTs. However, we believe we are the first to propose the demeaning transformation at the unit level, using the pre-treatment averages, and then combining the transformation with various treatment effects estimators after establishing unconfoundedness. Our approach makes it straightforward to apply different strategies based on different transformations and different control periods and control groups.

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.

To illustrate where the efficiency and robustness of our method originate, we compare the information sets utilized by various estimators when estimating treatment effects. Suppose there are three treatment cohorts—groups 4, 5, and 6—and our goal is to estimate the average treatment effects on the treated (ATT) for group 4 at time $t = 5$, $ATT(4, 5)$.

Table 2.1 summarizes the subset of data each estimator employs for this estimation task. For pre-treatment periods up to $t = 3$, the areas highlighted in dark gray represent the data used by each estimator, while light gray regions indicate observations that are not used. Compared to the estimator proposed by Callaway and Sant'Anna (2021), our unit-specific time-series transformation (rolling) method leverages a larger set of pre-treatment observations for each unit. This broader utilization enhances efficiency, recovering some of the precision otherwise lost in the CS (2021) framework.

Table 2.1 Pre-Treatment Period Observations Used by Each Method

	CS (2021)			Rolling Method			Wooldridge (2021)		
time	$g = 4$	$g = 5$	$g = 6$	$g = 4$	$g = 5$	$g = 6$	$g = 4$	$g = 5$	$g = 6$
1									
2									
3									
4	-			-			-		
5	★	-		★	-		★	-	
6	-	-	-	-	-	-	-	-	-

	Utilization of Information
	Loss of Information

★	Target Estimate: $ATT(4, 5)$
-	$ATT(g, t)$

However, the rolling method uses slightly less information than Wooldridge (2021)'s approach.

Specifically, estimating $ATT(4,5)$ under the rolling method requires dropping observations from units already treated by period $t = 5$, including $\dot{Y}_{55}$, which retains pre-treatment history. This exclusion is necessary to avoid "bad comparisons." A similar logic underlies the CS (2021) approach, which likewise omits already-treated units.

In contrast, Wooldridge (2021)'s regression-based framework retains all pre-treatment observations by construction, achieving high efficiency under correct specification. However, this efficiency comes at the cost of increased vulnerability to model misspecification.

In this respect, our estimator can be viewed as a middle ground that combines the robustness of the CS method with the efficiency of the Wooldridge (2010) framework. For example, our method accommodates doubly robust estimation, offering greater robustness to outcome model misspecification while still preserving much of the efficiency of the Wooldridge (2021) framework—a benefit not shared by the CS (2021) framework.

In the special case of a single treatment group–that is, under common timing—the rolling method and Extended TWFE estimator (Wooldridge, 2021) use an equivalent set of pre-treatment observations, resulting in comparable efficiency when the model is correctly specified.

2.4.3 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 $Y_t(T)$: $E\left[Y_r(g) - Y_r(T) \,|D_g = 1\right]$, $g \in \{S, \ldots, T-1\}$, $r \in \{g, \ldots, T\}$. 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 $r < T$, $E\left[Y_r(g) - Y_r(T) \,|D_g = 1\right] = E\left[Y_r(g) - Y_r(\infty) \,|D_g = 1\right]$ 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 (2.36) 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.

2.5 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

$$E\left[Y_t(\infty)|\mathbf{D}, \mathbf{X}\right] = q_\infty\left(\mathbf{X}\right) + \sum_{g=S}^{T} D_g q_g\left(\mathbf{X}\right) + m_t\left(\mathbf{X}\right), t = 1, \ldots, T, \tag{2.39}$$

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

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

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

$$E\left[\dot{Y}_{rg}\left(\infty\right)|D_\infty = 1, \mathbf{X}\right] = m_r\left(\mathbf{X}\right) - \frac{1}{(g-1)} \sum_{s=1}^{g-1} m_s\left(\mathbf{X}\right) \equiv \dot{m}_{rg}\left(\mathbf{X}\right) \tag{2.40}$$

and so $\dot{m}_{rg}\left(\cdot\right)$ is the conditional mean function implicit in the methods from Section 2.4 that use a conditional mean specification (RA IPW or IPWRA). Because $\dot{m}_{rg}\left(\cdot\right)$ can take on positive and negative values, we essentially assumed in regression-based methods that $\dot{m}_{rg}\left(\cdot\right)$ can be approximated by a function linear in parameters (allowing controls to appear flexibly, as usual).

The representation in (2.39) suggests a way to relax parallel trends for cohorts where we have at least two pre-treatment time periods. We replace (2.39) with the following assumption.

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

$$E\left[Y_t(\infty)|\mathbf{D}, \mathbf{X}\right] = \eta_S \left(D_S \cdot t\right) + \cdots + \eta_T \left(D_T \cdot t\right) + q_\infty \left(\mathbf{X}\right) + \sum_{g=S}^{T} D_g q_g \left(\mathbf{X}\right) + m_t \left(\mathbf{X}\right). \quad \square \tag{2.41}$$

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

$$E\left[Y_t(\infty) - Y_{t-1}(\infty)|\mathbf{D}, \mathbf{X}\right] = \eta_g D_g + \cdots + \eta_T D_T + \left[m_t \left(\mathbf{X}\right) - m_{t-1} \left(\mathbf{X}\right)\right] \tag{2.42}$$

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 2.3 and 2.4 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

$$\begin{aligned} Y_{it}\left(\infty\right) &= \mathbf{h}\left(\mathbf{D}_i, \mathbf{X}_i\right) + \mathbf{D}_i \cdot t \cdot \eta + m_t\left(\mathbf{X}_i\right) + U_{it}\left(\infty\right) \\ E\left[U_{it}\left(\infty\right)|\mathbf{D}_i, \mathbf{X}_i\right] &= 0, \end{aligned} \tag{2.43}$$

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

Equation (2.43) is an example of a heterogeneous (or random) trend model of the kind discussed in Wooldridge (2010, Section 11.7.2).Appendix 2B details how unit-specific linear trends are removed by regressing the outcome on a constant and time dummies using pre-treatment periods. The resulting residuals define the detrended outcome variable, denoted by $\ddot{Y}_{irg}(\infty)$. We then show that the treatment cohort indicators, D_i, are unconfounded with respect to these detrended variable, conditional on X_i. Moreover, by Assumption CNAS, cohorts with $h > r$ (including $h = \infty$) can be used as part of the control group. Then, the argument is as in Section 2.4: 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):

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

$$Y_{it} \text{ on } 1, t, \ t = 1, \ldots, g-1 \tag{2.44}$$

For $r \in \{g, \ldots, T\}$, compute the out-of-sample predicted value $\widehat{Y}_{irg}$ and the prediction error (detrended variable) $\ddot{Y}_{irg} \equiv Y_{ir} - \widehat{Y}_{irg}$. (One need not do this for units treated prior to period g.)

Step 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 the NT group).

Step 3. Using the subset of data with $D_{ig} + D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1$, or a further subset, apply standard TE methods – such as linear RA, IPW, IPWRA or matching – to the cross section

$$\left\{\left(\ddot{Y}_{irg}, D_{ig}, \mathbf{X}_i\right) : i = 1, \ldots, N\right\}, \tag{2.45}$$

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 (2.43) is done with $g = S$ only and then the detrended outcomes $\ddot{Y}_{ir}$ 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

$$N_1^{-1} \sum_{i=1}^{N} D_i \ddot{Y}_{i3} - N_0^{-1} \sum_{i=1}^{N} (1 - D_i) \ddot{Y}_{i3}, \tag{2.46}$$

where $\ddot{Y}_{i3}$ is obtained as the prediction error in period three after the regression Y_{it} on $1, t$, $t = 1, 2$. After a little algebra, (2.46) can be shown to equal

$$\left[\left(\bar{Y}_{13} - \bar{Y}_{12}\right) - \left(\bar{Y}_{03} - \bar{Y}_{02}\right)\right] - \left[\left(\bar{Y}_{12} - \bar{Y}_{11}\right) - \left(\bar{Y}_{02} - \bar{Y}_{01}\right)\right], \tag{2.47}$$

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 (2.47) 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 trends 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 (2.44) does not mean there is an incidental parameters problem that can cause inconsistency in the $\widehat{\tau}_{gr}$ 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).

2.6 Violations 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, the transformed outcome ($\dot{Y}_{irg}$ or $\ddot{Y}_{irg}$) 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} (∞) – 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.

2.7 Application

In this section, we revisit the literature examining the effects of Walmart's entry on local labor markets (Basker, 2005, Neumark et al., 2008, Brown and Butts, 2023). Prior studies highlight concerns about potential violations of the parallel trends assumption. For example, Brown and Butts (2023) provide visual evidence suggesting that much of the estimated effect could be due to pre-existing trends rather than the impact of Walmart's entry (Rambachan and Roth, 2023). If store opening decisions are mainly driven by county-level economic fundamentals, observed employment differences between treated and control groups may not be attributable solely to Walmart's opening.

2.7.1 Data

As we show in Section 2.5, to illustrate that our rolling method yields robust results even when counties exhibit disparate trends after controlling for covariates, we use the dataset constructed by Brown and Butts (2023), which follows Basker (2005) and draws on the County Business Patterns (CBP) data from 1964 and 1977–1999.

We limit the dataset to counties that had more than $1,500$ of total employment in 1964 and had non-negative employment growth rates during 1964-1977, each of which was imputed from the 1977-1999 dataset (see, Brown and Butts, 2023). The Walmart store opening indicator is derived

from the geographic dataset of openings in Arcidiacono et al. (2020). We exclude counties whose first Walmart opened before 1985. As a result, we obtain a balanced panel of 1280 counties, each observed for nine pre-treatment years and fourteen years of post-treatment periods, for a total of 23 years.

Figure 2.1 plots the sample counties and their first Walmart opening year. Darker shades indicate later openings, light green shows never-treated counties, and gray marks excluded areas.

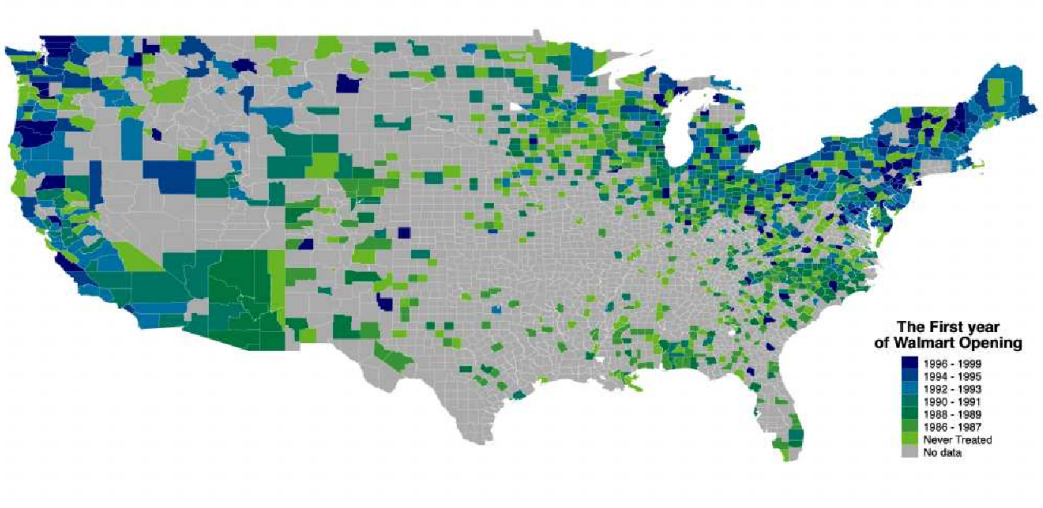


Figure 2.1 First Year of Walmart Store Openings by County

The treatment cohort is defined by the year of first opening. We utilize covariates, including the 1980 shares of the population employed in manufacturing, above the poverty line, and with a high school education. Never-treated counties comprise 31% of the sample. For descriptive statistics of the variables used in our analysis, see Appendix **??** Table 2D.1. We use the dataset to assess how Walmart openings affect county-level retail and wholesale employment.

2.7.2 Estimation Results

First, we estimate cohort-year specific average treatment effects on the treated (ATT), denoted as $ATT(g,t)$, for cohort g in year t. We then compute the weighted sum of ATTs by relative time (length of exposure) and calculate the 95% bootstrap confidence intervals. We define the weighted

ATTs, $WATT(r)$, as follows:

$$WATT\,(r) = \sum_{g \in G_r} w(g,r) \cdot ATT(g, g+r)$$

where $r = t - g$, for all $r \in \{0, 1, \ldots, T-S\}$, $g \in G_r = \{S, \ldots, T-r\}$, and $w(g,r) = \frac{N_g}{N_{G_r}}$ (N_g is the number of counties in treated-cohort g, N_{G_r} is the number of counties in G_r). For pre-treatment periods ($r < 0$), the procedure used to obtain the ATTs is detailed in Appendix **??**.

For example, $WATT(0)$ denotes the weighted average of immediate impact on the log of employment level, which we are averaging out ATTs of the first treated year across every treated-cohort; i.e., $ATT(g, g)$ for all $g \in G_{r=0}$.

Lastly, we compare the results from CS (2021) approach, rolling IPWRA estimator with unit-specific demeaning, and rolling IPWRA estimator with unit-specific detreading.

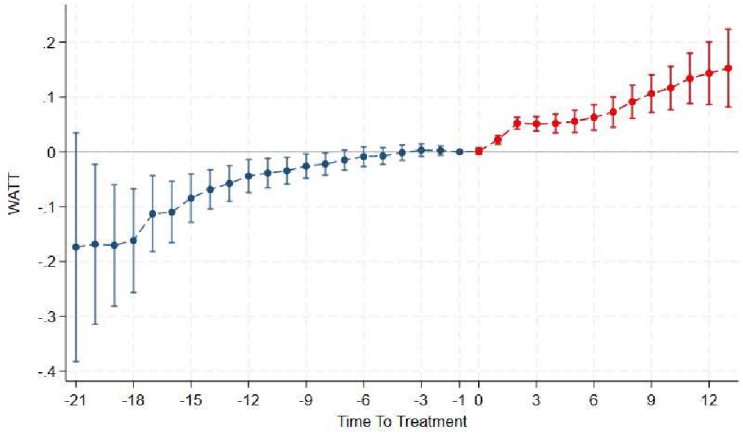
2.7.2.1 Retail Employment

In this section, we estimate the impact of Walmart's entry on county-level retail employment, using the log of employment as the outcome. Figure 2.2 reports weighted ATT estimates with bootstrapped 95% confidence intervals from the three estimators.

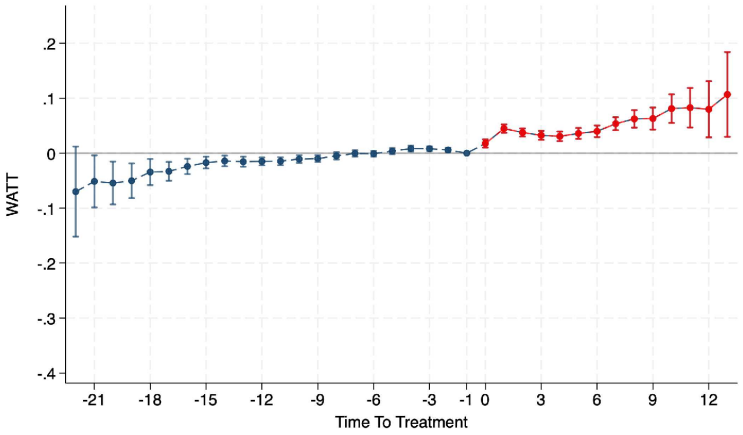
Panel (a) is an event-study plot generated using the CS (2021) approach. The blue line represents the pre-trends, which ideally should be zero; however, it shows a clear positive slope, indicating pre-existing differences between treated and control counties. As a result, the estimated treatment effects—shown in red—likely do not reflect the true impact of Walmart's entry on wholesale employment.

Panel (b) presents results from our rolling IPWRA estimator applied to the demeaned outcome variable. This approach smooths out much of the pre-trend bias relative to the CS estimator, but the pre-trends are still not fully addressed.

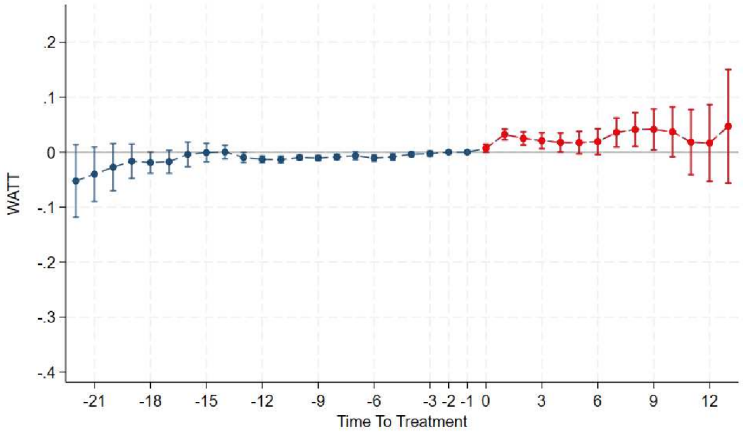
Panel (c) shows results from the rolling IPWRA estimator with detrending, which effectively eliminates county-level linear trends. The flatter pre-treatment trajectory indicates better adjustment for pre-existing differences. Post-entry, we observe modest but statistically significant increases in retail employment that decline over time and eventually converge to zero.



Panel (a) CS Approach



Panel (b) Rolling IPWRA estimator



Panel (c) Rolling IPWRA with unit-specific detrending

Figure 2.2 Effects of Walmart Opening on log(Retail Employment)

For example, ATT(1) from the rolling IPWRA with heterogeneous trends is 0.032 (SE = 0.005), indicating a 3.2% increase in retail employment one year after entry (see Appendix **??** for all

estimates). Given an average of 6,589 retail employees in treated counties, this translates to about 210 additional jobs. However, since Walmart stores typically employ 150–300 workers (Basker, 2005), it is difficult to attribute broader local employment gains beyond direct hiring. Results for wholesale employment appear in Appendix 2D.1.

2.8 Concluding Remarks

In this paper, we propose a flexible transformation approach for estimating treatment effects in panel data with staggered interventions and heterogeneous trends. A key advantage of this method is that once the transformed dependent variable is defined—whether in the common timing case, the staggered case, or after removing unit-specific trends—researchers can apply standard treatment effect estimators to the resulting cross-sectional data. This includes regression adjustment, inverse probability weighting, and doubly robust procedures. The transformation is easily adapted to accommodate unit-specific trendsand allows for dynamic and heterogeneous treatment effects across units and time periods. Our empirical application to Walmart's entry demonstrates how the approach improves estimation accuracy and robustness compared to existing methods. These features, combined with its simplicity of implementation, make the method a valuable tool for applied researchers.

BIBLIOGRAPHY

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, 72(1):1–19.

Arcidiacono, P., Ellickson, P. B., Mela, C. F., and Singleton, J. D. (2020). The competitive effects of entry: Evidence from supercenter expansion. *American Economic Journal: Applied Economics*, 12(3):175–206.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.

Basker, E. (2005). Job creation or destruction? labor market effects of wal-mart expansion. *Review of Economics and Statistics*, 87(1):174–183.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Borusyak, K. and Jaravel, X. (2018). Revisiting event study designs. *Available at SSRN 2826228*.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, page rdae007.

Brown, N. and Butts, K. (2023). Dynamic treatment effect estimation with interactive fixed effects and short panels.

Callaway, B. (2023). Difference-in-differences for policy evaluation. *Handbook of Labor, Human Resources and Population Economics*, pages 1–61.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, 110(9):2964–2996.

De Chaisemartin, C. and d'Haultfoeuille, X. (2023). Two-way fixed effects and difference-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 26(3):C1–C30.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Marcus, M. and Sant'Anna, P. H. (2021). The role of parallel trends in event study settings: An application to environmental economics. *Journal of the Association of Environmental and Resource Economists*, 8(2):235–275.

Negi, A. and Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534.

Neumark, D., Zhang, J., and Ciccarella, S. (2008). The effects of wal-mart on local labor markets. *Journal of Urban Economics*, 63(2):405–430.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, 225(2):175–199.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data.

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. *The Econometrics Journal*, 26(3):C31–C66.

APPENDIX 2A

PROOF OF THEOREM 2.2

We modify the argument in Wooldridge (2021, Theorem 8.1). The $\widehat{\tau}_r$ are obtained from regression (2.21). Because $\dot{Y}_{ir} = Y_{ir} - \bar{Y}_{i,pre}$, basic OLS algebra shows that all coefficients from (2.21) are obtained by differencing the coefficients from the two regressions

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

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

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

$$\widehat{\tau}_r = \widehat{\rho}_r - \widehat{\rho}_{pre} \tag{2A.3}$$

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

To show (2A.3) is the same as the coefficient on $D_i \cdot fr_t$ in (2.24), 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 $\widetilde{\tau}_r$ are the coefficients on $D_i \cdot fr_t$ in the regression

$$\begin{aligned} Y_{it} \text{ on } 1, \mathbf{X}_i, D_i, D_i \cdot \mathbf{X}_i, fS_t, \ldots, fT_t, fS_t \cdot \mathbf{X}_i, \ldots, fT_t \cdot \mathbf{X}_i \\ D_i \cdot fS_t, \ldots, D_i \cdot fT_t, D_i \cdot fS_t \cdot \mathbf{X}_i, \ldots, D_i \cdot fT_t \cdot \mathbf{X}_i \end{aligned} \tag{2A.4}$$

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

$$\mathbf{L}_{it} \equiv (fS_t, fS_t \cdot \mathbf{X}_i, D_i \cdot fS_t, D_i \cdot fS_t \cdot \mathbf{X}_i, \ldots, fT_t, , fT_t \cdot \mathbf{X}_i, D_i \cdot fT_t \cdot \mathbf{X}_i), \tag{2A.5}$$

a row vector with $2\,(T - S + 1)\,(K + 1)$ elements.

Note that for $q \neq r$, $(fq_t, fq_t \cdot \mathbf{X}_i, D_i \cdot fq_t, D_i \cdot fq_t \cdot \mathbf{X}_i)$ and $(fr_t, fr_t \cdot \mathbf{X}_i, D_i \cdot fr_t, D_i \cdot fr_t \cdot \mathbf{X}_i)$ are orthogonal in sample because $fq_t \cdot fr_t = 0$. The full set of regressors in (2A.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 (2A.4) can be written as

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

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

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\widetilde{\theta}) + \sum_{i=1}^{N}\sum_{t=1}^{T} p_t\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\widetilde{\theta}-\mathbf{L}_{it}\widetilde{\delta}) = \mathbf{0} \tag{2A.7}$$

$$\sum_{i=1}^{N}\sum_{t=1}^{T} p_t\mathbf{L}_{it}'(Y_{it}-\mathbf{H}_i\widetilde{\theta}-\mathbf{L}_{it}\widetilde{\delta}) = \mathbf{0} \tag{2A.8}$$

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

$$p_t\mathbf{H}_i = [p_t, p_t\cdot D_i, p_t\cdot\mathbf{X}_i, p_t\cdot D_i\cdot\mathbf{X}_i] = \sum_{q=S}^{T}[fq_t, fq_t\cdot D_i, fq_t\cdot\mathbf{X}_i, fq_t\cdot D_i\cdot\mathbf{X}_i], \tag{2A.9}$$

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 (2A.8) gives

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\widetilde{\theta}) + \mathbf{A}'\sum_{i=1}^{N}\sum_{t=1}^{T} p_t\mathbf{L}_{it}'(Y_{it}-\mathbf{H}_i\widetilde{\theta}-\mathbf{L}_{it}\widetilde{\delta}) = \mathbf{0} \tag{2A.10}$$

Along with (2A.8), (2A.10) implies that the FOCs for $\left(\widetilde{\theta},\widetilde{\delta}\right)$ are

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(1-p_t)\mathbf{H}_i'(Y_{it}-\mathbf{H}_i\widetilde{\theta}) = \mathbf{0} \tag{2A.11}$$

But (2A.11) means that $\widetilde{\theta}$ 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, $\widetilde{\theta}$ is the same as the cross-sectional regression

$$\bar{Y}_{i,pre} \text{ on } 1, D_i, \mathbf{X}_i, D_i\cdot\mathbf{X}_i \tag{2A.12}$$

In particular, the coefficient on D_i is precisely $\widehat{\rho}_{pre}$ in (2A.3).

Next, the FOC in (2A.3) shows that $\widetilde{\delta}$ is from a POLS regression using the post-treatment periods:

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

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, $\widetilde{\delta}_r$, $r = S, \ldots, T$, is obtained from a separate cross-sectional regression for each post-treatment period. Namely, because $fr_r = 1$, the regression for period r is

$$Y_{ir} - \mathbf{H}_i\widetilde{\theta} \text{ on } 1, D_i, \mathbf{X}_i, D_i \cdot \mathbf{X}_i, i = 1, \ldots, N \tag{2A.13}$$

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

$$\widetilde{\delta}_r = \left(\sum_{i=1}^{N} \mathbf{H}_i'\mathbf{H}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{H}_i' Y_{ir}\right) - \widetilde{\theta} \tag{2A.14}$$

The first term is the regression coefficients from

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

which is $\widehat{\rho}_r$ from (2A.3). We have shown that the coefficient corresponding to $D_i \cdot fr_t$ in the regression (2A.4) is $\widehat{\rho}_r - \widehat{\rho}_{pre}$, which establishes the equivalence of the pooled OLS estimator across all time periods, (2.24), 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 $D_i \cdot fr_t \cdot \mathbf{X}_i$ in (2.24) are the same as the coefficients on $D_i \cdot \mathbf{X}_i$ in (2.21). □

APPENDIX 2B

PROOF OF THE DETRENDING PROCEDURE

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

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

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

$$\mathbf{Y}_{i,g-1}(\infty) = \left[Y_{i1}(\infty), \ldots, Y_{i,g-1}(\infty)\right]'. \tag{2B.2}$$

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

$$\begin{aligned} \mathbf{Y}_{i,g-1}(\infty) &= \mathbf{J}_{g-1} \begin{pmatrix} \mathbf{h}(\mathbf{D}_i, \mathbf{X}_i) \\ \mathbf{D}_i \eta \end{pmatrix} + \mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty) \\ &\equiv \mathbf{J}_{g-1}\mathbf{Q}_i + \mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty) \end{aligned} \tag{2B.3}$$

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

$$\begin{aligned} \widehat{\mathbf{B}}_{i,g-1} &= \left(\mathbf{J}'_{g-1}\mathbf{J}_{g-1}\right)^{-1} \mathbf{J}'_{g-1}\mathbf{Y}_{i,g-1}(\infty) \\ &= \mathbf{Q}_i + \left(\mathbf{J}'_{g-1}\mathbf{J}_{g-1}\right)^{-1} \mathbf{J}'_{g-1}\left[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)\right] \end{aligned} \tag{2B.4}$$

For $r \geq g$, the prediction of $Y_{ir}(\infty)$ using the unit-specific linear trend up through period $g-1$ is

$$\widehat{Y}_{irg}(\infty) \equiv (1, r)\,\widehat{\mathbf{B}}_{i,g-1} \tag{2B.5}$$

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

$$\begin{aligned} \ddot{Y}_{irg}(\infty) \equiv Y_{ir}(\infty) - \widehat{Y}_{irg}(\infty) &= Y_{ir}(\infty) - (1, r)\,\widehat{\mathbf{B}}_{i,g-1} = (1, r)\,\mathbf{Q}_i + m_r(\mathbf{X}_i) + U_{ir}(\infty) \\ &\quad - (1, r)\left\{\mathbf{Q}_i + \left(\mathbf{J}'_{g-1}\mathbf{J}_{g-1}\right)^{-1} \mathbf{J}'_{g-1}\left[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)\right]\right\} \\ &= m_r(\mathbf{X}_i) + U_{ir}(\infty) - (1, r)\left\{\left(\mathbf{J}'_{g-1}\mathbf{J}_{g-1}\right)^{-1} \mathbf{J}'_{g-1}\left[\mathbf{M}_{i,g-1} + \mathbf{U}_{i,g-1}(\infty)\right]\right\} \end{aligned} \tag{2B.6}$$

The expression in (2B.6) shows that $\ddot{Y}_{ir}(\infty)$ does not depend on $\mathbf{D}_i$. In particular,

$$\begin{aligned} E\left[\ddot{Y}_{irg}(\infty)\,|\mathbf{D}_i, \mathbf{X}_i\right] &= E\left[\ddot{Y}_{irg}(\infty)\,|\mathbf{X}_i\right] \\ &= m_r(\mathbf{X}_i) - (1, r)\left(\mathbf{J}_{g-1}'\mathbf{J}_{g-1}\right)^{-1}\mathbf{J}_{g-1}'\mathbf{M}_{i,g-1} \end{aligned} \tag{2B.7}$$

This conclusion is practically important because it shows that the vector of treatment cohort indicators, $\mathbf{D}_i$, are unconfounded with respect to the detrended variable $\ddot{Y}_{ir}(\infty)$, conditional on $\mathbf{X}_i$. Note how this extends the argument in Section 2.4 where, instead of $\mathbf{J}_{g-1}$ having rows $(1, t)$, its rows simply consisted of unity.

Now the modification to the arguments in Section 2.4 are straightforward. In place of (2.28) we have

$$Y_{ir}(g) - Y_{ir}(\infty) = \ddot{Y}_{irg}(g) - \ddot{Y}_{irg}(\infty) + \left[\widehat{Y}_{irg}(g) - \widehat{Y}_{irg}(\infty)\right], \tag{2B.8}$$

where $\ddot{Y}_{irg}(g) \equiv Y_{ir}(g) - \widehat{Y}_{irg}(g)$ and $\widehat{Y}_{irg}(g)$ are the predicted values from (2B.6) and (2B.5) 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 $D_g = 1$:

$$\begin{aligned} \tau_{gr} &= E\left[Y_{ir}(g) - Y_{ir}(\infty)\,|D_{ig} = 1\right] \\ &= E\left[\ddot{Y}_{irg}(g) - \ddot{Y}_{irg}(\infty)\,|D_{ig} = 1\right] + E\left[\widehat{Y}_{irg}(g) - \widehat{Y}_{irg}(\infty)\,|D_{ig} = 1\right] \end{aligned} \tag{2B.9}$$

The second term in (2B.9) is zero by no anticipation because $\widehat{Y}_{irg}(g)$ and $\widehat{Y}_{irg}(\infty)$ are the same linear functions of the potential outcomes in periods $\{1, 2, \ldots, g-1\}$. Therefore,

$$\tau_{gr} = E\left[\ddot{Y}_{irg}(g) - \ddot{Y}_{irg}(\infty)\,|D_{ig} = 1\right] \tag{2B.10}$$

2B.1 Monte Carlo Simulations

In this supplementary 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 our rolling estimators: regression adjustment (RA), inverse-probability-weighted

regression adjustment (IPWRA), and propensity score matching (PSM). The final estimator is CS (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) whose transformation is in equation (2.38).

2B.2 Common Timing Case

We consider the common timing case. Recall from Theorem 2.1., that the POLS method in Wooldridge (2021) and regression adjustment using our rolling method are the same in the common timing case. Therefore, we have four estimators in the simulations. We assume that the 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 $S = 4$, which implies three post-treatment periods. Across Monte Carlo simulations we draw random samples of sizes 100, 500, and 1, 000. We report bias, Monte Carlo standard deviation, and the root mean squared error (RMSE) of each estimator. All simulations use 1, 000 Monte Carlo replications.

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

$$p\,(\mathbf{x}) = P(D = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp\,(\mathbf{Z}_1\gamma_1)}{1 + \exp\,(\mathbf{Z}_1\gamma_1)} \tag{2B.11}$$

where the propensity score index function, $\mathbf{Z}_1\gamma_1$, is

$$\mathbf{Z}_1\gamma_1 = -1.2 + \frac{(X_1 - 4)}{2} - X_2 \tag{2B.12}$$

Our second step is generating heterogeneous treatment effects as follows:

$$\tau_r\,(\mathbf{X}) = \theta \cdot \sum_{r=S}^{T} (r - S + 1)^{-1} + \lambda_r \cdot h(\mathbf{X}),\ r \in \{S, \ldots, T\}, \tag{2B.13}$$

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

$$h(\mathbf{X}) = \frac{(X_1 - 4)}{2} + \frac{X_2}{3} \tag{2B.14}$$

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

$$h(\mathbf{X}) = \frac{(X_1 - 4)}{2} + \frac{X_2}{3} + \frac{(X_1 - 4)^2}{4} + (X_1 - 4) \cdot \frac{X_2}{2} \tag{2B.15}$$

We generate the potential outcomes in the untreated state as

$$Y_t(0) = \delta_t + C + \beta_t \cdot f(\mathbf{X}) + U_t(0), \tag{2B.16}$$

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:

$$\beta' = (\beta_1, \beta_2, \ldots, \beta_T) = (1.0, 1.5, 0.8, 1.5, 2, 2.5) \tag{2B.17}$$

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

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2} \tag{2B.18}$$

and

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2} + \frac{(X_1 - 4)^2}{3} + (X_1 - 4) \cdot \frac{X_2}{4} \tag{2B.19}$$

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 (2B.14) and (2B.18) are used in simulations, the conditional mean is correctly specified. However, when the data are generated as in (2B.15) and (2B.19), quadratic term X_1^2 and an interaction term $X_1 \cdot X_2$ are included; the conditional mean is misspecified.

We also consider a case where $(X_1 - 4)^2/2$ is added to the index function in the propensity score, and so the estimated logit model, which is always estimated assuming an index linear in X_1 and X_2, is misspecified:

$$\mathbf{Z}_2\gamma_2 = -1.2 + \frac{(X_1 - 4)}{2} - X_2 + \frac{(X_1 - 4)^2}{2} \tag{2B.20}$$

Table 2B.1 describes basic setups for each scenario.

Table 2B.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	$(B.4)$	$(B.8)$	Yes	$(B.3)$
Scenario 2C	Yes	$(B.4)$	$(B.8)$	No	$(B.10)$
Scenario 3C	No	$(B.5)$	$(B.9)$	Yes	$(B.3)$
Scenario 4C	No	$(B.5)$	$(B.9)$	No	$(B.10)$

Simulation Results This section shows the simulation results, especially for two different scenarios listed in Table 2B.1: Scenario 1C and 3C. The results for Scenario 1C are shown in Table 2B.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.

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 the POLS.

In terms of RMSE, the POLS estimator is uniformly better in Table 2B.2 – again, this is not surprising because POLS is the BLUE. When $N = 1,000$, rolling 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 higher than that of POLS/RA estimator, 0.379.

Table 2B.2 Scenario 1C: When $E(Y_t|\mathbf{X} = \mathbf{x})$ 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

Note 1: The population R-squared values are about 0.39, 0.36, and 0.36, respectively.
Note 2: The average propensity score is about 0.26.

Table 2B.3 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 rolling 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.

Table 2B.3 Scenario 3C; When $E(Y_t|\mathbf{X} = \mathbf{x})$ 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

Note 1: The population R-squared values are about 0.41, 0.38, and 0.38, respectively.
Note 2: The average propensity score is about 0.17.

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

APPENDIX 2C

PRE-PERIOD DYNAMICS AND EVENT STUDY PLOT

This section describes how we compute transformed outcomes in the pre-treatment periods to estimate pre-trend treatment effects and plot them in the event study graph.

2C.1 Demeaning for Pre-treatment Periods, $\dot{Y}_{itg}$

For each unit i, group g, and pre-treatment time period $t \in \{1, 2, \ldots, g-1\}$, we define:

$$\dot{Y}_{itg} \equiv Y_{it} - \frac{1}{g-t-1} \sum_{q=t+1}^{g-1} Y_{iq} \tag{2C.1}$$

This transformation removes the average of future outcomes from the pre-treatment period (from $t+1$ to $g-1$), helping to anchor pre-treatment dynamics relative to a consistent future baseline, $t = g-1$. This is in the spirit of a "rolling" transformation.

For example, to obtain the transformed outcome at time 2 for group 6, $\dot{Y}_{i26}$, we subtract the average of future pre-treatment outcomes ($q = \{3, 4, 5\}$) from the outcome at $t = 2$. The final pre-treatment period, $g-1$, serves as the reference point, yielding a value of zero and anchoring the dynamics accordingly.

2C.2 Detrending for Pre-treatment Periods, $\ddot{Y}_{itg}$

To obtain detrended outcome variable, we follow the same logic described in Appendix 2C.1, but but now anchor the two most recent pre-treatment periods, $g-2$ and $g-1$, since at lease two periods are necessary to estimate the fitted value $\widehat{Y}_{itg}$.

Formally, for pre-treamtent periods $t \in \{1, 2, \ldots, g-3\}$, the detrended outcome is defined as:

$$\ddot{Y}_{itg} \equiv Y_{it} - \widehat{Y}_{itg} \tag{2C.2}$$

where $\widehat{Y}_{itg}$ is the fitted value from regressing Y_{iq} on q using future pre-treatment periods $q \in \{t+1, \ldots, g-1\}$.

This procedure removes potential linear (or more general) trends in the pre-treatment outcome trajectory, yielding a detrended outcome series. For valid extrapolation, at least two future pre-treatment periods are required (i.e., $g-t-1 \geq 2$).

2C.3 Estimation

Once the transformed outcome variables are constructed, the estimation procedure follows the same steps as Procedure 3.1, 4.1, and 5.1.

For identification, we dynamically define the control group based on whether pre-treatment periods t for group g precedes or follows the first treatment period S. Specifically, if $t < S$, the control group includes all not-yet-treated and never-treated units. If $t \geq S$, the control group consists only of units treated at $t + 1$ or later, as well as never-treated units. This staggered adoption rule ensures that the comparison group remains untreated at each point of comparison.

The weighted ATTs in the pre-treatment periods capture the pre-treatment effects. Under the no anticipation assumption, we expect $WATT(r)$ for $r < 0$ to be approximately zero. Significant deviations may indicate dynamic selection or violations of the identifying assumptions. When using demeaning transformations, the pre-treatment analysis applies to periods $r \leq -2$, since the transformation anchors outcomes at $t = g - 1$ (i.e., $r = -1$). Similarly, with detrending transformations, the analysis applies to periods $r \leq -3$.

APPENDIX 2D

ESTIMATION RESULTS

In this section, Table 2D.1 reports the descriptive statistics of variables used in our analysis.

Table 2D.1 Descriptive Statistics of variables

Variable	Obs	Mean	Std.	Min	Max
log(Retail Emp)	29,440	7.754502	1.281587		
log(Wholesale Emp)	29,440	6.413699	1.482646		
Share of Population Poverty (above)	29,440	.8470385	.0619999		
Share of Population in Manufacture	29,440	.0998018	.0501518		
Share of Population Graduate High School	29,440	.092258	.0256816		
Treated Cohort				1986	1999
Counties	1280				

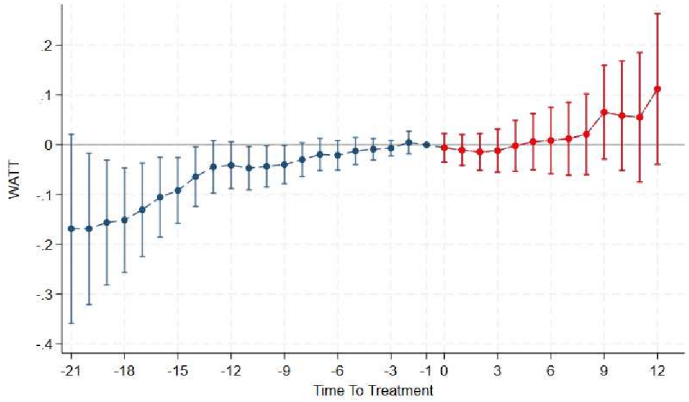
Table 2D.2 presents estimation results, corresponding to the visual representation in Figure 2.2; the retail labor market case. We estimate standard errors of the weighted ATTs with bootstrap, repeating 100 times. The first three columns display estimates of weighted ATTs from the ETWFE (Wooldridge, 2021), CS (2021), and our rolling IPWRA estimator (without unit-specific detrending), each of which is biased under the violation of parallel trend assumption. The fourth and fifth columns present the point estimates from our rolling RA and rolling IPWRA estimators, which appropriately account for potential heterogeneous linear trends.

Table 2D.2 Effects of Walmart Opening on log(Retail employment)

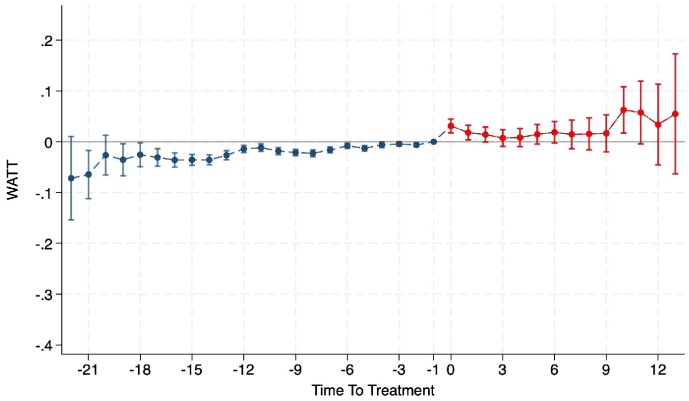
							Heterogeneous Trend			
	ETWFE		CS (2021)		Rolling IPWRA		Rolling RA		Rolling IPWRA	
r	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE
0	0.041	(0.006)	0.023	(0.003)	0.018	(0.004)	0.002	(0.004)	0.007	(0.004)
1	0.073	(0.007)	0.054	(0.004)	0.045	(0.004)	0.035	(0.005)	0.032	(0.005)
2	0.073	(0.008)	0.054	(0.005)	0.038	(0.004)	0.029	(0.006)	0.025	(0.006)
3	0.075	(0.009)	0.055	(0.006)	0.032	(0.004)	0.014	(0.007)	0.021	(0.007)
4	0.081	(0.01)	0.059	(0.007)	0.031	(0.004)	0.019	(0.009)	0.018	(0.009)
5	0.091	(0.011)	0.067	(0.008)	0.036	(0.005)	0.015	(0.01)	0.017	(0.01)
6	0.101	(0.012)	0.077	(0.009)	0.040	(0.005)	0.022	(0.012)	0.019	(0.012)
7	0.119	(0.013)	0.096	(0.01)	0.054	(0.006)	0.005	(0.013)	0.036	(0.013)
8	0.132	(0.014)	0.110	(0.011)	0.062	(0.008)	0.016	(0.016)	0.041	(0.016)
9	0.137	(0.016)	0.120	(0.013)	0.063	(0.01)	0.015	(0.019)	0.041	(0.019)
10	0.158	(0.019)	0.138	(0.015)	0.081	(0.013)	0.032	(0.023)	0.037	(0.023)
11	0.166	(0.023)	0.146	(0.018)	0.083	(0.018)	0.007	(0.031)	0.018	(0.03)
12	0.167	(0.03)	0.153	(0.023)	0.080	(0.026)	-0.008	(0.036)	0.017	(0.036)
13	0.206	(0.043)	0.191	(0.032)	0.107	(0.039)	0.046	(0.054)	0.047	(0.053)

2D.1 Wholesale Employment

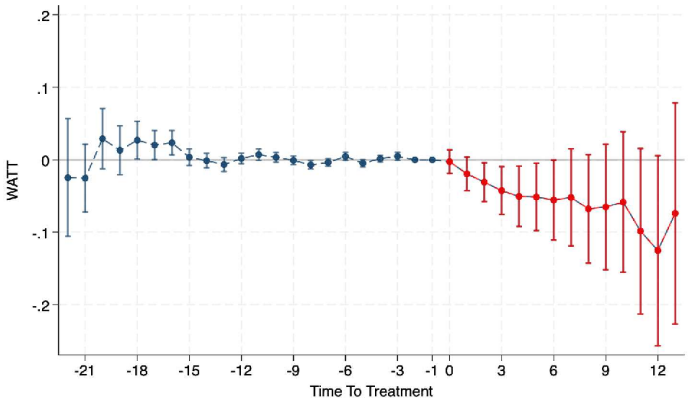
In this section, we report the estimation results for the county-level wholesale employment level. Figure 2D.1 presents the estimates from CS (2021) approach and our rolling IPWRA estimators.



Panel (a) CS Approach



Panel (b) Rolling IPWRA with unit-specific demeaning



Panel (c) Rolling IPWRA with unit-specific detrending

Figure 2D.1 Effects of Walmart Opening on log(Wholesale Employment)

In the top panels of Figure 2D.1, the weighted ATT estimates from the CS (2021) approach and the rolling IPWRA estimator (without unit-specific detrending) mirror the retail employment results, showing an upward pre-treatment trend.

In contrast, the panel (c) of Figure 2D.1 displays almost flat pre-treatment trends, indicating no significant pre-treatment effects. Additionally, it is noteworthy that the effects of Walmart entry on wholesale employment levels turn negative when the rolling IPWRA estimator is applied to the proposed dataset after removing county-specific linear trends. This reversal in the sign of the coefficients aligns with the findings of Brown and Butts (2023), who use factor models to relax the parallel trends assumption.

Table 2D.3 shows estimates in line with Figure 2D.1.

Table 2D.3 Effects of Walmart Opening on log (Wholesale employment)

							Heterogeneous Trend			
	ETWFE		CS (2021)		Rolling IPWRA		Rolling RA		Rolling IPWRA	
r	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE	ATT(r)	SE
0	0.041	(0.011)	0.008	(0.007)	0.031	(0.007)	-0.009	(0.008)	-0.003	(0.008)
1	0.032	(0.012)	-0.002	(0.008)	0.018	(0.007)	-0.018	(0.012)	-0.019	(0.012)
2	0.030	(0.014)	-0.005	(0.011)	0.014	(0.007)	-0.035	(0.013)	-0.031	(0.014)
3	0.027	(0.016)	-0.009	(0.013)	0.007	(0.008)	-0.052	(0.016)	-0.043	(0.017)
4	0.031	(0.018)	-0.007	(0.015)	0.008	(0.009)	-0.021	(0.021)	-0.051	(0.021)
5	0.039	(0.02)	0.004	(0.019)	0.015	(0.01)	-0.071	(0.023)	-0.051	(0.024)
6	0.047	(0.022)	0.013	(0.021)	0.019	(0.011)	-0.038	(0.027)	-0.055	(0.028)
7	0.047	(0.024)	0.016	(0.024)	0.015	(0.014)	-0.096	(0.033)	-0.052	(0.034)
8	0.052	(0.026)	0.019	(0.026)	0.015	(0.016)	-0.043	(0.037)	-0.068	(0.038)
9	0.058	(0.029)	0.028	(0.028)	0.016	(0.019)	-0.120	(0.043)	-0.065	(0.044)
10	0.109	(0.034)	0.073	(0.034)	0.063	(0.023)	-0.045	(0.049)	-0.058	(0.049)
11	0.113	(0.04)	0.062	(0.037)	0.057	(0.031)	-0.149	(0.059)	-0.099	(0.058)
12	0.098	(0.049)	0.057	(0.043)	0.034	(0.04)	-0.239	(0.067)	-0.125	(0.067)
13	0.130	(0.062)	0.112	(0.051)	0.055	(0.06)	-0.219	(0.075)	-0.074	(0.078)

Our rolling estimators, with county-level detrending, account for heterogeneous linear trends in a straightforward manner, allowing for the application of various treatment effect estimators. This underscores the importance of capturing variations in trends across different counties—trends that may not be captured by observed covariates but can significantly influence estimation results. Estimators that do not explicitly account for heterogeneous trends yield substantially different results, which are less convincing as causal estimates.

CHAPTER 3

SIMPLE APPROACHES TO INFERENCE WITH DIFFERENCE-IN-DIFFERENCES ESTIMATORS WITH SMALL CROSS-SECTIONAL SAMPLE SIZES

(CO-AUTHORED WITH JEFFREY M. WOOLDRIDGE)

3.1 Introduction

Difference-in-differences methods with panel data have a long history for evaluating policy interventions. In the case of common intervention timing, various methods are available for allowing treatment effects to vary by treatment period, as well as by control variables. Wooldridge (2021, 2025) shows how to estimate flexible linear models by pooled OLS or fixed effects – they are equivalent – that allows straightforward inference, provided there are enough cross-sectional units and the data are independently sampled across units. With a relatively small time series sample size (T) and large enough numbers of control (N_0) and treated units (N_1), standard errors robust to arbitrary serial correlation and heteroskedasticity are readily available. Other methods – such as the Callaway and Sant'Anna (2021) long differencing methods – also rely on large cross-sectional sample sizes for inference.

As shown in Lee and Wooldridge (2023) [LW (2023)], regression-based difference-in-differences (DID) estimators can be obtained via cross-sectional regressions after simple time-series transformations of the data at the unit level. This characterization of DiD estimators permits application of many widely used treatment effects estimators in the DiD setting, including matching, doubly robust estimators using the propensity score, and even machine learning causal methods. Here we use the representations in LW (2023) to obtain simple inference when the usual inference obtained from clustering at the individual unit may be problematical.

Our approach in the current paper combines the algebraic equivalence in LW (2023) and an idea from Donald and Lang (2007), who propose collapsing individual-level data to aggregated groups when the assignment of an intervention is at the group level. Here, the collapsing of the

The co-author has approved that the co-authored chapter is included. The co-author's contact: Jeffrey M. Wooldridge, Department of Economics, 486 W. Circle Drive, 110 Marshall-Adams Hall, Michigan State University, East Lansing, MI 48824-1038. Email: wooldri1@msu.edu

data is at the unit level across time in order to exploit the equivalences of panel data DiD estimators and those based on cross-sectional regressions. In addition to the standard DiD estimators, the approach extends easily to allow for unit-specific trends. Also, time-constant control variables can be added if the cross-sectional section sample sizes are large enough to avoid degeneracies.

The method can also be useful in cases where N_0 and N_1 (the sizes of the control and treatment groups, respectively) are not particularly small but where the number of time series is reasonably large. In such cases, using the panel structure and clustering by cross-sectional unit to account for serial dependence can cause distortions in the inference. The approach we suggest here requires no modification for large T. In fact, because inference is based on a cross-sectional regression, we need not worry about strong dependence in the time series dimension.

Recently, Simonsohn (2021), in commenting on Young (2019), argues that a particular version of the heteroskedasticity-robust standard error, proposed by Davidson, MacKinnon et al. (1993) (and labeled "hc3" in the popular Stata statistical package), can produce satisfactory results even in Young's setting with $N_0 = 18$ and $N_1 = 2$. Therefore, in some cases one might feel comfortable using heteroskedasticity-robust inference.

A popular method for studying interventions with a small number of treated units (with one being the most common) is the synthetic control method (SCM), pioneered by Abadie et al. (2010), where pre-intervention data are used to obtain a weighted average of "donor" control units to obtain a single synthetic control. More recently, Arkhangelsky, Athey, Hirshberg, Imbens and Wager (2021) propose a unification of traditional DiD and SC methods, called *Synthetic difference-in-differences* (SDiD). The SDiD approach allows more flexibility than either DiD or SC, and inference methods are available. However, SDiD assume that the series are weakly dependent – often called "integrated of order zero," or $I(0)$ – and that N_0, T_0, and T_1 are suitably "large." The authors also impose normality in obtaining inference, although it is not clear how important that is in practice.

Compared with the SDiD approach, the method we propose in this paper has some advantages and disadvantages. The advantages are that one need not have a very large control group nor a large number of time periods. Because inference is based on a cross-sectional regression under the

assumption of the classical linear model assumptions, there are minimal restrictions on N_0, N_1, T_0, and T_1. Moreover, we can use a cross-sectional regression for each treated time period, allowing for different estimated effects along with confidence intervals. We can also aggregate to estimate a single effect, which is what the SDiD approach produces. The SDiD approach does not explicitly allow for heterogeneous trends, something that is easily accommodated using the exact approach in this paper.

The downside to the cross-sectional approach here is that it can be be more biased than SDiD when differences in pre-trends are complicated, although SDiD is more biased for simple heterogeneous trends. Our cross-sectional approach also can be considerably less efficient than SDiD. We provide evidence of that via simulations. Nevertheless, sometimes SDiD will be less efficient. Moreover, the popular SDiD packages that produce the estimators and standard errors can produce confidence intervals that are too optimistic.

We view the methods proposed in this paper as a complement to SDiD, providing another tool for estimating treatment effects when the number of treated or control units is small.

We start with the common timing case in Section 3.2, first summarize the algebraic equivalences that allow one to apply cross-sectional regression to obtain the DiD estimates(s) and standard errors. Section 3.3 extends the framework to accommodate heterogeneous trends and seasonal effects, demonstrating how to implement our approach in these more complex settings. In Section 3.4, we compare our method to the Synthetic Control (SC) and Synthetic Difference-in-Differences (SDiD) approaches, highlighting differences in their underlying assumptions.

Section 3.5 presents simulation results evaluating the performance of these estimators. We find that SC and SDiD, especially the former, can exhibit substantially more bias than removing heterogeneous trends. In Section 3.6, we revisit the California smoking restrictions passed in 1989, using the data in Abadie, Diamond, and Hainmueller (2010). Section 3.7 addresses the staggered rollout case and applies our method to measure the effects of castle laws on homicides. We conclude in Section 3.8.

3.2 The Common Timing Case

3.2.1 Estimating a Single ATT

Initially, the setting is that a total of T time periods are available, and an intervention occurs at time S, $S \in \{2, ..., T\}$. A cross-sectional unit, i, is either subjected to the intervention, which remains in place periods S through T, or it is a control unit. Treatment status is indicated by the binary variable

$$\begin{aligned} D_i &= 0 \text{ if a control unit} \\ D_i &= 1 \text{ if a treated unit} \end{aligned} \tag{3.1}$$

The time-varying treatment indicator is

$$W_{it} = D_i \cdot post_t \tag{3.2}$$

where $post_t = 1$ if $t \in \{S, S+1, ..., T\}$ and zero otherwise. In what follows, we assume that the data draws are independent and identically distributed across i but allow general dependence and changing distributions across t. Late, we provide a discussion of how one can accomodate clustering or spatial correlation if the cross-sectional sample sizes are reasonably large.

The simple DiD estimator can be obtained as the coefficient $\widehat{\tau}_{DD}$ on W_{it} from the pooled regression

$$Y_{it} \text{ on } 1, D_i, post_t, W_{it}, \quad i = 1, ..., N; \; t = 1, ..., T \tag{3.3}$$

Equivalently, one can replace D_i and $post_t$ with a full set of unit and time period fixed effects, resulting in the popular two-way fixed effects (TWFE) estimator. [The equivalence follows from Wooldridge (2021, 2010).] When N is small (or N_1 is small), inference using the TWFE estimator is tricky because one should allow serial correlation in the underlying time-varying errors, U_{it}, that are implicit in the equation. The usual method of clustering by unit i generally results in poor performance with small N (or with small N_0 or N_1).

As is fairly well known, and recently shown in Lee and Wooldridge (2023) in cases with covariates, $\widehat{\tau}_{DD}$ also can be obtained as follows. First, for each unit i, create

$$\begin{aligned} \Delta\bar{Y}_i &\equiv \frac{1}{(T-S+1)}\sum_{t=S}^{T} Y_{it} - \frac{1}{(S-1)}\sum_{t=1}^{S-1} Y_{it} \\ &\equiv \bar{Y}_{i,post} - \bar{Y}_{i,pre}, \end{aligned} \tag{3.4}$$

which is a simple transformation of the data within each unit. This results in a simple cross-sectional data set $\left\{\left(\Delta\bar{Y}_i, D_i\right) : i = 1, ..., N\right\}$, where N can be large or small. In the second step, $\widehat{\tau}_{DD}$ is obtained as the coefficient on D_i from the simple regression

$$\Delta\bar{Y}_i \text{ on } 1,\ D_i,\ \ i = 1, ..., N. \tag{3.5}$$

Of course, this regression leads to the formula for a difference in means:

$$\begin{aligned} \widehat{\tau}_{DD} &= N_1^{-1}\sum_{i=1}^{N} D_i \cdot \Delta\bar{Y}_i - N_0^{-1}\sum_{i=1}^{N}(1-D_i)\cdot\Delta\bar{Y}_i \\ &= \Delta\bar{Y}_{treat} - \Delta\bar{Y}_{control}. \end{aligned} \tag{3.6}$$

The benefit of the previous characterization of the DiD estimator is that we can think of a simple underlying model,

$$\Delta\bar{Y}_i = \alpha + \tau D_i + U_i \tag{3.7}$$

$$E\left(U_i|D_i\right) = 0, \tag{3.8}$$

where the zero conditional mean assumption holds when the difference-in-differences design identifies the average treatment effect on the treated, τ. If N is reasonably large, and we have several treated and control units, we can rely on asymptotic analysis and justify a heteroskedasticity-robust standard error for $\widehat{\tau}_{DD}$ for computing a confidence interval for τ. But when N is small, or, say, N_0 is large but N_1 is small, we cannot rely on asymptotics. Nevertheless, it is possible that (3.7) satisfies the classical linear model (CLM) assumptions:

$$U_i|D_i \sim Normal\left(0, \sigma_U^2\right) \tag{3.9}$$

Under (3.7) and (3.9), conditional on D_i, $\widehat{\tau}_{DD}$ has an exact normal distribution, and

$$\frac{\left(\widehat{\tau}_{DD} - \tau\right)}{\text{se}\left(\widehat{\tau}_{DD}\right)} \sim \mathcal{T}_{N-2} \tag{3.10}$$

This result means that we can obtain exact tests of any null hypothesis, such as $H_0 : \tau = 0$. Moreover, the usual confidence, obtained using percentiles of the $\mathcal{T}_{N-2}$ distribution, have exact coverage. Relatedly, Hagemann (2025) assumes normality in the context of testing for a treatment effect when a single cluster is treated with many units within the cluster, allowing for different variances in the distributions of the control and treated units. Here, we do not require any particular number of time periods (units within a cluster) provided the normality assumption (3.9) holds.

In looking at the expression for $\widehat{\tau}_{DD}$ in (3.6), it is clear that asymptotic theory will not apply when, say, N_0 is very large if N_1 is still small: we need the central limit theorem to apply to both terms in (3.6), and it will not generally be a good approximation to the distribution if N_1 is small. The same is true if N_0 is small.

Assuming (3.9) holds, the approach works in any setting with $N_0 \geq 1$, $N_1 \geq 1$, and $N = N_0 + N_1 \geq 3$. In particular, we can have a single treated unit, $N_1 = 1$, and many or few control units. When there is only a single treated unit, the t-statistic $\widehat{\tau}_{DD}/\text{se}\left(\widehat{\tau}_{DD}\right)$ is known as the "studentized residual," which plays a role in outlier analysis. See, for example, Wooldridge (2020, Section 9.5). Viewing the t statistic for $\widehat{\tau}_{DD}$ as an outlier diagnostic makes intuitive sense: we are trying to determine whether the single treated unit, with $D_i = 1$, is an outlier compared with the control units.

The approach just outlined is essentially that taken by Donald and Lang (2007), but here we apply it to panel data with either a short or long stretch of time. If T_0 and T_1 are reasonably large, the normality in (3.9) may be an acceptable assumption because, if the data are weakly dependent across time, the central limit theorem implies that $\bar{Y}_{i,post}$ and $\bar{Y}_{i,pre}$ are approximately normal. Moreover, because we only require conditional normality of $\Delta\bar{Y}_i = \bar{Y}_{i,post} - \bar{Y}_{i,pre}$, strong dependence (such as a time-constant unobserved effect) is eliminated by differencing the post-intervention and pre-intervention averages. Underlying unit root processes also need not cause $\Delta\bar{Y}_i$ to deviate substantially from normality.

In addition to (technically) maintaining exact normality, another drawback of the exact inference approach is that it maintains homoskedasticity – that is, $Var\left(U_i|D_i\right) = Var\left(U_i\right)$. Nevertheless,

recent simulations by Simonsohn (2021) are promising in that one version of the heteroskedasticity-robust standard error, often called "hc3" in the literature, can work well even with a small number of treated and control units. Thus, one may feel comfortable using heteroskedasticity-robust inference except in cases with a very small number of treated and/or control units.

In addition, a notable advantage of our method is its compatibility with randomization inference (RI), which does not rely on the normality assumption. RI enables the computation of exact p-values for testing the null hypothesis that all treatment effects are zero. The p-value is calculated as the proportion of permutation-based test statistics that are as extreme or more extreme than the observed test statistic. Specifically, if the number of such permutations is c, and the total number of permutations is N, then the two-sided randomization p-value is defined as c/N. In practice, the user-written Stata command `ritest` can be employed to implement this procedure and obtain exact p-values. See, for example, Heß (2017).

There is another characterization of $\widehat{\tau}_{DD}$ that will be useful when we allow unit-specific pre-trends in Section 3.3.

Procedure 2.1 (Unit-Specific Demeaning):

1. Namely, think of obtaining $\bar{Y}_{i,pre}$, for each i, using the pre-intervention regression:

$$Y_{it} \text{ on } 1, t = 1, \ldots, S-1, \tag{3.11}$$

where the coefficient on unit (the only coefficient) gives $\bar{Y}_{i,pre}$.

2. Next, in each post-intervention period, we obtain out-of-sample residuals, or prediction errors:

$$\dot{Y}_{it} = Y_{it} - \bar{Y}_{i,pre} = Y_{it} - \frac{1}{(S-1)} \sum_{r=1}^{S-1} Y_{ir}, \ t = S, \ldots, T \tag{3.12}$$

3. It is easy to see that, when these out-of-sample residuals are average, we obtain the same regressand as in (3.5):

$$\overline{\dot{Y}}_i \equiv \frac{1}{(T-S+1)} \sum_{t=S}^{T} \dot{Y}_{it} = \bar{Y}_{i,post} - \bar{Y}_{i,pre} = \Delta \bar{Y}_i$$

4. Obtain an average effect, $\widehat{\tau}_{DM}$, and its standard error and confidence interval, from

$$\overline{\dot{Y}}_i \text{ on } 1, D_i, i = 1, ..., N. \ \square \tag{3.13}$$

3.2.2 Adding Control Variables

As discussed in Lee and Wooldridge (2023), Procedure 2.1 uncovers the ATT under a no anticipation (NA) assumption and a parallel trends (PT) assumption. With $Y_{it}(0)$ and $Y_{it}(1)$ denoting the potential outcomes for unit i in period t, the weakest version of NA is

$$E\left[Y_{it}\left(1\right) - Y_{it}\left(0\right)|D_i = 1\right] = 0, \quad t = 1, \ldots, S-1, \tag{3.14}$$

so the potential outcomes are the same, on average for the treated units. Sufficient, of course, is $Y_{it}(1) = Y_{it}(0), \quad t = 1, \ldots, S-1$.

The parallel trends assumption is that, for all $t = 2, \ldots, T$,

$$E\left[Y_{it}\left(0\right) - Y_{i1}\left(0\right)|D_i\right] = E\left[Y_{it}\left(0\right) - Y_{i1}\left(0\right)\right] \equiv \delta_t, \tag{3.15}$$

where δ_t is a constant that is unrestricted over time. This assumption allows an unrestricted trend in the control state, but that trend must be the same across control and treated units. In effect, it allows treatment assignment, D_i, to be correlated with the level, say $Y_{i1}(0)$, but it rules out cases where the assignment is based on differential trends in the control state.

As shown in Lee and Wooldridge (2023), the PT assumption implies that, for some constant α,

$$E\left[\Delta\bar{Y}_i\left(0\right)|D_i\right] = \alpha, \tag{3.16}$$

where $\Delta\bar{Y}_i\left(0\right) = \bar{Y}_{i,post}\left(0\right) - \bar{Y}_{i,pre}\left(0\right)$ is the same transformation appearing in (3.4) but for the potential outcomes. Equation (3.16) means that, $\Delta\bar{Y}_i\left(0\right)$ is mean independent of D_i; as discussed in LW (2023), along with the NA assumption this implies that τ is identified and $\widehat{\tau}_{DD}$ is (conditionally) unbiased and consistent. Here, we are relying on unbiasedness because of the small N_1 or small N_0.

LW (2023) also show that the PT assumption is easily relaxed by conditioning on controls. Given time-constant controls $\mathbf{X}_i$, a $1 \times K$ vector, the conditional PT assumption is

$$E\left[Y_{it}\left(0\right) - Y_{i1}\left(0\right)|D_i, \mathbf{X}_i\right] = E\left[Y_{it}\left(0\right) - Y_{i1}\left(0\right)|\mathbf{X}_i\right] \equiv \alpha_t + \mathbf{X}_i\beta_t, \tag{3.17}$$

where we have imposed linearity because we have little choice in a small-N setting. Assumption (3.17) is an unconfoundedness assumption on D_i once we condition on $\mathbf{X}_i$, but it is in terms of the

differences $Y_{it}(0) - Y_{i1}(0)$, rather than the levels $Y_{it}(0)$. (This is what gives the DiD procedure its main advantage over cross-sectional regression.) In terms of the cross-sectional regression, it is easy to adapt LW (2023) to see that we should add $\mathbf{X}_i$ to the equation. To use exact inference, we would write

$$\Delta\bar{Y}_i = \alpha + \tau D_i + \mathbf{X}_i\beta + U_i \tag{3.18}$$

$$U_i | D_i, \mathbf{X}_i \sim Normal\left(0, \sigma_U^2\right) \tag{3.19}$$

Provided $N > K + 2$, (3.18) can be estimated by OLS, and, under the conditional normality assumption, the t statistic has an exact $\mathcal{T}_{N-K-2}$ distribution and the corresponding confidence intervals are exact. Again, it is intriguing to think that, even without N being too large, we might use heteroskedasticity-robust inference.

When N_0 and N_1 are both sufficiently large, and the support of $\mathbf{X}$ for $D = 1$ is contained in that for $D = 0$, full regression is preferred on theoretical grounds. Namely, $\widehat{\tau}_{DD}$ is the coefficient on D_i from the regression that includes interactions:

$$\Delta\bar{Y}_i \text{ on } 1,\ D_i,\ \mathbf{X}_i,\ D_i \cdot \left(\mathbf{X}_i - \bar{\mathbf{X}}_1\right),\quad i = 1, \ldots, N,$$

which is identical to running separate regressions for $D_i = 0$ and $D_i = 1$. Here, $\bar{\mathbf{X}}_1$ is the average of the covariates over $D_i = 1$. Unfortunately, this cannot be done with small N_1, as it requires $N_0 > K + 1$ and $N_1 > K + 1$.

3.2.3 Estimating Separate Effects for Each Treated Period

Rather than estimate a single ATT, it is little additional effort to estimate ATTs separately in each treated period. Define $\dot{Y}_{it}$ as in (3.12), and this is used as the dependent variable in period t. Then $\widehat{\tau}_{t,DD}$ for $t = S, \ldots, T$ is obtained from

$$\dot{Y}_{it} \text{ on } 1, D_i, i = 1, \ldots, N \tag{3.20}$$

In order to conduct small N (or small N_1 inference), we cannot rely on the central limit theorem in the post-treatment period. Also, we cannot easily obtain standard errors for linear combinations for the $\widehat{\tau}_{t,DD}$ because they will be dependence in complicated ways due to serial correlation. (Unless

we make the ideal assumptions that rule out serial correlation.) Nevertheless, we can obtain a CI for each $\tau_t = E\left[Y_{it}\left(1\right) - Y_{it}\left(0\right)|D_i = 1\right]$.

The coefficients one obtains from (3.20) are identical to running a standard panel DiD using the control and treatments for the time periods $\{1, 2, \ldots, S-1, t\}$ where $S \leq t \leq T$. That is, we include all pre-treatment variables.

Again, with N large enough, we can add controls $\mathbf{X}_i$. Note that, like the treatment effects themselves, the intercept and slope parameters on $\mathbf{X}_i$ vary freely over the treatment periods.

3.2.4 Using Different Pre-Treatment Periods

The suggestion in the previous subsection is to use all pre-treatment periods in computing the pre-treatment averages. Under NA and parallel trends, that approach uses the most information. Nevertheless, one might want to ignore information in the pre-treatment time periods. The standard event-study approach uses only the period just before the treatment. This results in the transformation

$$\mathring{Y}_{it} = Y_{it} - Y_{i,S-1},\, t = S, \ldots, T, \tag{3.21}$$

which can then be used in place of $\dot{Y}_{it}$ in (3.20). This is the transformation used by Callaway and Sant'Anna (2021). Of course, one can use any average of $\{Y_{it} : t = 1, \ldots, S-1\}$ – even a weighted average – and the method would be still valid. If there is concern about violation of no anticipation, the average $\bar{Y}_{i,pre}$ could be replaced with

$$\bar{Y}_{i,S_0} = \frac{1}{S_0}\sum_{r=1}^{S_0} Y_{ir} \tag{3.22}$$

where $S_0 < S-1$. A long-difference also can be used: $Y_{it} - Y_{i,S_0}$.

3.3 Heterogeneous Trends and Seasonality in the Common Timing Case

LW (2023) show that, when the units have unit-specific trends, these can be removed before applying a modification of Procedure 2.1. For concreteness, the following procedure removes linear trends; it is easily modified to remove higher-order unit-specific trends.

Procedure 3.1 (Unit-Specific Detrending):

1. For each i, obtain $\widehat{A}_i$, $\widehat{B}_i$ from the pre-treatment periods by regressing on a constant and time:

$$Y_{it} \text{ on } 1, t, \ t = 1, \ldots, S-1 \tag{3.23}$$

2. For the post-treatment periods, remove the pre-treatment trends:

$$\ddot{Y}_{it} = Y_{it} - \widehat{Y}_{it} \equiv Y_{it} - \widehat{A}_i - \widehat{B}_i \cdot t, t = S, \ldots, T, \tag{3.24}$$

where $\widehat{Y}_{it} \equiv \widehat{A}_i + \widehat{B}_i \cdot t$ is the projected value of Y_{it} in a treated period using a trend obtained from pre-treatment periods.

3. For each unit, average the adjusted outcomes:

$$\overline{\ddot{Y}}_i \equiv \frac{1}{(T-S+1)} \sum_{t=S}^{T} \ddot{Y}_{it} \tag{3.25}$$

4. Obtain an average effect, $\widehat{\tau}_{DT}$, and its standard error and confidence interval, from

$$\overline{\ddot{Y}}_i \text{ on } 1, D_i, i = 1, ..., N. \ \square \tag{3.26}$$

We use "DT" to denote that $\widehat{\tau}_{DT}$ is obtained from a detrending procedure. As before, even with $N_1 = 1$, we can perform inference under

$$\begin{aligned} \overline{\ddot{Y}}_i &= \alpha + \tau_{DT} \cdot D_i + U_i \\ U_i | D_i &\sim Normal\left(0, \sigma^2\right) \end{aligned}$$

With large enough N we could include covariates, as in (3.17), which would allow for confounded assignment even after removing unit-specific linear trends.

Because Procedure 3.1 uses unit-specific removal trends, it may initially appear that the final regression used to obtain $\widehat{\tau}_{DT}$ (or $\widehat{\tau}_{DD}$ in Procedure 2.1) might suffer from the so-called "incidental parameters" problem. This is not the case, as we are simply transforming each unit using its own time series data. That is, $\Delta\bar{Y}_i$ and $\overline{\ddot{Y}}_i$ are just linear functions of $\{Y_{it} : t = 1, 2, \ldots, T\}$. Under independence across i, the result is cross-sectional data where the observations are independent. That allows us to at least argue for classical linear model analysis when N is small. Without

control variables, Lee and Wooldridge (2023) show that a sufficient condition is the independence of the treatment indicator from transformed outcome variable– that is, the outcome after removing unit-specific pre-treatment averages or trends.

As in the case without trends, we can easily estimate separate effects for each t: for each $t \in \{S, ..., T\}$ obtain $\widehat{\tau}_{t,DT}$ from the sequence of regressions

$$\ddot{Y}_{it} \text{ on } 1, D_i, \ i = 1, ..., N$$

and possibly include $\mathbf{X}_i$. As before, under the CLM assumptions we can perform inference on the $\widehat{\tau}_{t,DT}$. The other strategies discussed in Section 3.2, such as leaving a gap before the intervention if one is concerned about anticipation, apply here as well.

Rather than using all of the time periods to remove a trend, one could use second differencing, say,

$$\ddot{Y}_{it} \equiv \left(Y_{it} - Y_{i,S-1}\right) - \left(Y_{i,S-1} - Y_{iR}\right),$$

where $1 \leq R < S - 1$. The resulting estimator is a difference-in-difference-in-differences estimator. Such an estimator does not exploit averaging across pre- and post-treatment periods, likely making it more sensitive to violations of the normality assumption.

When Y_{it} is the log of a positive outcome, a linear time trend is attractive. Nevertheless, with large enough S, we can make the pre-trend removal more flexible; most obviously by including higher powers, such as t^2 and t^3. However, removing too much of the variation in the outcome may make it difficult to detect effects of the intervention.

With quarterly or monthly data, and maybe even with weekly data, it might make sense to remove seasonality at the unit level (perhaps in addition to a trend). In the first step of Procedure 2.1 or 3.1, we would simply include quarterly, monthly, or weekly dummy variables and obtain the deseasonalized/detrended outcomes. After that, the procedure is exactly the same.

3.4 Comparison with Synthetic Control Methods

With a single treated unit, ADH (2010) propose algorithms to use pre-treatment data to obtain weights on "donors" to create a synthetic control (SC) for the treated unit. The weights are chosen

by an optimization algorithm so that all weights are nonnegative and sum to unity. In early SC applications, inference is essentially visual, relying on taking each control unit as a placebo treated unit and comparing the post-treatment gaps with those of the actual treated unit.

More recently, Arkhangelsky et al. (2021) study the SC methods, and extensions, when the target parameter is the post-treatment time average of the ATTs – rather than estimating a different parameter in each time period. They propose the Synthetic difference-in-differences (SDiD) estimator, which can be interpreted as a weighted two-way fixed effects (TWFE) estimator. Specifically, SDiD assigns unit weights to balance the pre-treatment trends of treated and control units, and time weights to balance pre- and post-treatment periods within control units.

These weights are obtained by solving regularized optimization problems that minimize discrepancies in pre-treatment outcome trajectories, thereby improving robustness to violations of the parallel trends assumption and enhancing efficiency. This approach allows SDiD to blend the strengths of both SC and DID while retaining valid large-sample inference guarantees.

When it comes to staggered interventions, Arkhangelsky et al. (2021) suggest splitting the sample by adoption date into subsets and applying SDiD method separately to each subset. See, for example, AAHIW (Appendix: Staggered Adoption, 2021). For instance, if units 5 and 6 begin treatment in period 5, units 3 and 4 begin in period 3, and units 1 and 2 are never treated, then, the sample can be divided into to sub-samples: one including units $1, 2, 5$, and 6 and another including units $1, 2, 3$, and 4.

Inference for the SDiD estimator is developed under large N_0, large T asymptotics. The key assumptions include independence of the error terms across i, allowing for standard central limit arguments as the number of control units increases. Within each unit, the errors are permitted to exhibit serial correlation over time, provided they satisfy weak dependence conditions. It also allows arbitrary correlation across time within units while maintaining independence across units. Additionally, the validity of their placebo-based standard error estimation relies on approximate normality of the residuals.

Procedures 2.1 and 3.1 of our paper use relatively less sophisticated strategies in effectively

choosing a synthetic Control. However, it is important to remember that the chosen control is based on the outcome after the pre-treatment averages or pre-treatment trends have been removed. The average of the resulting residuals across all included control units effectively plays the role of the synthetic control for the residualized outcomes of the treated unit. In other words, the cross sectional average of the controls,

$$\left\{ N_0^{-1} \sum_{i=1}^{N_0} \dot{Y}_{it} : t = 1, ..., T \right\}, \tag{3.27}$$

is the synthetic control for the cross-sectional averages of the treated units,

$$\left\{ N_1^{-1} \sum_{i=N_0}^{N_0+N_1} \dot{Y}_{it} : t = 1, ..., T \right\}. \tag{3.28}$$

If we remove pre-treatment trends then $\ddot{Y}_{it}$ replaces $\dot{Y}_{it}$. The hope is that after transforming Y_{it} in the pre-treatment periods, the control residuals do a good job of tracking the treated residuals pre-treatment. When the procedure is successful, there should be close agreement in the two average residual series over the periods $t = 1, 2, \ldots, S-1$. The estimated treatment effects are obtained from the differences for $t \geq S$, and these are exactly the coefficients in (3.5) or (3.26).

3.5 Monte Carlo Simulations

In this section, we design simulations to assess the performance of our proposed estimator in the presence of treatment effect heterogeneity under the common timing case. Our goal is to understand how well the estimator recovers the average treatment effects under realistic data-generating processes characterized by serial correlation, unit-specific heterogeneity, and dynamic treatment effects.

In each simulation, we generate data for $N = 20$ units observed over $T = 20$ time periods, with treatment beginning at a common timing in period $t = 11$, resulting in 10 pre-treatment and 10 post-treatment periods for each unit.

3.5.1 Data Generating Process

This section outlines the specific steps used to generate the simulated data in detail.

Step 1. Unit characteristics

Each unit is endowed with two latent characteristics, unit-specific heterogeneity and unit-specific time trend slope: $c_i \sim \mathcal{N}(0, \sigma_c^2)$ and $g_i \sim \mathcal{N}(1, \sigma_g^2)$, respectively, where $\sigma_c = 2$ and $\sigma_g = 1$ in our simulation. These characteristics are constant across time.

Step 2. Idiosyncratic error term

We generate serially correlated errors u_{it} using an AR(1) process,

$$u_{i1} \sim N\left(0, \sqrt{\frac{2}{1-\rho^2}}\right),$$

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, 2), \rho = 0.75$$

This structure introduces persistent shocks in the outcome evolution.

Step 3. Potential outcomes

The potential outcome in the control state and treated state is defined as

$$Y_{it}(0) = \lambda_t \cdot fs_t - c_i + g_i t + u_{it},$$

$$Y_{it}(1) = Y_{it}(0) + \delta_t \cdot fs_t + \nu_{it}, \quad \nu_{it} \sim N(0, \sqrt{2}),$$

where fs_t is time dummy, λ_t represents time fixed effects and δ_t is time-varying treatment effects. We assume that $\delta_t = 0$ for $t < 11$. Then, the observed outcome is determined using a potential outcomes framework:

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

where $D_i \in \{0, 1\}$ is a binary indicator for treatment status. Treatment is assigned based on a logistic selection rule,

$$\Pr(D_i = 1) = \mathbb{I}\left(\alpha_0 - \alpha_1 \cdot c_i + \alpha_2 \cdot g_i + \epsilon_i > 0\right), \ \epsilon_i \sim Logistic(0, 1)$$

The average post-treatment effect is set to approximately 2 by construction.

The detailed settings for each scenario are summarized in Table 3.1. Specifically, the time fixed effects λ_t and time-varying treatment effects δ_t are held constant, while the parameters governing

the treatment assignment $\Pr(D_i = 1)$ vary across three scenarios. The probability of being treated decreases from Scenario 1 to Scenario 3.

Table 3.1 Simulation Setup Parameters Across Three Scenarios

Parameters	Scenario 1	Scenario2	Scenario 3
Treatment Rule $(\alpha_0, \alpha_1, \alpha_2)$	$(-1, -\frac{1}{3}, \frac{1}{4})$	$(-1.5, \frac{1}{3}, \frac{1}{4})$	$(-2, \frac{1}{3}, \frac{1}{4})$
Time Fixed Effects $(\lambda_1, \lambda_2, \ldots, \lambda_{20})$	(0, 0, 0, 0, 0.2, 0.6, 0.7, 0.8, 0.6, 0.9, 0.9, 1, 1.1, 1.3, 1.2, 1.5, 0.6, 1.4, 1.8, 1.9)		
Treatment Effects $(\delta_{11}, \ldots, \delta_{20})$	(1, 2, 3, 3, 3, 2, 2, 2, 1, 1)		
Note: only the treatment rule parameters vary across scenarios.			

3.5.2 Simulation Results

Table 3.2 presents the simulation results across three scenarios outlined in Table 3.1. Our proposed Detrending estimator, both in its standard error ($HC0$) and with heteroskedasticity-consistent standard errors ($HC3$), outperforms alternative methods such as Synthetic Control (SC) and Synthetic difference-in-differences (SDiD) methods.

Across all scenarios, the Detrending estimator achieves the lowest root mean squared error (RMSE), indicating superior finite-sample performance. Specifically, in Scenario 1 where the probability of treatment is higher ($P(D_i = 1) = 0.32$), the Detrending estimator exhibits a negligible bias and an RMSE of approximately 1.73, significantly outperforming SC and SDiD, which exhibit large biases (-0.375 and 0.392, respectively) and correspondingly higher RMSEs. Similar patterns are observed in Scenarios 2 and 3.

In particular, while SC and SDiD exhibit a notable gap between the empirical standard deviation (*SD*) of the estimates and the average estimated standard errors (*Avg SE*) reported within each replication, our Detrending estimator maintains a close alignment between the two, leading to more reliable inference and coverage rates close to the nominal 95%.

For example, in Scenario 1, the *SD* for SC is 1.73, while the average standard error is 2.66, indicating that the method tends to overestimate its own uncertainty and produce wider-than-necessary confidence intervals. This conservative estimation contributes to coverage rates that remain close to or slightly above the nominal 95%.

Table 3.2 Simulation Results Across Three Scenarios

Scenario 1. $P(D_i = 1) = 0.32$

	Average Effects	Bias	SD	RMSE	Coverage Rate	Avg SE
Sample AE	1.991					
Demeaning	3.905	1.914	5.10	5.445	0.94	4.96
Detrending	2.000	0.009	1.73	1.734	0.96	1.74
Detrending (hc3)	2.000	0.009	1.73	1.734	0.96	1.87
SC	1.616	-0.375	2.14	2.177	0.97	3.27
SDiD	2.383	0.392	1.77	1.808	0.96	2.56

Scenario 2. $P(D_i = 1) = 0.24$

	Average Effects	Bias	SD	RMSE	Coverage Rate	Avg SE
Sample AE	2.011					
Demeaning	4.264	2.253	5.67	6.101	0.93	5.48
Detrending	1.969	-0.042	1.89	1.892	0.95	1.91
Detrending (hc3)	1.969	-0.042	1.89	1.892	0.93	2.04
SC	1.622	-0.389	2.35	2.384	0.94	2.73
SDiD	2.395	0.384	1.89	1.925	0.95	2.25

Scenario 3. $P(D_i = 1) = 0.17$

	Average Effects	Bias	SD	RMSE	Coverage Rate	Avg SE
Sample ATT	1.996					
Demeaning	4.566	2.570	6.78	7.254	0.94	6.43
Detrending	2.161	0.165	2.37	2.380	0.95	2.26
Detrending (hc3)	2.161	0.165	2.37	2.380	0.91	2.60
SC	2.962	0.966	2.92	3.078	0.92	2.85
SDiD	2.615	0.619	2.35	2.435	0.94	2.33

Each simulation study has $1,000$ replications with $N = 20$

Theoretically, substantial bias would be expected to shift confidence intervals away from the true treatment effect and lower coverage rates. Although SC and SDiD exhibit noticeably larger biases compared to our Detrending method, in our simulations, the magnitude of these biases is not sufficiently large relative to their conservative standard errors to meaningfully reduce coverage. As a result, despite their higher bias, SC and SDiD still achieve coverage rates comparable to or slightly exceeding the nominal 95% level.

Moreover, in Scenario 3 where treatment assignment becomes rarer ($P(D_i = 1) = 0.17$), all estimators experience greater challenges, but our detrending method still maintains the smallest

RMSE and the most reliable inference properties compared to SC and SDiD.

Overall, these simulation results suggest that the detrending approach can offer substantial improvements in bias reduction, efficiency, and inference accuracy in scenarios with heterogeneous linear trends. However, its advantages are not universal. When a unit's pre-trends exhibit more complex patterns that cannot be adequately captured by linear detrending, our method will not always work better than existing alternatives.

3.6 Application to California Smoking Restrictions

We apply our methods to the problem analyzed in ADH (2010), estimating the effect of California's tobacco control program. The first year of the program is 1989, with 19 pre-treatment years (1970-1988) and 12 treatment years (1989-2000). We use the log of per capita cigarette sales as the outcome variable; ADH (2010) used the level of this variable, but the log seems more natural – partly to obtain a percentage effect and partly to make normality of the transformed outcome a better approximation. California is the only treated state. As in ADH (2010), we use $N_0 = 38$ potential control states after eliminating states that implemented some sort of anti-smoking program over the period.

As discussed in the previous section, SCM uses pre-treatment outcomes (and sometimes other variables) to create a synthetic control – in this case, a synthetic California. The weights for the 38 controls are chosen to make the SC as close to California as possible based on pre-treatment observables (primarily Y_{it}). These weights are then used to project out what would have been the untreated outcome for CA after the policy change; this is then compared with the actual outcome in post-intervention.

3.6.1 Results

Figure 3.1 shows the graphs of (3.27) and (3.28) using Procedure 2.1, allowing for separate effects in each treated period. The solid pink line represents the outcome for the treated unit (California), while the dashed blue line corresponds to the synthetic control group. The vertical dashed line marks the intervention year (1989).

It is evident that the tracking is not particularly good, with the average of the controls initially

being below the treated and then above the treated unit in 1980. Since it includes all of the controls, it is not surprising that perhaps just removing a pre-treatment average is not enough to make them similar to California prior to the intervention – even averaged across $N_0 = 38$. One possibility is to choose a subset of the controls that seem most similar to California. In effect, that is what the SCM does in a systematic way.

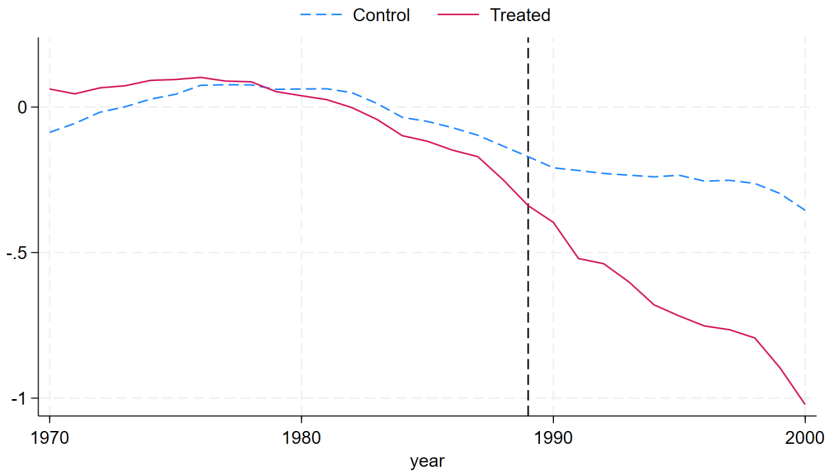


Figure 3.1 Removing Pre-Treatment Averages, All Controls

On the other hand, Figure 3.2 plots the average residuals for the controls against California after removing the state-specific trends, obtained using Procedure 3.1. This provide a much better fit up until the policy change.

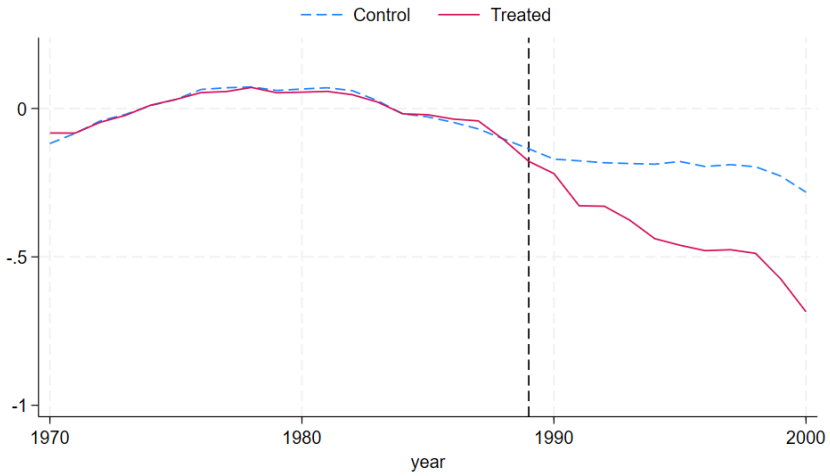


Figure 3.2 Removing Pre-Treatment Linear Trends, All Controls

Now, we can interpret the gaps after the intervention data as the effects of the intervention at different time horizons. Figure 3.3 shows the estimated gap between detrended California and the average of the 38 detrended controls. Prior to the intervention in 1989, the gap ranges from about -1.4% to 3.5%, with an average very close to zero. After the intervention, the treated unit exhibits a sharper decline compared to the synthetic control, indicating a potential negative effect of the policy.

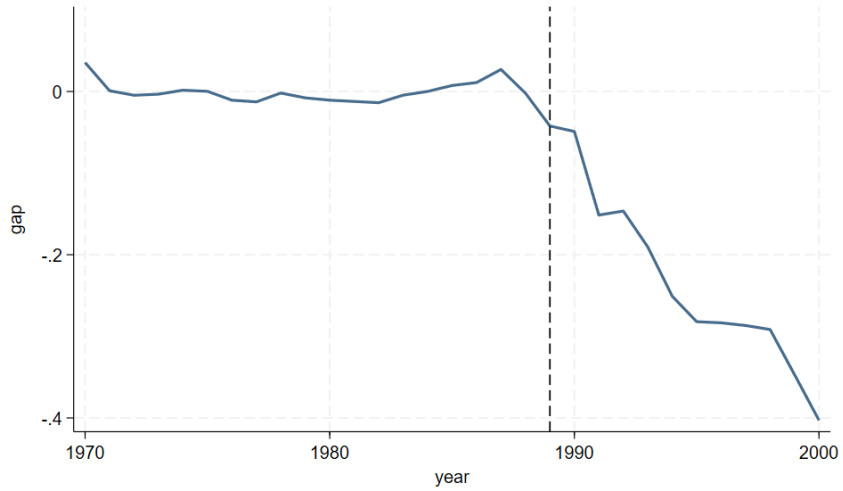


Figure 3.3 Gap Between California and Average of All Controls, State-Specific Linear Trend

Similarly, both the Synthetic Control and Synthetic DiD methods, using all 38 states as controls, display a similar pre-treatment trend between California and Synthetic California (Control), as visualized in Figure 3.4 and Figure 3.5, respectively.

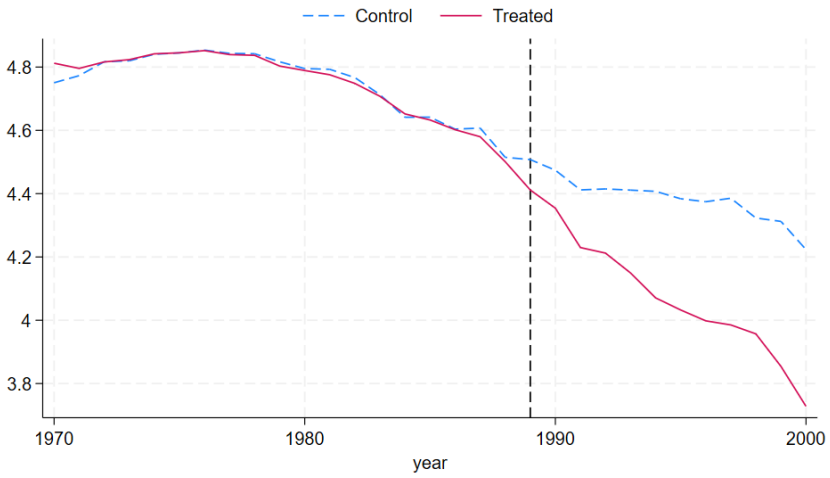


Figure 3.4 Synthetic Control Method, All Controls

Unlike Synthetic Control, Synthetic DiD in Figure 3.5 allows for level differences (drift) in pre-period trend between the treated and control units.

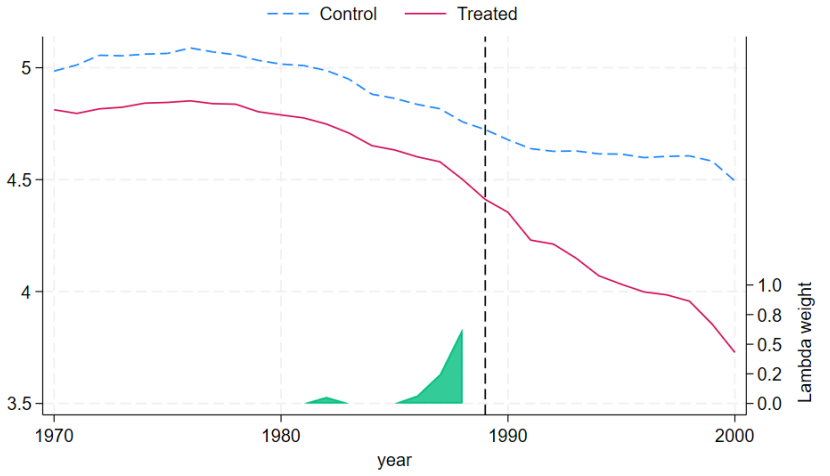


Figure 3.5 Synthetic DID, All Controls

Table 3.3 reports the estimated effect and associated standard errors from these four different estimation method: (i) Demeaning – Procedure 2.1, (ii) Detrending – Procedure 3.1, (iii) Synthetic Control Method, and (iv) Synthetic DiD. The first row, *Average Effect*, presents a single estimated effect averaged over all treated periods.

Table 3.3 Estimated ATTs, 38 Control (Donor) State

	Procedure 2.1 (DiD)	Procedure 3.1 (Unit-Specific Detrending)	SC	Synthetic DiD
Average Effect	-0.422***	-0.227**	-0.304***	-0.286***
	(0.121)	(0.094)	(0.112)	(0.097)
τ_{1989}	-0.168*	-0.043		
	(0.096)	(0.059)		
τ_{1995}	-0.484***	-0.282**		
	(0.137)	(0.112)		
τ_{2000}	-0.667***	-0.403**		
	(0.164)	(0.152)		

Note 1: standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Note 2: For Procedure 3.1, the p-value for the average effect under the normality assumption is 0.021. Based on randomization inference (RI) with 1,000 replications, the p-value is 0.041. This result is reasonably close to the conventional value, supporting the robustness of the inference.

Table 3.3 also shows the estimated effect and associated standard errors for several time horizons, since our cross-sectional regression method can easily estimate separate effects for each t. As is evident from the picture, the estimated effect grows over time, starting off small but ending at $year = 2000$ with -0.403 ($t = -2.65$).

3.6.2 Robustness Check

To further evaluate the robustness of our state-specific detrending procedure, we contrast the results from Procedure 3.1 with those from synthetic control methods when the donor pool is restricted to a small set of states that, a priori, does not seem "similar" California. While the previous section used all 38 potential control states, here we consider two alternative donor pools. The first group consists of Alabama (AL), Arkansas (AR), Louisiana (LA), and Mississippi (MS)—four southern states that differ substantially from California in socioeconomic characteristics. The second group includes Midwestern states—Illinois (IL), Indiana (IN), Iowa (IA), and Ohio (OH). By limiting the set of controls in this way, we assess how each estimator performs when the availability of similar donor units is limited, providing additional insights into the robustness of our procedure.

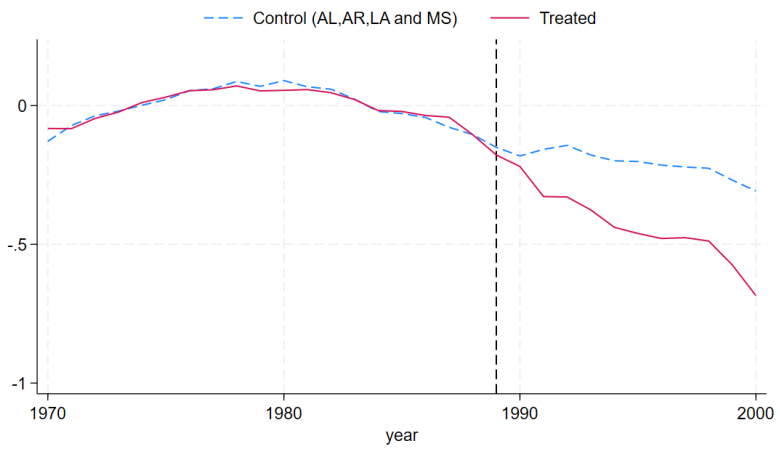
3.6.2.1 AL, AR, LA and MS as Controls

Table 3.4 summarizes the estimates using Alabama (AL), Arkansas (AR), Louisiana (LA), and Mississippi (MS) as control states across four different estimation methods. The results from the detrending procedure closely resemble those obtained when all 38 states are used as controls, both in magnitude and statistical significance. In contrast, the synthetic control (SC) and synthetic DiD methods yield larger estimates in absolute value, reflecting poorer pre-treatment fits observed in Figure 3.6.

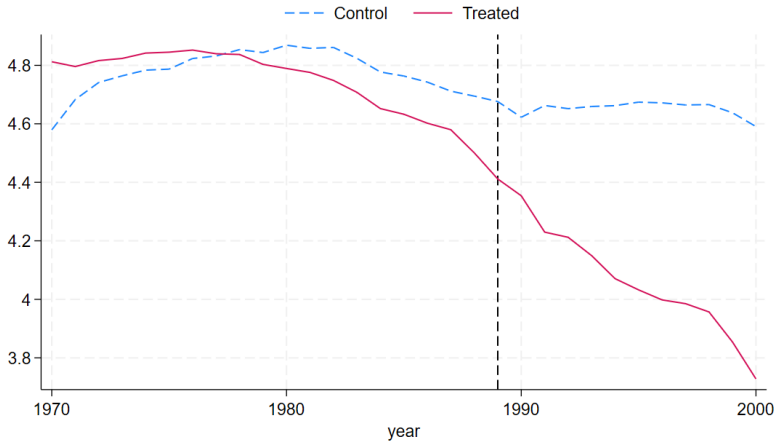
Table 3.4 Estimated ATTs, AL, AR, LA, MS as Controls

	Procedure 2.1 (DiD)	Procedure 3.1 (Unit-Specific Detrending)	SC	Synthetic DiD
Average	-0.556***	-0.215**	-0.571***	-0.392***
	(0.080)	(0.039)	(0.034)	(0.030)
1989	-0.247	-0.027		
	(0.107)	(0.052)		
1995	-0.611***	-0.259**		
	(0.077)	(0.055)		
2000	-0.839***	-0.377**		
	(0.032)	(0.115)		
Note: standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1				

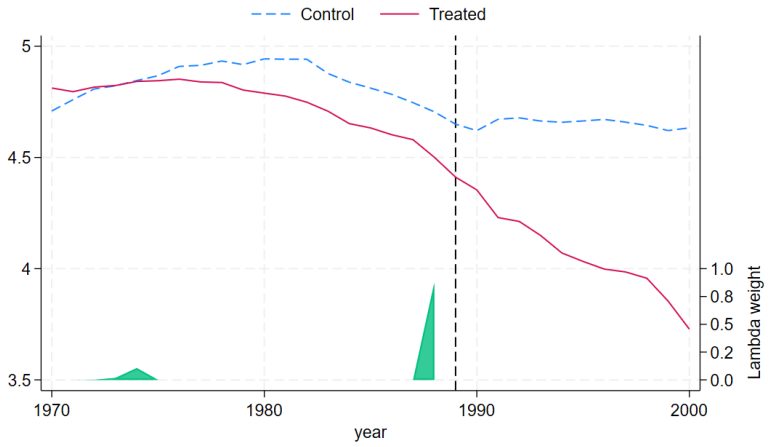
Panel (a) of Figure 3.6 shows how the removing unit-specific linear trends fits when the controls consist of AL, AR, LA, and MS. The average residuals across these four states provide a remarkably good fit to the detrended California. By contrast, Panel (b) and (c) of Figure 3.6 shows that both SC method and Synthetic DiD cannot recover a synthetic California that provides a close enough fit in order to produce a convincing estimate of the causal effect.



(a) State-Specific Detrending



(b) Synthetic Control



(c) Synthetic DiD

Figure 3.6 AL, AR, LA and MS as Controls

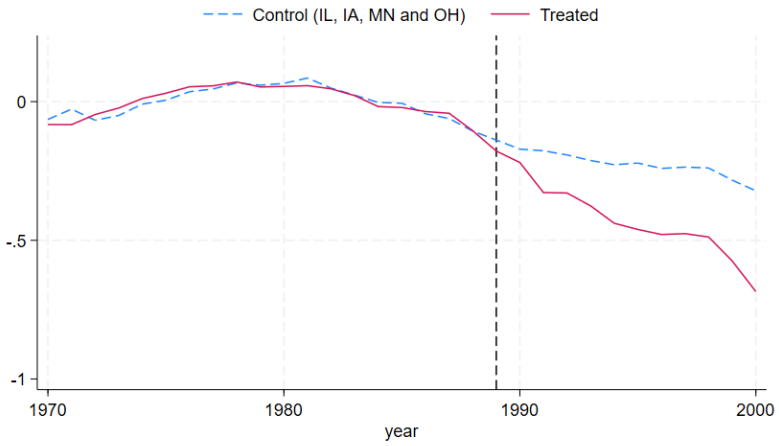
3.6.2.2 IL, IA, MN and OH as Controls

Similarly, we next employ four Midwestern states—Illinois (IL), Iowa (IA), Minnesota (MN), and Ohio (OH)—as control units. The corresponding results are presented in Table 3.5. As in the previous case with AL, AR, LA, and MS as controls, our unit-specific detrending approach achieves a strong pre-treatment fit, while the synthetic control and synthetic DiD methods perform poorly. As shown in Panel (a) of Figure 3.7, the detrending method closely aligns with the pre-treatment trends, whereas the alternative methods exhibit clear discrepancies.

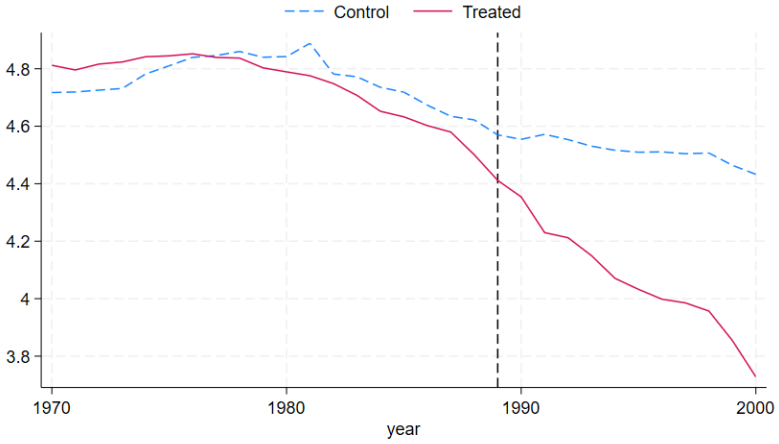
Table 3.5 Estimated ATTs, IL, IA, MN and OH as Controls

	Procedure 2.1 (DiD)	Procedure 3.1 (Unit-Specific Detrending)	SC	Synthetic DiD
Average	-0.413**	-0.198*	-0.437**	-0.275*
	(0.118)	(0.079)	(0.184)	(0.154)
1989	-0.178*	-0.040		
	(0.071)	(0.045)		
1995	-0.462**	-0.239*		
	(0.133)	(0.088)		
2000	-0.655**	-0.363*		
	(0.183)	(0.136)		
Note: standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1				

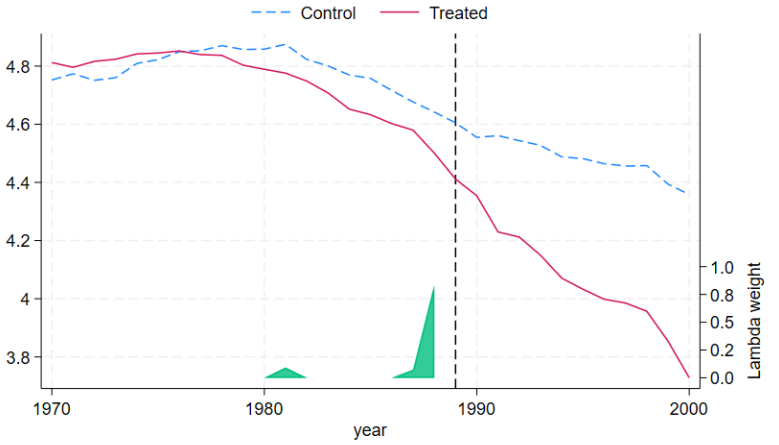
These findings further support the robustness of our approach, even in settings with a limited donor pool. Notably, while both synthetic DiD and synthetic control method typically rely on a large number of control units to construct a reliable counterfactual, our method requires only minimal conditions (e.g., at least one treated and one control unit, with a total of three or more units), making it particularly useful when the donor pool is small.



(a) State-Specific Detrending



(b) Synthetic Control



(c) Synthetic DiD

Figure 3.7 IL, IA, MN and OH as Controls

3.7 Staggered Rollouts

We can extend the previous methods to the case of staggered interventions, where units are allowed to be first treated at different times. But one must use care in choosing a suitable control group.

3.7.1 The Setup and Estimation

Suppose now that the first treatment period is S, but treatment can first occur in periods $S+1, \ldots, T$. Index the treatment cohorts or groups by $g \in \{S, S+1, \ldots, T\}$, the time of first treatment. Let D_g , $g = S, S+1, \ldots, T$ be the cohort indicators. Specifically, $D_{ig} = 1$ if and only if unit i was first treated in period g. The never treated units are indicated by $D_{i\infty} = 1$.

There are N_g units in each cohort, with N_∞ being the number of never treated units. With $N_\infty \geq 2$, the other treated cohorts may have only one unit. Naturally, if there are periods without new treated units, there will be no treatment effects estimated in those periods.

The potential outcomes are $Y_t(g)$, $g = S, \ldots, T, \infty$, where $Y_t(g)$ is the outcome in period t if the first period of treatment is g. $Y_t(\infty)$ is the never treated state (or control state). The treatment effects are now indexed by cohort and time:

$$\tau_{gt} = E\left[Y_t(g) - Y_t(\infty) \,|D_g = 1\right], \; t = g, \ldots, T. \tag{3.29}$$

We are also interested in estimating the cohort-specific average effects,

$$\tau_g = \frac{1}{(T-g+1)} \sum_{t=g}^{T} \tau_{gt}. \tag{3.30}$$

As is common, we make a no anticipation Assumption:

$$Y_t(g) = Y_t(\infty), \, t < g,$$

which says that all potential outcomes are the same as in the never-treated state in periods before the treatment occurs.

For a given cohort g, one now removes the average or a trend using data up through $g-1$, the final pre-treatment period. It is easiest to obtain the transformation for each unit in the sample, and

then sort out the valid control units afterwards. For each i, one runs the regressions

$$Y_{it} \text{ on } 1, \;\; t = 1, 2, \ldots, g-1 \tag{3.31}$$

Because we have different treatment cohorts, we add a cohort subscript when projecting out to post-treatment periods:

$$\dot{Y}_{itg} = Y_{it} - \bar{Y}_{i,pre(g)}, \;\; t = g, \ldots, T \tag{3.32}$$

where $\bar{Y}_{i,pre(g)}$ is the average of Y_{it} for $t = 1, \ldots, g-1$. To remove a linear time trend, the regressions are

$$Y_{it} \text{ on } 1, t, \;\; t = 1, 2, \ldots, g-1 \tag{3.33}$$

and then the out-of-sample residuals are

$$\ddot{Y}_{itg} = Y_{it} - \widehat{A}_{ig} - \widehat{B}_{ig} \cdot t, \;\; t = g, \ldots, T \tag{3.34}$$

Once the transformed variables have been obtained, the key is choosing a valid control group at time t for treatment cohort g (which could have as few as one unit). One can use only the never treated units at time t, provided $N_\infty \geq 2$. However, as shown in LW (2023), under suitable NA and PT assumptions, one can using any unit not-yet treated (NYT). Then for cohort g in time period t, the units represented by

$$D_{i,t+1} + D_{i,t+2} + \cdots + D_{iT} + D_{i\infty} = 1 \tag{3.35}$$

are valid controls. For efficiency reasons, one should use all possible controls (and this tends to help the normality assumption be more realistic). Nevertheless, one can use any subset (including the NT group). Let $C_{i,t+1}$ be the control group chosen for period t, with the subscript indicating that these groups cannot have been treated prior to $t+1$. Then $\widehat{\tau}_{gt}$ is the coefficient on D_{ig} in the cross-sectional regression

$$\dot{Y}_{itg} \text{ on } 1, D_{ig} \;\; \text{using } D_{ig} + C_{i,t+1} = 1 \tag{3.36}$$

Or, replace $\dot{Y}_{itg}$ with the detrended version, $\ddot{Y}_{itg}$. Under a homoskedastic normality assumption, exact inference can be used in (3.36) even if $N_g = 1$ and the number of control units is small.

For many purposes, we are more interested in the τ_g, which we can aggregate to obtain a single, weighted effect – more on this below. Then, because we want inference without many units in the treated cohort, we use the never treated units as the control and run a single, aggregated regression. Specifically, define

$$\overline{\dot{Y}}_{ig} \equiv \frac{1}{(T-g+1)} \sum_{t=g}^{T} \dot{Y}_{it} \text{ and } \overline{\ddot{Y}}_{ig} \equiv \frac{1}{(T-g+1)} \sum_{t=g}^{T} \ddot{Y}_{it}. \tag{3.37}$$

Then run the regression

$$\overline{\dot{Y}}_{ig} \text{ on } 1, D_{ig}, i = 1, ..., N, D_{ig} + D_{i\infty} = 1 \tag{3.38}$$

to obtain $\widehat{\tau}_g$ as the coefficient on D_{ig}, or use $\overline{\ddot{Y}}_{ig}$ in place of $\overline{\dot{Y}}_{ig}$. Again, this is a cross-sectional regression with independent observations. If N_g and N_∞ are suitably large to employ asymptotics, then we can use a heteroskedasticity-robust standard error (hc3) to obtain a t statistic or confidence intervals. We can appeal to exact theory if N_g is small, including $N_g = 1$.

If we want to aggregate the $\widehat{\tau}_g$ to obtain a single treatment effect, a challenge is to obtain a valid standard error when cohort sizes are not particularly large. A natural parameter is to weight the cohort ATTs, τ_g, but the cohort shares. This leads to the estimator

$$\widehat{\tau}_\omega = \sum_{g=S}^{T} \widehat{\omega}_g \widehat{\tau}_g \tag{3.39}$$

where

$$\widehat{\omega}_g = \frac{N_g}{N_S + N_{S+1} + \cdots + N_T} \tag{3.40}$$

are the cohort shares. A standard error for $\widehat{\tau}_\omega$ must account for the covariances among the $\widehat{\tau}_g$, which is difficult if we cannot appeal to asymptotics. Even with a somewhat larger number of total treated units, a trick is helpful. We know from simple regression analysis that $\widehat{\tau}_g$ is a difference in means:

$$\widehat{\tau}_g = \frac{1}{N_g} \sum_{i=1}^{N} D_{ig} \overline{\dot{Y}}_{ig} - \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i\infty} \overline{\dot{Y}}_{ig} \tag{3.41}$$

and so, by simple algebra,

$$\widehat{\tau}_\omega = \frac{1}{(N_S + N_{S+1} + \cdots + N_T)} \sum_{g=S}^{T} \sum_{i=1}^{N} D_{ig} \overline{\dot{Y}}_{ig} - \frac{1}{N_\infty} \sum_{g=S}^{T} \sum_{i=1}^{N} D_{i\infty} \widehat{\omega}_g \overline{\dot{Y}}_{ig} \tag{3.42}$$

Define a treatment indicator

$$D_i = D_{iS} + \cdots + D_{iT} \tag{3.43}$$

the number of treated units as

$$N_{treat} = N_S + N_{S+1} + \cdots + N_T.$$

Rearrange (3.42) to obtain

$$\begin{aligned}
\widehat{\tau}_\omega &= \frac{1}{N_{treat}} \sum_{i=1}^{N} D_i \cdot \left(\sum_{g=S}^{T} D_{ig} \overline{\dot{Y}}_{ig} \right) - \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i\infty} \cdot \left(\sum_{g=S}^{T} \widehat{\omega}_g \overline{\dot{Y}}_{ig} \right) \\
&\equiv \frac{1}{N_{treat}} \sum_{i=1}^{N} D_i \cdot \overline{\dot{Y}}_i - \frac{1}{N_\infty} \sum_{i=1}^{N} D_{i\infty} \cdot \overline{\dot{Y}}_i \qquad (3.44) \\
&= \frac{1}{N_{treat}} \sum_{i=1}^{N} D_i \cdot \overline{\dot{Y}}_i - \frac{1}{N_{control}} \sum_{i=1}^{N} (1 - D_i) \cdot \overline{\dot{Y}}_i \qquad (3.45)
\end{aligned}$$

where

$$\overline{\dot{Y}}_i \equiv D_{iS} \cdot \overline{\dot{Y}}_{iS} + \cdots + D_{iT} \cdot \overline{\dot{Y}}_{iT} + D_{i\infty} \cdot \left(\sum_{g=S}^{T} \widehat{\omega}_g \overline{\dot{Y}}_{ig} \right). \tag{3.46}$$

Note that the representation in (3.45) uses the fact that $D_i \cdot D_{ig} = D_{ig}$. Also, $D_i + D_{i\infty} = 1$, and so (3.45) is a simple difference in sample mean of $\overline{\dot{Y}}_i$ between units that are eventually treated and the never treated group. In other words, run the cross-sectional regression

$$\overline{\dot{Y}}_i \text{ on } 1, D_i, \ i = 1, \ldots, N, \tag{3.47}$$

and then $\widehat{\tau}_\omega$ is the coefficient on D_i. Naturally, if we use unit-specific detrending for each cohort date g to obtain the $\overline{\ddot{Y}}_{ig}$, then we replace $\overline{\dot{Y}}_{ig}$ with $\overline{\ddot{Y}}_{ig}$ everywhere.

Obtaining $\widehat{\tau}_\omega$ from the regression in (3.47) has some advantages. It automatically accounts for the correlations among the $\widehat{\tau}_g$, and it can be used whether or not N_{treat} is small or large (and same for $N_{control}$). If N_{treat} and $N_{control}$ are even moderately large, a heteroskedasticity-robust standard error from (3.47) is justified via asymptotic analysis.

Arkhangelsky et al. (2021) discuss how SDiD can be applied to staggered interventions. They suggest splitting the sample by adoption date and applying the SDiD method separately to each

treatment cohort, using the never treated cohort as the control [see the appendix in Arkhangelsky et al. (2021)]. In other words, do exactly what we propose with our unit-specific demeaning or detrending. Therefore, our proposed methods and the way SDiD is implemented for staggered designs are directly comparable.

3.7.2 Application: Estimating the Effects of Castle Laws on Homicides

We apply the previous methods to the data set used in Cunningham (2021) on the adoption of so called "castle" laws – or "hold-your-ground" laws – on homicide rates in the United States. A castle law typically allows individuals to use force, including deadly force, to defend themselves against an intruder in their home – without a duty to retreat. The data set covers the 50 United States from 2000 through 2010. In 2005, one state adopted a castle law. In 2006, 13 more states did. The remaining treated cohorts are 2007 (four states), 2008 (two states), and 2009 (one state). Therefore, there are $N_{treat} = 21$ eventually treated units and $N_{control} = 29$ control units. the outcome variable is the log of the annual homicides.

Using the regression in (3.47) with $\overline{\dot{Y}}_i$ defined in (3.46), the estimated aggregated treatment effect, $\widehat{\tau}_\omega$, is about 0.092, or about 9.2% more homicides from a state adopting a castle law. The usual OLS standard error is 0.057, which gives $t \approx 1.61$. This is not quite significant at the 10% level against a two-sided alternative. The hc3 t statistic is 1.50. Replacing $\overline{\dot{Y}}_i$ with $\overline{\ddot{Y}}_i$ – obtained from linear detrending – decreases the estimate to 0.067, but the hc3 standard error is 0.055, and $t \approx 1.21$.

The synthetic DiD estimate, obtained using the sdid package in Stata 18, is 0.099 and the standard error, using the placebo method, is 0.069 ($t = 1.41$). Both the estimate and its precision are in close agreement with the demeaning method described in Section 3.7.1.

3.8 Additional Practical Considerations

3.8.1 Choosing the Number of Time Periods

With synthetic control-type approaches and the approaches we suggest here, one can study the robustness of the findings by adjusting the number of pre-treatment time periods. For example, when using the unit-specific detrending method, we can vary the starting point of the data as a way

of evaluated the way the estimates change as the unit-specific trends are removed using different stretches of data. This is in the spirit of Rambachan and Roth (2023), who formalize the idea of allowing a range of violations of parallel trends and studying the sensitivity to the estimates. In many examples, the policy intervention is likely based on past outcomes (in the untreated state). Does one need to go back, say, 20 years to remove unit specific intercepts and trends to account for the selection into treatment? In many cases, fewer pre-treatment periods might suffice. Because our approach does not rely on large T_0 or T_1 (but does rely on normality), the robustness of the estimates can be studied by varying T_0 in particular.

3.8.2 Clustering and Spatial Correlation with Larger Cross Sections

Although our motivation for the previous procedures is motivated by settings with few control or treated units, or few of both, our approach has benefits in settings where $N_{control}$ and N_{treat} are large enough to rely on large-N asymptotic analysis. Recall that the basic DiD estimator is obtained from

$$\dot{\overline{Y}}_i \text{ on } 1, D_i, i = 1, ..., N$$

and the one that removes pre-treatment unit-specific trends is

$$\ddot{\overline{Y}}_i \text{ on } 1, D_i, i = 1, ..., N.$$

We can even add controls with large enough N. In Section 3.7 we showed that the same regressions can be used in the case of staggered assignment by defining D_i to be an "ever treated" indicator – see (3.43) – and by modifying $\dot{\overline{Y}}_i$ or $\ddot{\overline{Y}}_i$ as in equation (3.46). Because these are cross-sectional regressions, it is straightforward to compute standard errors clustered at a level higher than i. For example, if i is a county, but we we are studying a policy that varies only at the state level, we can cluster at the state level if we have a sufficiently large number of treated and control states. Again, it does not matter how large T_0 and T_1 are. Abadie, Athey, Imbens and Wooldridge (2023) discuss why standard errors should be clustered when the intervention is at a higher level than the unit. As another example, i could be a household whereas a policy is applied at the village level. Clustering can be done separately by time period, too. The key is that it is a cross sectional regression and

then the usual clustering methods can apply.

In situations with a spatial structure – for example, the assignment of treated units in implementing a new policy may be spatially correlated – we can obtain standard errors using a covariance matrix estimator robust to heteroskedasticity and spatial correlation. These so-called "SHAC" standard errors are proposed in Conley (1999).

3.9 Concluding Remarks

Building on Lee and Wooldridge (2023), we have proposed a simple approach to inference when using panel data with few treated units or few control units. In the common timing case and without controls, the unit-specific demeaning reproduces the standard difference-in-differences estimator for each treated period, and also averaged across the treated periods. We also show how the unit-specific trending method of LW (2023) can be implemented.

Estimation is simple and inference follows under normality using the classical linear model assumptions taught in introductory econometrics. Because the method uses averages across time, the central limit theorem across the time dimension often can be used to justify the normality assumption. Heteroskedasticity is always an issue in cross-sectional regressions, and we propose using what is known as the hc3 version provided there are at least a handful of treated units.

The approach here is not intended to replace the very popular synthetic DiD method of AAHIW (2021) in cases where SDiD has natural advantages. However, our simulations show that, when there is sufficient heterogeneity in, say, linear time trends, our method of unit-specific detrending can have substantially less bias. In some cases the inference is more reliable. Moreover, SDiD is not intended to apply to situations with few time periods, or few donor units among which to choose controls. The SDiD asymptotics assumes N_0, T_0, and T_1 all increase – although our simulations suggest SDiD tends to work well more generally, provided trends are not to heterogeneous.

Our approach also allows the estimation of separate effects in the treated periods, a feature not allowed by SDiD methods. Also, we showed in Section 3.7 that allowing for staggered interventions is straightforward, and obtaining an overall weighted estimate with valid standard error can be done by a single cross-sectional regression.

Overall, we view our approach as complementary to SDiD methods for many applications. When applied to the California smoking data, the state-specific detrending, using all 38 control states, produces estimates and inference similar to SDiD when restricting attention to the overall average affect. In applying our approach to the staggered rollout of so-called castle laws, the our rolling method based on unit-specific demeaning gives a very similar point estimate and standard error to SDiD. For cases with a small number of time periods, a small number of controls, or both, exact inference in a cross-sectional regression using our transformation is appealing.

BIBLIOGRAPHY

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2023), "When should you adjust standard errors for clustering? ", *The Quarterly Journal of Economics* 138(1), 1–35.

Abadie, A., Diamond, A. and Hainmueller, J. (2010), "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program", *Journal of the American statistical Association* 105(490), 493–505.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W. and Wager, S. (2021), "Synthetic difference-in-differences", *American Economic Review* 111(12), 4088–4118.

Callaway, B. and Sant'Anna, P. H. (2021), "Difference-in-Differences with Multiple Time Periods", *Journal of econometrics* 225(2), 200–230.

Conley, T. G. (1999), "GMM estimation with cross sectional dependence", *Journal of econometrics* 92(1), 1–45.

Cunningham, S. (2021), *Causal inference: The mixtape*, Yale university press.

Davidson, R., MacKinnon, J. G. et al. (1993), *Estimation and inference in econometrics*, Vol. 63, Oxford New York.

Donald, S. G. and Lang, K. (2007), "Inference with difference-in-differences and other panel data", *The review of Economics and Statistics* 89(2), 221–233.

Hagemann, A. (2025), "Inference with a single treated cluster", *Review of Economic Studies* p. rdaf002.

Heß, S. (2017), "Randomization inference with Stata: A guide and software", *The Stata Journal* 17(3), 630–651.

Lee, S. J. and Wooldridge, J. M. (2023), "A Simple Transformation Approach to Difference-in-Differences Estimation for Panel Data", *Available at SSRN 4516518* .

Rambachan, A. and Roth, J. (2023), "A more credible approach to parallel trends", *Review of Economic Studies* 90(5), 2555–2591.

Simonsohn, U. (2021), "Just Run Robust Standard Errors: A Commentary on Young (2019)". Working paper, available at `http://urisohn.com/43`.

Wooldridge, J. M. (2010), "Econometric Analysis of Cross Section and Panel Data".

Wooldridge, J. M. (2020), *Introductory Econometrics: A Modern Approach*, Cengage learning:: Boston, 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. (2025), "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators", *latest version of Wooldridge(2021)* .

Young, A. (2019), "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results", *The quarterly journal of economics* 134(2), 557–598.

CHAPTER 4

ROLLING APPROACH TO DIFFERENCE-IN-DIFFERENCES: EXPLORING TREATMENT REVERSIBILITY AND MODERATING EFFECTS

4.1 Introduction

Difference-in-differences (DiD) methods have become indispensable for estimating causal effects in settings with staggered interventions and heterogeneous responses. Yet, recent work has highlighted a critical limitation: when treatment effects vary across groups or over time, the traditional two-way fixed effects (TWFE) estimator can produce biased estimates under staggered intervention designs. To overcome this issue, several alternative estimators have been proposed to capture treatment effect heterogeneity more faithfully; see, for example, Callaway and Sant'Anna (2021), Wooldridge (2021), Sun and Abraham (2021), Borusyak et al. (2024), and De Chaisemartin and d'Haultfoeuille (2020).

Most of these approaches, however, rely on the assumption of *absorbing treatments*—once a unit receives treatment, it remains treated in all subsequent periods. But, in many real-world settings, treatment status is neither permanent nor monotonic. For example, firms may adopt a policy temporarily, governments may roll out and later withdraw interventions, or—as in the setting examined in this paper—pharmacy chains may open and later close locations. Most existing methods are not well suited to handle such dynamic treatment regimes, and few provide formal guidance for estimation and aggregation under this complexity.

Recently, De Chaisemartin and d'Haultfoeuille (2024) propose a framework designed to identify treatment effects in contexts with *non-binary*, *non-absorbing*, and potentially *lagged* treatments, relating to their prior work in De Chaisemartin and d'Haultfoeuille (2020). Their method accommodates reversible treatment states, allowing units to both enter and exit treatment over time, and explicitly accounts for dynamic effects that unfold across different exposure durations.

However, their approach defines treatment relative to exposure duration and classifies units based on their initial transition into treatment. Once a unit exits treatment, its subsequent observations are excluded from the analysis. As a result, the framework does not permit the estimation of

post-exit effects, and re-entry into treatment is not considered. This limitation hinders the ability to assess whether treatment effects persist after discontinuation or to evaluate the impact of repeated exposures.

In contrast, this paper introduces a more flexible framework—building on Lee and Wooldridge (2023)—that enables the estimation of treatment effects both during and after treatment episodes. By aligning event time relative to treatment exposure (e.g., $t - g$, where g is the treatment start period), I incorporate both on-treatment and post-exit periods in the analysis, facilitating the identification of lingering or decaying effects over time. This extension is particularly useful for evaluating whether the impact of an intervention persists or fades after discontinuation.

Furthermore, I explicitly account for re-entry into treatment—a case ignored in prior implementations—by identifying multiple treatment paths and estimating the effects of subsequent exposures separately from the initial one. This allows for direct comparisons between first-time and repeated treatments and offers a more comprehensive perspective on dynamic, non-monotonic treatment patterns. Together, these contributions enhance the applicability of dynamic DID methods in real-world settings where treatment status can change multiple times, and allowing for evaluating whether being treated again has the same, weaker, or stronger impact than the first treatment.

In addition, I propose an extended doubly robust estimator that incorporates moderating effects across subgroups using an augmented inverse probability weighting and regression adjustment (IPWRA) approach. This method allows for explicit estimation of treatment effect heterogeneity across subgroups while preserving the desirable double robustness property. It also enables researchers to examine whether treatment effects vary across subgroups, allowing for more nuanced evaluations of treatment or policy interventions in the presence of demographic or structural heterogeneity.

The remainder of the paper is organized as follows. Section 4.2 outlines the setup and identifying assumptions underlying the proposed framework. In Section 4.3, I demonstrate how to incorporate moderating effects while leveraging a doubly robust estimator. Section 4.4 presents Monte Carlo simulation results to evaluate the performance of the proposed method. In Section 4.5, I apply the approach to data on the entry and exit of chain pharmacy stores to examine their impact on

independently owned pharmacies in rural areas. Section 4.6 concludes.

4.2 Setup and Identification

This framework is designed to accommodate treatment patterns involving both staggered adoption and reversible treatment spells, aligning with and extending the work of Lee and Wooldridge (2023). By incorporating exit and re-entry episodes, we provide a comprehensive toolkit for analyzing dynamic treatment effects where treatment status may change multiple times throughout the study period.

Our approach is initially focused on binary treatment indicators, classifying units as either treated or untreated at each time point. I define the treatment indicator $D_{ig} \in \{0, 1\}$, where $D_{ig} = 1$ if unit i is treated at time g, and $D_{ig} = 0$ otherwise. For notational simplicity, I denote the treatment indicator as $D_g \equiv D_{ig}$, suppressing the unit index when focusing solely on treatment timing. However, future work will extend this framework to discrete treatment intensities, where units can receive varying levels of treatment $D_{g,t} \in \{0, 1, \ldots, K\}$, allowing for analysis of policies where treatment intensity (e.g., dosage or funding levels) is a crucial factor.

We consider the following treatment patterns:

- **Case 1: Binary Treatment with Staggered Adoption** — This case mirrors the structure in Lee and Wooldridge (2023), where treatment is introduced at different times across units without accounting for exit or re-entry.

- **Case 2: Binary Treatment with Exit and Re-entry** — Units may exit treatment and potentially re-enter at a later time, resulting in multiple treatment spells. This setting allows for robust comparisons by aligning units based on their treatment histories and adjusting for pre-treatment dynamics.

By accommodating a wider array of treatment patterns—including reversible and intermittent treatments—our framework provides a unified, flexible approach for researchers aiming to assess the dynamic and heterogeneous effects of policy interventions in realistic settings.

4.2.1 Identification Assumptions

To identify treatment effects under these settings, we introduce the following assumptions, which follows Lee and Wooldridge (2023). Following the potential outcome framework, let $Y_t(g)$ denote the potential outcome at time t for a unit treated at time g. For units that are never treated, we define the untreated state using ∞, such that $Y_t(\infty)$ represents the potential outcome under the control state at time t.

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

$$E\left[Y_t\left(g\right)|D_g = 1, \mathbf{X}\right] = E\left[Y_t\left(\infty\right)|D_g = 1, \mathbf{X}\right]. \square \tag{4.1}$$

This implies treatment effects prior to the intervention are all zero.

Assumption 2. Conditional Parallel Trends for Treatment, Exit and Re-entry Groups (CPT for T.E.R): For $\mathbf{D} = (D_S, \ldots, D_T)$ and $t = 1, 2, \ldots, T$,

$$E[Y_t(\infty) - Y_1(\infty)|\mathbf{D}, \mathbf{X}] = E[Y_t(\infty) - Y_1(\infty)|\mathbf{X}], \quad t = 2, \ldots, T. \square \tag{4.2}$$

Units with identical outcome paths up to $t = g-1$ are assumed to follow the same expected trends in the evolution of control state outcomes, regardless of whether they exit or re-enter treatment after their initial treatment at time g. This allows for constructing a control group based on pre-treatment outcome paths, facilitating identification of treatment, exit, and re-entry effects.

Assumption 3. Overlap: For cohorts $g \in \{S, \ldots, T\}$ and time periods $r \in \{g, g+1, \ldots, T\}$,

$$P\left(D_g = 1|D_g + A_{r+1} = 1, \mathbf{X} = \mathbf{x}\right) < 1 \text{ for all } \mathbf{x} \in \text{Supp}\left(\mathbf{X}\right). \square \tag{4.3}$$

This condition ensures that within the subpopulation consisting of cohort g, as well as the never-treated and not-yet-treated units at time r (denoted as A_{r+1}), every treated unit has a comparable control unit with the same level (magnitude) of covariates.

4.2.2 Estimation Strategy: Treatment, Exit, and Re-entry Effects

The estimation procedure is structured in two main stages:

First Stage: General Effect Estimation

For each treatment group g, i.e., whose initial treatment occur at time g, we define a control group at time t based on units with the same outcome paths up until time $g-1$ and not yet treated at time t, which including never treated units. At this stage, I focus solely on each group's first treatment timing g, without accounting for subsequent treatment, exit, or re-entry status. It means we will estimate the treatment effects on the treated ATT(g,t) as if each group is treated at g, and then stay treated up to T, which means their realized potential outcome is considered as $Y_{it}(g)$. However, they could "exit" or "re-enter" down the road. Thus, let me define these estimated group-time specific effect as a general treatment effect on the treated (GTT) at time t for g as follow:

$$GTT(g,t) = E\left[Y_{it}(g) - Y_{it}(\infty)|D_g = 1\right] \tag{4.4}$$

To get this $GTT(g,t)$, I simply apply **Procedure 4.1 (Rolling Methods, Staggered Interventions)** in Lee and Wooldridge (2023):

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

$$\dot{Y}_{irg} \equiv Y_{ir} - \frac{1}{(g-1)}\sum_{s=1}^{g-1} Y_{is} \equiv Y_{ir} - \bar{Y}_{i,pre(g)} \tag{4.5}$$

Step 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 the Never Treated (NT) group).

Step 3. Using the subset of data with

$$D_{ig} + D_{i,r+1} + D_{i,r+2} + \cdots + D_{iT} + D_{i\infty} = 1, \tag{4.6}$$

apply standard treatment effect (TE) methods—such as linear regression adjustment (RA), inverse probability weighting (IPW), IPWRA, and matching—to each cross-sectional dataset separately.

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

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

Lee and Wooldridge (2023) establish that the coefficient on D_{ig} from these cross-sectional regressions identifies the average treatment effect on the treated (ATT) for group g at time t, denoted as $ATT(g,t)$, under the assumption of treatment irreversibility. In addition, if you are concerned about unit-specific heterogeneous linear trends, you can apply Procedure 5.1 in Lee and Wooldridge (2023), which allowing for unit-specific heterogeneous linear trends.

However, in the present framework, I allow for the possibility that group g may exit treatment at certain periods t ($> g$). Thus, while I initially follow the Lee and Wooldridge (2023) approach to estimate the effect for these periods, I subsequently reclassify these periods as exit periods in Step 2 of our estimation procedure.

Consequently, instead of defining the effect as $ATT(g,t)$, I adopt a broader concept, denoted as $GTT(g,t)$, representing the generalized treatment effect for group g at time t, which can encompass not only the standard treatment effect but also exit and re-entry effects.

Second Stage: Classification and Aggregation by Treatment Path

After estimating $GTT(g,t)$ for all $g \in \{S,\ldots,T\}$ and $t \in \{g,\ldots,T\}$, we classify each period based on treatment path, allowing for a more nuanced interpretation of treatment effects. The classification includes three distinct categories based on the observed treatment status:

$$\alpha \in \{Treated,\ Exit,\ Re-entry\},$$

In this context, *Treated* refers to time periods immediately following the initial treatment, indicated by a positive treatment status relative to the first treatment timing. *Exit* represents periods after treatment cessation, during which the treatment status returns to zero following the initial treatment. *Re-entry* denotes periods after the re-initiation of treatment following an exit episode, allowing for multiple treatment spells.

Table 4.1 illustrates the classification for two groups, $g = 4$ and $g = 3$, each representing distinct treatment paths. The table presents the relative time for each treatment spell, distinguishing periods of treatment, exit, and re-entry based on the specific treatment path of each group.

Table 4.1 Classification of Treatment Paths for Groups $g = 4$ and $g = 3$

Group $g = 4$	Treatment Path	0	0	0	1	1	1	0	0	0
	Calendar Time (t)	1	2	3	4	5	6	7	8	9
1. Treatment Period	Relative Time (r)	-3	-2	-1	0	1	2			
2. Exit Period	Relative Time (r)							0	1	2
3. Re-entry Period	Relative Time (r)									

Group $g = 3$	Treatment Path	0	0	1	1	0	0	0	1	1
	Calendar Time (t)	1	2	3	4	5	6	7	8	9
1. Treatment Period	Relative Time (r)	-2	-1	0	1					
2. Exit Period	Relative Time (r)					0	1	2		
3. Re-entry Period	Relative Time (r)					-3	-2	-1	0	1

In Group $g = 4$, treatment is initiated at $t = 4$, continues for three periods, and then exits at $t = 7$. Exit effects are observed from $t = 7$ to $t = 9$. No re-entry occurs in this group. In contrast, Group $g = 3$ initiates treatment earlier at $t = 3$, exits at $t = 5$, and then re-enters at $t = 8$. The re-entry period spans $t = 8$ to $t = 9$, allowing for the estimation of re-treatment effects distinct from the initial treatment effects.

4.2.3 Aggregation Strategy

Define relative time $r = t - g$. Then, we compute the weighted average of the group-time general treatment effects $GTT(g, t)$ based on relative time r for each $\alpha \in \{\text{Treated, Exit, Re-entry}\}$:

$$WGTT(\alpha, r) = \sum_{g \in G_{r,\alpha}} w(g, r) \cdot GTT(g, g + r), \ where \ \ r = t - g \tag{4.7}$$

where $G_{r,\alpha}$ denotes the set of groups with relative time r and treatment state α at time t, and $w(g, r)$ represents the weight assigned to group g. While various weighting strategies could be considered, this paper adopts a simple proportional weighting scheme based on group sizes:

$$w(g, r) = \frac{N_g}{N_{G_{r,\alpha}}},$$

where N_g is the number of units in group g, and $N_{G_{r,\alpha}}$ is the total number of units across all groups in $G_{r,\alpha}$ used for estimating $WGTT(\alpha, r)$. For example, if only groups b and c, each with 10 units, contribute to the estimation at $r = 1$, then $w(b, 1) = w(c, 1) = \frac{10}{20} = 0.5$.

This aggregation yields three distinct $WGTT$ series, which can be visualized through event-study plots to analyze dynamic treatment effects across different treatment states.

De Chaisemartin and d'Haultfoeuille (2024) also propose an estimation method for settings with dynamic treatment patterns, such as treatment entry and exit. However, their approach addresses this complexity by focusing exclusively on the "treatment period," and explicitly excluding observations during "exit" and "re-entry" periods. This restriction is driven by their no-crossing condition, which requires treatment levels to remain strictly above or below the period-one status ($t = 1$), effectively dropping post-exit observations from the analysis. As a result, their framework does not accommodate re-entry into treatment or estimate effects beyond the treatment spell.

In contrast, I leverage control groups to estimate effects across all three treatment states. By explicitly classifying periods according to treatment status, the proposed framework enables a comprehensive analysis of how treatment effects evolve across multiple spells and interruptions. This extension provides researchers with a deeper understanding of treatment dynamics over time, including potential lingering effects after discontinuation and resurgent effects following re-entry.

In addition, a key strength of the framework proposed by De Chaisemartin and d'Haultfoeuille (2024) lies in its ability to accommodate discrete treatment intensities and flexibility regarding the initial treatment status. Specifically, even when units are already treated in the initial period, their approach allows for the estimation of treatment effects for "switchers"—units whose treatment intensity strictly increases or decreases relative to the baseline level. By using status-quo units as the control group, their method enables identification of dynamic treatment effects for such transitions under certain assumptions.

While my current study focuses on binary treatment effects, the proposed classification-based framework is naturally extensible to settings with discrete or continuous treatment intensities. Incorporating these variations into the treatment-exit-reentry structure will be a key direction for future research. Such an extension would enable more comprehensive evaluation of dynamic policy interventions, where treatment may vary not only in timing but also in dosage or intensity.

4.2.4 Subgroup-Specific Generalized Treatment Effects (SGTTs)

In previous discussions, the treatment group g was assumed to exit treatment uniformly, implying that all units within the group followed an identical treatment trajectory over time. However, in many empirical settings, units within the same treatment group may experience heterogeneous treatment patterns. Specifically, within group g, some units may remain treated throughout, others may exit treatment at certain periods, and some may re-enter treatment after a period of exit.

To account for these varied treatment patterns, I define three subgroup-specific generalized treatment effects (SGTTs) for each treatment group g at time t, as follows:

$$SGTT(g,t,\alpha) = E[Y_{it}(g) - Y_{it}(\infty)|D_g = 1, \alpha], \quad \alpha \in \{Treated,\ Exit,\ Re-entry\} \tag{4.8}$$

Each SGTT captures the generalized treatment effect for units that remain treated at time t, for units that have exited treatment by time t, and for units that have re-entered treatment at time t.

Similarly, at the first stage, following (2023, Procedure 4.1), estimate the subgroup-specific generalized treatment effects $SGTT(g,t,\alpha)$. These SGTTs will subsequently be utilized in the second stage of the estimation procedure, where these effects are aggregated based on the observed treatment status.

$$WGTT(\alpha, r) = \sum_{g_\alpha \in G_{r,\alpha}} w(g_\alpha, r) \cdot SGTT(g, g+r, \alpha),\ where\ \ r = t - g \tag{4.9}$$

where $w(g_\alpha, r) = \frac{N_{g\alpha}}{N_{G_{r,\alpha}}}$ denotes the weight assigned to group g in state $\alpha \in \{Treated, Exit, Re-entry\}$ at relative time r.

4.3 Moderating Effects

This section extends the rolling transformation framework in Lee and Wooldridge (2023) to incorporate moderating effects—treatment effect heterogeneity driven by observable unit characteristics. Such heterogeneity arises when the impact of a treatment varies across subgroups defined by factors such as income, race, gender, or baseline access levels. For instance, in the Empirical Application section, I consider two subgroups: $b_i = 1$ if a town has more than 20% of residents aged 65 or older, $b_i = 0$, otherwise. The treatment effect may differ across these subgroups.

To address treatment effect heterogeneity, I outline a Two-stage estimation procedure for a generalized version of the Inverse Probability Weighted Regression Adjustment (IPWRA) method, accommodating staggered intervention settings with multiple treatment groups.

Stage 1: Propensity Score Estimation

To account for staggered treatment timing, we estimate the propensity score using a multinomial logit model, where the treatment status D_g is defined as the first treatment period $g \in \{S, S+1, \ldots, T\}$. The propensity score for each unit i is given by:

$$p_g(\mathbf{X}) = \Pr(D_g = 1|\mathbf{X}) = \frac{\exp(\mathbf{X}'\delta_g)}{1 + \sum_{k=1}^{G} \exp(\mathbf{X}'\delta_k)} \tag{4.10}$$

where δ_g are the parameters estimated for each treatment group g. The control group's (i.e., $g = \infty$) coefficient is normalized to zero.

Stage 2: IPWRA Estimation with Moderating Effects

After obtaining the estimated propensity scores $\widehat{p}_g(\mathbf{X_i})$, I proceed to the IPWRA estimation that incorporates moderating effects. Specifically, the treatment effect for group g at time t is estimated by solving the following weighted least squares problem:

$$\underset{\tau_{g,t},\alpha,\beta,\gamma_{g,t}}{\arg\min} \sum_{i=1}^{N} \left(D_g + \frac{(1-D_g)\widehat{p}_g(\mathbf{X})}{1-\widehat{p}_g(\mathbf{X})} \right) \left(\dot{Y}_{g,t} - \tau_{g,t} \cdot D_g - \alpha - X_i^{'}\beta - D_g(X_i - \bar{X}_g)^{'} \cdot \gamma_{g,t} \right)^2 \tag{4.11}$$

Here, $\gamma_{g,t}$ captures the moderating effects—heterogeneous treatment effects that vary with deviations from the treated group's average covariates $\bar{\mathbf{X}}_g$. This formulation allows for a flexible and interpretable estimation of treatment effect heterogeneity across subgroups.

Importantly, this structure reproduces the same point estimate for τ as the baseline IPWRA estimator in Lee and Wooldridge (2023), given by:

$$\underset{\tau_{g,t},\alpha,\beta}{\arg\min} \sum_{i=1}^{N} \left(D_g + \frac{(1-D_g)\widehat{p}_g(\mathbf{X})}{1-\widehat{p}_g(\mathbf{X})} \right) \left(\dot{Y}_{g,t} - \tau_{g,t} \cdot D_g - \alpha - X_i^{'}\beta \right)^2 \tag{4.12}$$

However, the inclusion of the γ coefficients provides additional insights by quantifying the extent to which the treatment effect differs across levels of baseline covariates, thereby offering a more comprehensive understanding of policy impacts across heterogeneous subgroups.

This framework offers a practical advantage over standard built-in command in `Stata` such as the `teffects ipwra`, which does not support estimation of moderating effects. By applying the transformed-outcome approach and two-stage IPWRA estimation procedure, we can estimate moderating effects while leveraging the efficiency of doubly robust estimation.

4.3.1 Comparison with the Pooled OLS Estimator

Wooldridge (2021) proposes a pooled OLS method for estimating treatment effect heterogeneity over time, while also allowing for the identification of moderating effects. For illustration, consider the common timing case in which all treated units received their initial treatment at the same period $t = q$. The specification includes a full set of time-fixed effects and interaction terms between the treatment indicator and covariates, including demeaned covariates (centered at the treated group mean). As shown in Equation (5.46) of Wooldridge (2021), the pooled OLS regression includes terms of the form:

$$\begin{aligned} y_{it} \text{ on } 1,\ & d_i,\ \dot{\mathbf{x}}_i,\ d_i \cdot \dot{\mathbf{x}}_i,\ f_{q,t},\ \ldots,\ f_{T,t},\ f_{q,t} \cdot \dot{\mathbf{x}}_i,\ \ldots,\ f_{T,t} \cdot \dot{\mathbf{x}}_i, \\ & d_i \cdot f_{q,t},\ \ldots,\ d_i \cdot f_{T,t},\ d_i \cdot f_{q,t} \cdot \dot{\mathbf{x}}_i,\ \ldots,\ d_i \cdot f_{T,t} \cdot \dot{\mathbf{x}}_i, \end{aligned} \tag{4.13}$$

where $f_{q,t}, \ldots, f_{T,t}$ denote time indicators such that $f_{t',t} = 1$ if $t = t'$ and 0 otherwise, where $t' \in \{q, \ldots, T\}$, post-treatment periods. The covariates $\dot{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}_1$ represent deviations from the treated group's mean covariate values. The terms $d_i \cdot f_{t,t} \cdot \dot{\mathbf{x}}_i$ capture period-specific moderating effects, allowing the treatment effect to vary flexibly with covariates across time. Each coefficient on these triple interaction terms identifies $\gamma(t)$, the marginal effect of covariates on the treatment effect in period t. While straightforward and easy to implement, this regression-based approach relies on correct specification of the outcome model. If relevant interaction or nonlinear terms from the true model—which are rarely known to researchers in practice—are omitted, the estimator may suffer from bias due to model misspecification.

In contrast, my proposed two-stage IPWRA estimator provides greater robustness through its doubly robust structure. It remains consistent as long as either the outcome model or the propensity score model is correctly specified. Importantly, it also enables the estimation of moderating effects

even when the outcome model is misspecified.

In Section 4.4, I demonstrate through simulation studies that the two-stage IPWRA approach successfully recovers both the average treatment effects on the treated (ATTs) and the corresponding moderating effects. Compared to POLS, the two-stage IPWRA yields substantially lower bias and RMSE for heterogeneous treatment effects under outcome model misspecification, while maintaining accurate estimation of average treatment effects. These findings underscore the practical value of the proposed method for evaluating differential policy impacts across heterogeneous populations, especially in settings where model misspecification is a concern.

4.4 Monte Carlo Simulations

This section outlines the data generating process (DGP) employed in the simulation study to evaluate the performance of the proposed two-stage IPWRA estimator described in Section 4.3. Although the study primarily focuses on staggered intervention settings, the simulation design is restricted to a common timing case for simplicity. The findings demonstrate that the proposed two-stage IPWRA estimator successfully incorporates both moderating and treatment effects. Regarding treatment effects, the estimator accurately replicates the baseline IPWRA estimates in Lee and Wooldridge (2023).

4.4.1 Data Generating Process (DGP)

The data generating process (DGP) is structured to simulate treatment effects under a common timing framework, involving a single treatment group and a control group. The DGP is replicated following the procedure outlined in **Appendix 2D** of Lee and Wooldridge (2023), and the relevant variables are defined as described below.

The dataset spans six time periods ($T = 6$), representing the years 2001 to 2006, with the treatment initiated at time $S = 4$. The control group ($D_i = 0$) consists of never-treated units, while the treatment group ($D_i = 1$) includes units treated at $S = 4$.

The detailed structure of the variables is as follows. I generate two time-invariant covariates $\mathbf{X} = (X_1, X_2)$,

$$X_1 \sim Gamma(2, 2) \quad \text{and} \quad X_2 \sim Bernoulli(0.6)$$

The treatment indicator is defined through a logistic propensity score model:

$$p(\mathbf{X}) = \Pr(D = 1 \mid \mathbf{X}) = \frac{\exp(\mathbf{X}'\mathbf{z})}{1 + \exp(\mathbf{X}'\mathbf{z})}, \tag{4.14}$$

where the index function is specified as $\mathbf{X}'\mathbf{z} = -1.2 + \frac{(X_1-4)}{2} - X_2$.

To investigate the moderating effects, the treatment effects are generated to exhibit heterogeneity not only across time but also over covariates $\mathbf{X}$. The treatment effect function is specified as follows:

$$\tau_r(\mathbf{X}) = \theta \cdot \sum_{r=S}^{T} (r - S + 1)^{-1} + \lambda_r \cdot h(\mathbf{X}), \tag{4.15}$$

where $\theta = T - S + 1$, $\lambda_r = (0.5, 0.6, 0.8)$. The functional form of $h(\mathbf{X})$ is:

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

In this specification, $\theta \cdot \sum_{r=S}^{T} (r - S + 1)^{-1}$ captures the baseline treatment effect accumulated over time, while $\lambda_r \cdot h(\mathbf{X})$ represents the period-specific moderating effect, allowing treatment effects to vary with covariates across time.

Lastly, I define a potential outcome in the control state:

$$Y_t(0) = \delta_t + \alpha_i + \beta_t \cdot f(\mathbf{X}) + U_t(0), \tag{4.16}$$

where $\delta_t = t$, $\alpha_i \sim \mathcal{N}(2, 1)$, and $U_t(0) \sim \mathcal{N}(0, 4)$. The coefficient vector β' is set as:

$$\beta' = (1.0, 1.5, 0.8, 1.5, 2, 2.5).$$

$f(\mathbf{X})$ has two functional forms:

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2}, \tag{4.17}$$

$$f(\mathbf{X}) = \frac{(X_1 - 4)}{3} + \frac{X_2}{2} + \frac{(X_1 - 4)^2}{3} + (X_1 - 4) \cdot \frac{X_2}{4} \tag{4.18}$$

For the simulation studies, I assume that the outcome model is linear in $\mathbf{X}$. Under this specification, the conditional mean function $E(Y|\mathbf{X})$ is correctly specified when using equation (4.17), but becomes misspecified under equation (4.18) due to the inclusion of interaction and quadratic

terms in the true data-generating process. However, since the IPWRA estimator is doubly robust, it remains consistent as long as the propensity score model is correctly specified. Therefore, even under outcome model misspecification, I expect the IPWRA estimator to accurately recover the treatment effect estimates.

Finally, a potential outcome in the treated state is

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

where $U_t(1) \sim \mathcal{N}(0, 4)$.

4.4.2 Simulation Results

Table 4.2 presents the simulation results under correct specification of the outcome model, where the conditional mean function $E(Y|\mathbf{X})$ is linear in covariates. I compare the performance of four estimators: Pooled OLS (POLS) following Wooldridge (2021), the baseline IPWRA proposed in Lee and Wooldridge (2023), the doubly robust estimator developed by Callaway and Sant'Anna (2021) (denoted as CS), and the proposed two-stage IPWRA–these estimators are designed to estimate average treatment effects on the treated (ATTs), particularly in the presence of treatment effect heterogeneity across units and time. Estimator performance is evaluated based on bias, standard deviation (SD), and root mean squared error (RMSE), focusing on both average treatment effects (τ_4, τ_5, τ_6) and moderating effects associated with covariates x_1 and x_2.

Under correct specification, all estimators perform reasonably well in recovering the average treatment effects on the treated (ATT). Both the baseline IPWRA and two-stage IPWRA estimators yield identical estimates with virtually zero bias and low RMSE across all time periods. This confirms that the two-stage extension preserves the desirable statistical properties of the standard IPWRA estimator when estimating group-time average treatment effects.

Table 4.2 Simulation Results Under Correct Specification of $E(Y|\mathbf{X})$

	τ_4			τ_5			τ_6		
Sample ATT	3.220			4.753			5.838		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
POLS (RA)	-0.002	0.395	0.395	-0.013	0.410	0.410	-0.008	0.413	0.413
CS	-0.003	0.503	0.503	-0.015	0.503	0.504	-0.009	0.518	0.518
IPWRA	-0.001	0.403	0.403	-0.014	0.413	0.413	-0.007	0.421	0.421
Two-Stage IPWRA	-0.001	0.403	0.403	-0.014	0.413	0.413	-0.007	0.421	0.421

	$\gamma(X_1, 4)$			$\gamma(X_1, 5)$			$\gamma(X_2, 6)$		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample Moderating Effects	0.250			0.300			0.400		
POLS	-0.006	0.324	0.325	0.000	0.328	0.328	0.001	0.332	0.332
Two-Stage IPWRA	-0.006	0.370	0.370	-0.002	0.361	0.361	0.003	0.365	0.365
	$\gamma(x_2, 4)$			$\gamma(x_2, 5)$			$\gamma(x_2, 6)$		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample Moderating Effects	0.167			0.200			0.267		
POLS	0.038	0.803	0.804	0.037	0.811	0.811	-0.023	0.836	0.837
Two-Stage IPWRA	0.030	0.848	0.849	0.045	0.850	0.851	-0.019	0.869	0.869

Importantly, both POLS and the two-stage IPWRA also demonstrate strong performance in estimating the moderating effects $\gamma(x_1, t)$ and $\gamma(x_2, t)$. For example, when the true moderating effect is 0.25 for $\gamma(x_1, 4)$, the two-stage IPWRA achieves a bias of 0.002 and an RMSE of 0.370, while POLS also performs comparably well. This is expected, as the outcome model is correctly specified in both frameworks.

Table 4.3 presents simulation results under misspecification of the outcome model, where the true conditional mean function $E(Y|\mathbf{X})$ includes nonlinear interaction terms that are omitted in estimation. In this setting, while the propensity score model remains correctly specified, the outcome model deviates from the true data-generating process. This setup provides a useful test for evaluating the robustness of each estimator.

The upper panel of the table reports estimates for the average treatment effects on the treated (τ_4, τ_5, τ_6). Since the propensity score model is correctly specified, both the baseline IPWRA and the two-stage IPWRA recover the treatment effects with negligible bias and low RMSE—consistent with the doubly robust property of IPWRA. Notably, this property holds even when the outcome model is misspecified, confirming that the two-stage extension maintains the validity and consistency of the original IPWRA framework in estimating ATT.

Table 4.3 Simulation Results under Misspecification of $E(Y|\mathbf{X})$

	τ_4			τ_5			τ_6		
Sample ATT	3.220			4.753			5.838		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
POLS (RA)	0.044	0.397	0.400	0.091	0.418	0.428	0.154	0.429	0.456
CS	-0.003	0.509	0.510	-0.015	0.517	0.517	-0.008	0.543	0.543
IPWRA	0.000	0.405	0.405	-0.011	0.417	0.417	-0.002	0.431	0.431
Two-Stage IPWRA	0.000	0.405	0.405	-0.011	0.417	0.417	-0.002	0.431	0.431

	$\gamma(x_1,4)$			$\gamma(x_1,5)$			$\gamma(x_1,6)$		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample Moderating Effects	0.250			0.300			0.400		
POLS	0.155	0.329	0.364	0.362	0.344	0.500	0.564	0.380	0.681
Two-Stage IPWRA	0.002	0.374	0.374	0.016	0.391	0.391	0.031	0.438	0.439
	$\gamma(x_2,4)$			$\gamma(x_2,5)$			$\gamma(x_2,6)$		
	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Sample Moderating Effects	0.167			0.200			0.267		
POLS	0.114	0.807	0.815	0.208	0.819	0.845	0.245	0.855	0.889
Two-Stage IPWRA	0.028	0.853	0.853	0.040	0.866	0.867	-0.027	0.905	0.905

The distinction between estimators becomes more pronounced when examining moderating effects. The two-stage IPWRA continues to estimate $\gamma(x_1,t)$ and $\gamma(x_2,t)$ with minimal bias and relatively low RMSE. In contrast, the POLS estimator suffers from substantial bias in nearly every case, particularly when the true moderating effect is large. For example, when the true value of $\gamma(x_1,6)$ is 0.4, POLS overestimates it by more than 0.56, yielding an RMSE of 0.681, whereas the two-stage IPWRA produces a much smaller bias of 0.031 and an RMSE of 0.439.

This pattern holds across other periods and covariates. As a regression-based estimator, POLS is sensitive to outcome model misspecification—particularly when covariate interactions or nonlinear terms from the true model are omitted. In contrast, the two-stage IPWRA leverages its doubly robust property and remains consistent as long as the propensity score is correctly specified. By incorporating covariate-treatment interactions within a weighted least squares framework, it flexibly recovers heterogeneous effects without requiring correct specification of the outcome model.

In summary, the two-stage IPWRA estimator not only replicates the average treatment effect estimates of the standard IPWRA under outcome model misspecification, but also provides a reliable and robust framework for estimating heterogeneous treatment effects—something the standard IPWRA is not equipped to do. This makes it particularly valuable in empirical settings where

treatment effect heterogeneity is of interest, but the outcome model's functional form is uncertain or potentially misspecified.

4.5 Empirical Application

In this section, I revisit Kim (2023), who studies the impact of chain pharmacy entry on the number of local independent pharmacies in rural townships. While the primary contribution of that paper lies in modeling the strategic interaction between chain and independent pharmacies using a static game framework, the author first presents reduced-form evidence using recent staggered difference-in-differences (DID) methods. Although Kim (2023) notes that chain pharmacies frequently enter and exit markets, the analysis ultimately treats entry as a one-time event, citing the absence of a suitable framework to handle dynamic path of treatment exposure.

However, my estimation strategy directly addresses this limitation by allowing treatment status to vary over time–specifically, to turn on, off, and back on–offering a more realistic depiction of competitive exposure in local markets. Through a two-stage estimation procedure, I estimate seperate effects for three different observed treatment states, *Treated, Exit, Re-entry*.

I also focus on a key source of treatment effect heterogeneity: the share of the elderly population in a township. My framework extends the standard doubly-robust estimator—combining inverse probability weighting with regression adjustment (IPWRA)—to incorporate moderating effects. Following Kim (2023), I classify townships into two groups based on their elderly population share in the year 2000: (i) *high elderly population towns* (those with $\geq$ 20% of residents aged 65 or older), and (ii) *non-high elderly population towns* (those with $<$ 20% aged 65+). This approach enables the estimation of *subgroup-specific dynamic treatment effects*, facilitating a more nuanced understanding of how the competitive effects of chain entry vary across demographic contexts. The method not only assesses whether average effects differ across subgroups but also tracks how those effects evolve over time within each group.

4.5.1 Data and Setting

The empirical analysis employs the dataset constructed by Kim (2023), which combines the *Data Axle Historical Business Database* (1997–2021)—a comprehensive record of business estab-

lishments, including pharmacy operations across the United States—with demographic variables from the U.S. Census and the American Community Survey (ACS). The data are matched at the township-year level and focus on rural townships in the Midwestern United States, where access to pharmacies is a key healthcare concern.

Independent pharmacies in rural areas often serve as more than just retail outlets; they function as essential community hubs, providing prescription fulfillment, over-the-counter medications, and informal health advice for minor ailments. since the 1970s, however, the pharmacy landscape has been reshaped by the rapid expansion of chain pharmacies such as Walgreens, CVS, and Rite Aid, as well as mass merchandisers like Walmart and Target. These developments have threatened the survival of locally owned pharmacies, especially in underserved rural areas where the closure of a single pharmacy can have substantial consequences for healthcare access.

The geographic unit of analysis is the township. Each township is characterized by the number of active independent pharmacies, the presence of chain pharmacies within a 15-mile radius, and market characteristics such as population size, elderly share, and poverty rate. The outcome of interest is the number of independent pharmacies operating in each township-year. The treatment variable is a binary indicator equal to one if at least one chain pharmacy is present within 15 miles of the township in year t.

While Kim (2023) includes 802 towns in the original dataset, I restrict the sample to 596 townships by excluding those that were already exposed to chain pharmacies in the initial year. This restriction ensures that treatment does not occur at baseline; the number of chain presence (within 15 mi) is zero across all units in the initial year. The resulting panel covers the periods 2000 through 2019. For further details on township definitions and sampling criteria, see Section 2.2, *Final sample*, in Kim (2023).

Table 4.4 presents summary statistics for the 596 townships included in the analysis, covering a total of 11,920 township-year observations from 2000 to 2019. All summary measures, unless otherwise noted, reflect township characteristics in the initial year of the panel, 2000.

Again, townships are classified as “High Elderly” if 20% or more of the population was aged

Table 4.4 Summary Statistics of Full sample in Year 2000

	Full Sample	**Non-High Elderly**	**High Elderly**
Number of Townships	596	181	415
Number of Observations	11,920	3,820	8,100
Independent Pharmacies (mean)	0.865	0.674	0.949
Chain Presence (within 15 mi, mean, in 2000)	0	0	0
(averaged over 2000–2019)	0.26	0.37	0.22
Log(Population) (mean)	7.16	7.29	7.10
Elderly Share (Aged 65 or older, mean)	22.9%	16.1%	26.0%
Poverty Share (mean)	12.4%	13.9%	11.6%

65 or older in 2000. Based on their elderly population share in 2000, 415 townships (68%) are classified as *high elderly population* and 181 (32%) as *non-high elderly population*. On average, high elderly population townships have smaller populations than their non-high elderly counterparts, which is often associated with lower market demand. Since chain pharmacies are more likely to enter markets with greater demand and population density, chain presence is more common in non-high elderly population townships over the full sample period (2000–2019), with an average presence rate of 37% compared to 22% in high elderly townships. Nevertheless, in the initial year of the panel (2000), high elderly population townships already exhibited a higher average number of independent pharmacies. This highlights the essential role that independent pharmacies have historically played in these communities—particularly in serving older populations and addressing healthcare needs in areas.

4.5.2 Results

Table 4.5 presents the estimated weighted generalized treatment effects, $WGTT(\alpha, r)$, over relative time r for each treatment state $\alpha \in \{\text{Treat, After Exit, After Re-entry}\}$. Each estimate is accompanied by a 95% bootstrap confidence interval. While the proposed framework allows for various treatment effect estimators, the results reported here are obtained using the rolling regression adjustment (RA) estimator, following Lee and Wooldridge (2023, Procedure 4.1) with each subsample categorized according to its respective treatment state.

Table 4.5 Estimated Weighted General Treatment Effects $WGTT(\alpha, r)$

	Estimates (95% CI, bootstrapped)		
Relative Time, r	α = Treat	α = After Re-entry	α = After Exit
0	-0.355 (-0.427, -0.283)	-0.291 (-0.418, -0.165)	-0.253 (-0.364, -0.143)
1	-0.368 (-0.449, -0.287)	-0.33 (-0.448, -0.212)	-0.257 (-0.374, -0.141)
2	-0.389 (-0.487, -0.29)	-0.307 (-0.459, -0.155)	-0.212 (-0.368, -0.057)
3	-0.339 (-0.441, -0.237)	-0.317 (-0.47, -0.164)	-0.247 (-0.42, -0.074)
4	-0.34 (-0.452, -0.229)	-0.286 (-0.424, -0.148)	-0.166 (-0.354, 0.022)
5	-0.344 (-0.472, -0.215)	-0.283 (-0.435, -0.132)	-0.249 (-0.454, -0.044)
6	-0.347 (-0.482, -0.212)	-0.319 (-0.476, -0.163)	-0.236 (-0.521, 0.05)
7	-0.375 (-0.538, -0.212)	-0.291 (-0.467, -0.114)	-0.116 (-0.556, 0.325)
8	-0.389 (-0.531, -0.246)	-0.367 (-0.571, -0.162)	-0.161 (-0.652, 0.331)
9	-0.324 (-0.527, -0.121)	-0.355 (-0.538, -0.171)	-0.275 (-0.71, 0.159)
10	-0.339 (-0.547, -0.131)	-0.315 (-0.506, -0.124)	-0.313 (-0.875, 0.25)

Figure 4.1 visualizes these results, showing dynamic treatment effect patterns by subgroup over time using the rolling regression adjustment (RA) estimator following LW (2023). The y-axis represents $WGTT(\alpha, r)$, and the x-axis denotes relative time since initial treatment, exit, or re-entry.

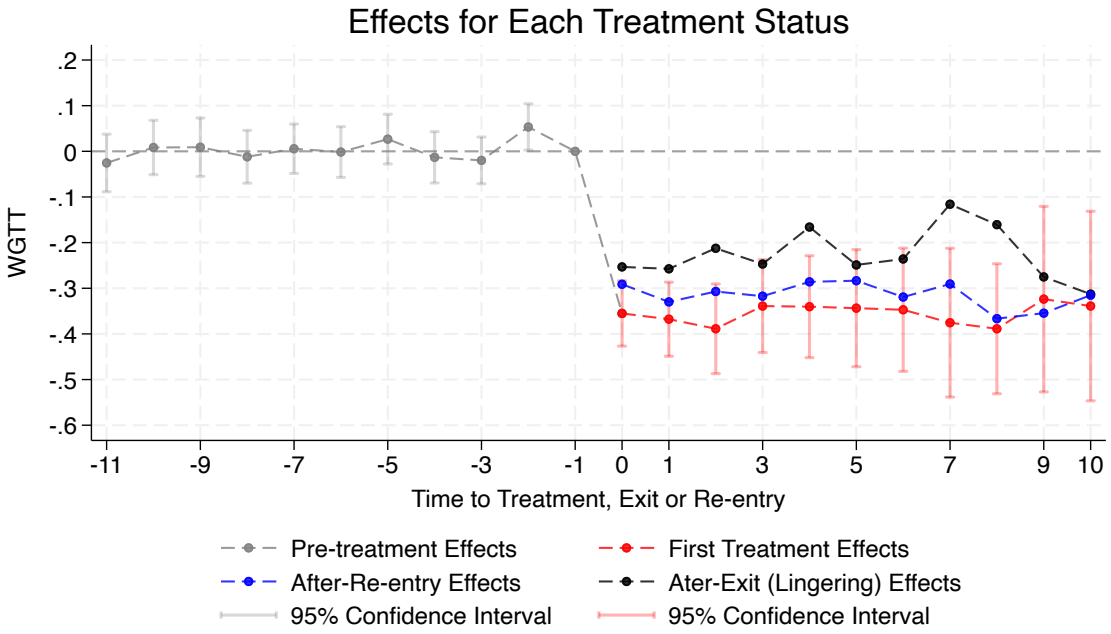


Figure 4.1 WGTT Estimates Using Rolling Regression Adjustment Estimator

Overall, the effects are consistently negative and statistically significant for the *Treat* group, indicating a persistent adverse impact on local pharmacy access. In contrast, the effects for the *After*

Exit group diminish over time and become statistically insignificant, suggesting potential recovery. For the *After Re-entry* group, the effects remain negative but are somewhat attenuated, implying that re-entry still harms access, though less severely than initial exposure.

These patterns likely reflect underlying market dynamics. The initial closure may involve substantial financial losses and signal poor profitability, discouraging future entrants. In small rural markets with thin margins, the departure of a chain pharmacy may not be enough to restore viable conditions for independent pharmacies. Thus, exit does not necessarily reverse the decline in access, highlighting the persistent and asymmetric nature of competitive disruption.

The attenuation observed for the *Re-entry* group (blue line) may reflect adaptive responses by independent pharmacies, market saturation, or reduced vulnerability after prior exposure.

Together, these results underscore the importance of modeling treatment as a dynamic, evolving process. By distinguishing between initial exposure, exit, and re-entry phases, the proposed estimator captures nuanced, heterogeneous treatment effects that would be obscured under static or binary frameworks.

4.6 Concluding Remarks

Building on the rolling approach developed by Lee and Wooldridge (2023), this paper extends the framework to accommodate dynamic treatment paths—including not only initial treatment but also treatment exit and re-entry—thereby offering a generalized method for evaluating complex intervention patterns over time.

In addition, I propose a two-stage IPWRA estimator that allows for the identification of moderating effects, enabling researchers to assess how treatment effects vary across subpopulations defined by covariates. By incorporating covariate-treatment interactions into a weighted least squares framework and leveraging the doubly robust structure of IPWRA, the proposed method recovers moderating effects even when the conditional outcome model is misspecified, as long as the propensity score is correctly specified.

Simulation results demonstrate that the two-stage IPWRA estimator performs well in recovering not only average treatment effects on the treated (ATTs) but also moderating effects, under both

correct and misspecified outcome models. In particular, for moderating effects, it outperforms the pooled OLS estimator when the outcome model is misspecified. These findings underscore the robustness and flexibility of the proposed approach in evaluating differential policy impacts across heterogeneous populations.

To illustrate its empirical relevance, I apply the proposed method to examine the effects of pharmacy chain entry on independently owned pharmacies. Beyond estimating the average treatment effect of initial chain entry, the analysis also uncovers dynamic treatment patterns following exit and re-entry. Specifically, treatment effects in the *After-Exit* phase remain persistently negative, indicating that the departure of chain pharmacies does not reverse the decline in access to independent pharmacies. These lingering effects suggest long-term disruptions in local markets, particularly in rural areas with limited demand.

In contrast, treatment effects in the *Re-entry* phase—reflecting repeated exposures to chain competition—are more attenuated, potentially due to adaptive responses or market saturation. These results highlight the importance of modeling treatment as a dynamic process and demonstrate the value of the proposed framework in capturing complex and evolving treatment effects that would be overlooked in static or binary settings.

While the current paper focuses on binary treatment settings, the framework is naturally extensible to discrete or continuous treatment intensities, including dynamic transitions such as treatment exit and re-entry. Future work will build on this foundation to further expand the applicability of the estimator in complex policy environments.

BIBLIOGRAPHY

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, page rdae007.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

De Chaisemartin, C. and d'Haultfoeuille, X. (2024). Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*, pages 1–45.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, 110(9):2964–2996.

Kim, H. (2023). Rural pharmacy access and competition: Static games with machine learning. *Available at SSRN 4377695*.

Lee, S. J. and Wooldridge, J. M. (2023). A simple transformation approach to difference-in-differences estimation for panel data. *Available at SSRN 4516518*.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, 225(2):175–199.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.