TI 2023-004/III
Tinbergen Institute Discussion Paper

Time-Weighted Difference-in-Differences: Accounting for Common Factors in Short T Panels

Timo Schenk[1]

[1] University of Amsterdam and Tinbergen Institute

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Time-Weighted Difference-in-Differences: Accounting for Common Factors in Short T Panels

Timo Schenk[*]

February 2, 2023

Abstract

This paper proposes a time-weighted difference-in-differences (TWDID) estimation approach that is robust against interactive fixed effects in short T panels. Time weighting substantially reduces both bias and variance compared to conventional DID estimation through balancing the pre-treatment and post-treatment unobserved common factors. To conduct valid inference on the average treatment effect, I develop a correction term that adjusts conventional standard errors for weight estimation uncertainty. Revisiting a study on the effect of a cap-and-trade program on NOx emissions, TWDID estimation reduces the standard errors of the estimated treatment effect by 10% compared to a conventional DID approach. In a second application I illustrate how to implement TWDID in settings with staggered adoption of the treatment.

Keywords: synthetic difference-in-differences, dynamic treatment effects, interactive fixed effects, panel data

1 Introduction

The presence of interactive fixed effects in the untreated potential outcomes leads to biased difference-in-difference (DID) estimates of average treatment effects. While the estimators of Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) and Chan and Kwok (2021) address this issue in large T panels, the question remains how to account for common factors in short T panels.

[*] Amsterdam School of Economics, University of Amsterdam & Tinbergen Institute (`t.d.schenk@uva.nl`). I thank my advisors Frank Kleibergen, Andreas Pick as well as Pedro Sant'Anna, Arturas Juodis and conference and seminar participants of the 2022 International Econometrics PhD Conference (Rotterdam), the Amsterdam Panel Data Workshop, ESEM, IAAE, IPDC, the NESG Meeting, and internal seminars for helpful comments and discussions. Financial support by the IAAE travel grant is highly acknowledged.

In this paper, I suggest a time-weighted DID (TWDID) approach to estimate the average treatment effect on the treated (ATT) of a binary treatment. Assume that at least two pre-treatment periods and a group of untreated units are observed. As this paper shows, weighting pre-treatment observations according to similarity between pre- and post-treatment outcomes makes estimation and inference robust against the presence of interactive fixed effects even when only a few number of pre-treatment periods are available.

Interactive fixed effects are time-varying, unobserved common factors that have a time-invariant but heterogeneous effect on the outcome of interest. DID estimation compares pre- and post-treatment averages of treated and untreated units. This results in a biased estimate of the ATT if the pre-treatment factors differ from the post-treatment factors and their effect on the outcome correlates with the treatment assignment. As a consequence, DID approaches require the latter correlation to be zero by imposing a parallel trend assumption, which can lead to pre-testing issues (Roth, 2022).

The TWDID approach, instead, allows for non-parallel trends. Under strict balancing conditions on the factors it completely eliminates the bias. In practice, weights that are estimated from the control unit data succeed in reducing the bias substantially, even when the strict balancing conditions are not met.

A second effect of the factor imbalance is that it amplifies the variance of the DID estimator even when the parallel trend condition holds. By balancing the factors, time weighting reduces the variance and leads to more accurate estimates. In fact, when the number of units is large compared to the number of periods, the estimated weights converge to pseudo-true weights which minimize the variance of the estimated treatment effect. Simulations show that the amount by which the variance is reduced outweighs the additional variance caused by the weight estimation uncertainty.

As a consequence of the bias, inference based on DID estimation is substantially oversized, which the bias reduction of TWDID alleviates. However, the presence of estimated weights still leads to empirical size in excess of the nominal size when using standard covariance estimators. To eliminate the remaining size distortions, I provide analytical two-step standard errors (Newey and McFadden, 1994) obtained from the asymptotic variance of the TWDID estimator. First, the cluster-covariance matrix estimator (Arellano, 1987) is applied to the weighted sample to estimate the variance of the estimated treatment effect under pseudo-true weights. For the second step, I develop a correction term that accounts for the presence of estimated weights. It uses the fact that, in short T panels, the time weights are asymptotically normal around the pseudo true weights.

TWDID circumvents another issue of DID estimation, which occurs with staggered adoption and heterogeneous treatment effects. When implementing DID with a simple two-way fixed effects regression, units which have already received treatment effectively still act as a control group. In general, this will not return an interpretable average treatment effect (Borusyak, Jaravel, and Spiess, 2022; Goodman-Bacon, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2020; Imai and Kim, 2021; Athey and Imbens, 2022). Instead, the TWDID approach requires the researcher to explicitly define a control group, which prevents potentially misleading comparisons. In particular, TWDID can be used to estimate group-time average treatment average effects, as Callaway and Sant'Anna (2020) suggest. These can then be aggregated to measures of interest that are interpretable when treatment effects are heterogeneous.

I revisit two applications to show how to implement the TWDID estimator in practice. First, I consider a study by Deschenes, Greenstone, and Shapiro (2017), who use a triple DID design to estimate the effect of a cap-and-trade program on NOx emissions. Interactive fixed effects are present here if common shocks (e.g. business cycle or weather) have heterogeneous effects on the NOx emissions in different counties. While the point estimates differ only mildly, TWDID reduces the standard errors of the estimated average treatment effect by 10% compared to a conventional DID approach.

Second, I consider the application of Callaway and Karami (2022), who estimate the effect of early-career job displacement on earnings using data from the 1979 National Longitudinal Study of Youth (NLSY). Here the treatment has staggered adoption, because different workers have their first job displacement at different points in time. TWDID estimates suggest that displacement has lead to a smaller, but still statistically significant reduction in earnings.

The TWDID estimator can be viewed as a restricted version of the synthetic DID (SDID) estimator of Arkhangelsky et al. (2021). The SDID estimator uses time weights alongside unit weights to balance unobserved factor structures in large N, T panels. Standard errors are obtained using panel-jackknife or bootstrap methods, which tend to be conservative. Although SDID requires large T for consistency, in the Monte Carlo experiments of this paper it performs well in terms of bias and variance even with only a few pre-treatment periods. In fact, when strict balancing conditions on the factors are not met, SDID can yield more precise estimates than TWDID. However, due to the conservative standard errors, the resulting confidence intervals are still wider than those of TWDID.

This paper also relates to a number of findings in the literature on synthetic control (SC) estimators (Abadie, Diamond, and Hainmueller, 2010; Xu, 2017; Abadie and L'hour, 2020; Ferman, 2021; Ferman and Pinto, 2021; Ben-Michael, Feller, and Rothstein, 2021). They use unit weights to balance time-invariant unobserved characteristics between treated and untreated units, which are estimated with a time-series regression over the pre-treatment periods. Consistent estimation requires a large number of pre-treatment periods, strict balancing conditions on the loadings and restrictions on the serial dependence of the errors. Similarly, TWDID requires a large number of control units and balancing conditions on the factors. Exploiting independence over the cross-section, however, TWDID estimation remains reliable when the data exhibits strong serial dependence.

There are also GMM-type approaches to identify and estimate treatment effects in short T panels with interactive fixed effects, see for example Brown and Butts (2022). These approaches make use of instruments to address the endogeneity of the factor structure and therefore require fewer restrictions on the factors themselves. Callaway and Karami (2022) propose such an approach using time-invariant observable covariates with constant effect on the outcome as instruments. However, these instruments are not always available in practice.

When both N and T are large, one can use well established results from linear panel data models with factor structures (Pesaran, 2006; Bai, 2009) to recover a long run average treatment effect. For example, Gobillon and Magnac (2016) apply the estimator of Bai (2009) to estimate the average treatment effect jointly with the factor structure. Using principal component analysis, Chan and Kwok (2021) construct factor proxies, which can then be used in a factor-augmented regression. In short T panels, however, these estimators are generally inconsistent.

The remainder of the paper is structured as follows. Section 2 covers the Theory. Section 2.1 introduces the interactive fixed effects model and defines the TWDID estimator. Section 2.2 shows the bias and variance reduction properties. Section 2.3 covers inference. Sections 2.4 and 2.5 extend the results to settings with multiple treated periods and staggered adoption, respectively. Section 3 illustrates the theoretical results with simulations. Section 4 contains the applications and Section 5 concludes.

2 Theory

2.1 Setting

Using a panel data set for treated and untreated units, we wish to estimate the effect of a policy intervention starting in period $t = T_0$. That is, we seek to estimate the average treatment effect on the treated

$$\tau_t := ATT_t = \mathrm{E}[y_{it}(1) - y_{it}(0)|D_i = 1] \tag{1}$$

in the post-treatment periods $t = T_0 + 1, \ldots, T$, where $y_{it}(1), y_{it}(0)$ are the potential outcomes of unit i in period t, and $D_i \in \{0, 1\}$ indicating whether unit i is ever treated. The researcher observes $y_{it} = D_i y_{it}(1) + (1 - D_i) y_{it}(0)$ for a large number of units $i = 1, \ldots, N$ and a small number of periods $t = 1, \ldots, T$, covering at least two pre-treatment periods ($T_0 \geq 2$).

The untreated potential outcomes are generated by an interactive fixed effects model,

$$y_{it}(0) = \beta_i + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \varepsilon_{it} \tag{2}$$

where β_i are unit fixed effects, $\boldsymbol{f}_t$ and $\boldsymbol{\lambda}_i$ are r-dimensional vectors of common factors and loadings, and ε_{it} is an idiosyncratic error component. Such unobserved factor structures, $\boldsymbol{\lambda}_i' \boldsymbol{f}_t$, are present in many economic settings. In microeconomic applications, $\boldsymbol{\lambda}_i$ can be thought of a vector of unobserved, time-invariant characteristics of individual i. In contrast to the fixed effects β_i, they have a time-varying impact on the outcome y_{it} measured by $\boldsymbol{f}_t$. In macroeconomic applications, the factors $\boldsymbol{f}_t$ are unobserved common shocks (e.g. technology or weather shocks) that have an heterogeneous impact $\boldsymbol{\lambda}_i$ on unit i.

To simplify notation, let $\mathbb{N}_j = \{i\colon D_i = j\}$, $j = 0, 1$ denote the sets of untreated and treated units in the sample, respectively. Let $N_0 = \sum_{i=1}^N D_i$ and $n_0 = \frac{N_0}{N}$ denote the number and share of untreated units, respectively. The share of pre-treatment periods is $t_0 = \frac{T_0}{T}$.

I make the following assumptions.

Assumption 1 (No anticipation). *$y_{it}(1) = y_{it}(0)$ for all $t \leq T_0$ and all $i = 1, \ldots, N$.*

Assumption 2 (Correlated loadings). $\mathrm{E}[\boldsymbol{\lambda}_i | D_i = 1] - \mathrm{E}[\boldsymbol{\lambda}_i | D_i = 0] = \boldsymbol{\xi}_\lambda$ *with* $|\boldsymbol{\xi}_\lambda| < \infty$ *and* $\mathrm{Var}[\boldsymbol{\lambda}_i] = \boldsymbol{\Sigma}_{\lambda,i}$ *with* $\lim_{n\to\infty} \frac{1}{N_j} \sum_{i \in \mathbb{N}_j} \boldsymbol{\Sigma}_{\lambda,i} = \boldsymbol{\Sigma}_\lambda^{(j)}$ *for $j = 0, 1$, both positive definite $r \times r$ matrices.*

Assumption 3 (Convex hull condition). *The $T \times r$ factor matrix $\boldsymbol{F} = (\boldsymbol{F}_{\text{pre}}', \boldsymbol{F}_{\text{post}}')'$ satisfies*

1. $\operatorname{rank} \boldsymbol{F}_{\text{pre}} = k < T_0$,
2. *For all $t \in \{T_0 + 1, \ldots T\}$ $\exists \boldsymbol{v}_t \in \mathbb{V}\colon \boldsymbol{f}_t = \boldsymbol{F}_{\text{pre}}' \boldsymbol{v}_t$*

with $\mathbb{V} = \{\boldsymbol{v} \in \mathbb{R}^{T_0}\colon v_t \geq 0,\ \sum_{t=1}^{T_0} v_t = 1\}$ the set of non-negative weights that sum to one.

Assumption 4 (Selection on time-invariant unobservables). *For every i, $\mathrm{E}[\boldsymbol{\varepsilon}_i | \beta_i, D_i, \boldsymbol{\lambda}_i, \boldsymbol{F}] = \mathbf{0}$. Moreover, $\mathrm{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' | D_i] = \boldsymbol{\Sigma}_{\varepsilon,i}$ with $\lim_{n\to\infty} \frac{1}{N_j} \sum_{i \in \mathbb{N}_j} \boldsymbol{\Sigma}_{\varepsilon,i} = \boldsymbol{\Sigma}_\varepsilon^{(j)}$ some positive definite $T \times T$ matrices for $j = 0, 1$.*

Assumption 5 (Random sampling). *$(\boldsymbol{\varepsilon}_i, \boldsymbol{\lambda}_i)$ are independent over the cross section. $n_0, t_0 \in (0, 1)$ are both constant as $N \to \infty$ and $T_0 \geq 2$.*

Assumption 6 (Treatment effect heterogeneity). *The vector of individual treatment effects $\boldsymbol{\tau}_i = (\tau_{i,T_0+1}, \ldots, \tau_{i,T})'$, $\tau_{it} = y_{it}(1) - y_{it}(0)$, satisfies $\frac{1}{N} \sum_i D_i \boldsymbol{\tau}_i \xrightarrow{p} \boldsymbol{\tau}$ and either $\frac{1}{\sqrt{N}} \sum_i D_i (\boldsymbol{\tau}_i - \boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}[0, \frac{\boldsymbol{\Sigma}_\tau}{1-n_0}]$ or $\frac{1}{\sqrt{N}} \sum_i D_i (\boldsymbol{\tau}_i - \boldsymbol{\tau}) \xrightarrow{p} 0$, with $\boldsymbol{\Sigma}_\tau$ a $T_1 \times T_1$ positive semi-definite matrix.*

Assumption 1 ensures that we observe the untreated potential outcome of the treated units prior to the treatment. It would be violated in presence of anticipation effects, i.e. when the treatment affects the outcome before it actually starts. If sufficient pre-treatment observations are available, one can estimate the anticipation effects as it is commonly done in event-study designs.

Assumption 2 is the central characteristic of the model. It allows the loadings $\boldsymbol{\lambda}_i$ to differ systematically between treated and untreated units. The loading imbalance $\boldsymbol{\xi}_\lambda$ measures how much more (or less) the treated units are on average affected by the common factors $\boldsymbol{f}_t$. It also nests the two-way fixed effects model as a special case for $\xi_\lambda = 0$ and $\boldsymbol{\Sigma}_\lambda^{(1)} = \boldsymbol{\Sigma}_\lambda^{(0)} = 0$, since then $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ for all i. In that case the factor structure $\boldsymbol{\lambda}_i' \boldsymbol{f}_t$ reduces to a time fixed effect $\gamma_t = \boldsymbol{\lambda}' \boldsymbol{f}_t$.

The common factors $\boldsymbol{F}$ can be viewed as realizations from some deterministic or stochastic process, where the number of factors r is unknown and fixed. All results should be interpreted conditional on $\boldsymbol{F}$. Assumption 3.1 implies that all post-treatment factors can be written as a weighted average of pre-treatment factors with weights that sum to one but can be negative. To ensure this works with non-negative weights, Assumption 3.2, requires each post-treatment factor to be in the convex hull of the pre-treatment factors. It excludes factors containing monotonic deterministic trends. The synthetic control literature (Abadie et al., 2010; Ferman, 2021) requires similar balancing conditions on the loadings.

Under Assumption 4, the treatment assignment is strictly exogenous once conditioned on the loadings, fixed effects and the factors. Moreover, I allow for heteroskedasticity and arbitrary serial dependence of the idiosyncratic errors. Lastly, Assumption 5 imposes independence of the error component over the cross section. It requires the number of treated and untreated units to grow at the same rate.

Assumption 6 imposes high-level restrictions on the treatment effect heterogeneity required for consistency and asymptotic normality. To exclude cases in which the treatment effect heterogeneity dominates the limiting distribution, the sample average of the treated units' individual treatment effects must converge to the ATT at least at rate $\frac{1}{\sqrt{N}}$.

Remark 1. One can drop the non-negativity constraint on the weights. Instead of the restrictive convex hull condition, only sufficient rank of $\boldsymbol{F}_{\text{pre}}$ is required. Weights will have a closed form expression and inference remains unchanged. However, estimation becomes much more sensitive with respect to the number of pre-treatment periods, as the weights will lose their sparsity property.

Remark 2. Using potential outcomes notation, one may equivalently write Assumption 4 as $D_i \perp (\boldsymbol{y}_i(0), \boldsymbol{y}_i(1)) | (\beta_i, \boldsymbol{\lambda}_i)$. This is weaker than unconfoundedness as defined in Imbens and Wooldridge (2009), since here the treatment assignment may depend on unobserved characteristics as long as they are time-invariant.

The time-weighted DID estimator is computed in two steps. For notational convenience, I consider first the case in which treatment occurs only in the last period $t = T$.

1. Obtain a T_0-dimensional vector of time weights $\hat{\boldsymbol{v}} = (\hat{v}_1, \dots, \hat{v}_{T_0})'$ using the outcomes $y_{i,t}$ of the untreated units. Regress the post-treatment outcome $y_{i,T}$ on a constant and the pre-treatment outcomes $\boldsymbol{y}_{i,pre} = (y_{i,1}, \dots, y_{i,T_0})'$
$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v} \in \mathbb{V}, \alpha} \sum_{i:\, D_i = 0} (y_{i,T} - \alpha - \boldsymbol{y}'_{i,pre}\boldsymbol{v})^2 \tag{3}$$
with $\mathbb{V} = \{\boldsymbol{v} \in \mathbb{R}^{T_0} \colon v_t \geq 0,\ \sum_{t=1}^{T_0} v_t = 1\}$ the set of non-negative weights that sum to one.

2. Obtain the time-weighted DID estimator $\hat{\tau}(\hat{\boldsymbol{v}})$ as solution to the weighted two-way fixed effect regression
$$\min_{\tau, \boldsymbol{\mu}, \boldsymbol{\gamma}} \sum_{i=1}^{N} \sum_{t=1}^{T} v_t (y_{it} - \tau D_{it} - \mu_i - \gamma_t)^2 \tag{4}$$
with $v_T = 1$. The resulting estimator is
$$\hat{\tau}(\boldsymbol{v}) = \boldsymbol{v}'_a(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(0)}) = \Delta_T - \sum_{t=1}^{T_0} v_t \Delta_t \tag{5}$$
with augmented time weights $\boldsymbol{v}_a = (-\boldsymbol{v}', 1)'$, $\bar{\boldsymbol{y}}^{(j)} = \frac{1}{N_j}\sum_{i:\, D_i = j} \boldsymbol{y}_i$ the vectors of the treated $(j = 1)$ and untreated $(j = 0)$ units' average outcome in each period and $\Delta_t = \bar{y}_t^{(1)} - \bar{y}_t^{(0)}$.

In case of multiple treated periods, estimate the ATT for each post-treatment period separately with the two-step approach outlined above. Likewise, in case of staggered adoption one can apply the procedure for each treated group separately. Sections 2.4 and 2.5 discuss these cases in more detail.

The DID estimator is the special case of the TWDID estimator with equal weights $\bar{\boldsymbol{v}} = \frac{\iota_{T_0}}{T_0}$. The synthetic DID estimator of Arkhangelsky et al. (2021), in contrast, uses both unit weights and time weights and solves
$$\min_{\tau, \boldsymbol{\mu}, \boldsymbol{\gamma}} \sum_{i=1}^{N} \sum_{t=1}^{T} \omega_i v_t (y_{it} - \tau D_{it} - \mu_i - \gamma_t)^2$$
with $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{N_0})$ a vector of control unit weights and $\omega_i = 1$ for $i \in N_1$. The weights of the untreated units are non-negative sum to one and are estimated from the pre-treatment outcomes.

2.2 Bias and variance reduction through factor balancing

In the following, I will first consider the properties of $\hat{\tau}(\boldsymbol{v})$ under fixed weights to document the issue of factor imbalance. Subsequently, I will consider estimated time weights and their reduction of bias and asymptotic variance of estimated treatment effects compared to the standard DID approach. This part covers the case of one treated period and the cases of multiple treated periods and staggered adoption are discussed in Sections 2.4 and 2.5.

Consider the treatment effect estimate $\hat{\tau}(\boldsymbol{v})$ for a given vector of weights $\boldsymbol{v}$. For equal weights, $\boldsymbol{v} = \bar{\boldsymbol{v}}$, this is equivalent to the DID estimator obtained from a two-way fixed effects regression.

Theorem 1. *Suppose Assumptions 1,2 and 4-6 hold. Then for any $\boldsymbol{v} \in \mathbb{V}$,*

1. $\mathrm{E}[\hat{\tau}(\boldsymbol{v})|\boldsymbol{F}] = \tau + b(\boldsymbol{v})$ *with bias* $b(\boldsymbol{v}) = \boldsymbol{\xi}_\lambda' \boldsymbol{\xi}_f(\boldsymbol{v})$ *and the weighted factor imbalance* $\boldsymbol{\xi}_f(\boldsymbol{v}) = \boldsymbol{f}_T - \boldsymbol{F}_{\text{pre}}' \boldsymbol{v}$,
2. $\sqrt{N}(\hat{\tau}(\boldsymbol{v}) - \tau - b(\boldsymbol{v})) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{v})]$ *as* $N \to \infty$ *with limiting variance*

$$\boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{v}) = \boldsymbol{\xi}_f(\boldsymbol{v})' \boldsymbol{\Sigma}_\lambda \boldsymbol{\xi}_f(\boldsymbol{v}) + \boldsymbol{v}_a' \boldsymbol{\Sigma}_\varepsilon \boldsymbol{v}_a + \frac{\Sigma_\tau}{1 - n_0}$$

where $\boldsymbol{\Sigma}_\lambda = \frac{\boldsymbol{\Sigma}_\lambda^{(0)}}{n_0} + \frac{\boldsymbol{\Sigma}_\lambda^{(1)}}{1-n_0}$ *and* $\boldsymbol{\Sigma}_\varepsilon$ *defined accordingly.*

The proofs for this and the following theorems are in the Appendix. The weighted factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$ will play an important role. It is the difference between the post-treatment factor and the weighted average pre-treatment factors and affects the estimated treatment effect in two ways. First, the combination of a non-zero loading imbalance $\boldsymbol{\xi}_\lambda$ and factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$ leads to a first order bias term $b(\boldsymbol{v})$. For example, consider the case with one common factor f_t, which affects treated units on average more than untreated units ($\xi_\lambda > 0$). If f_t is in the post-treatment periods higher than in the pre-treatment periods, $\hat{\tau}(\boldsymbol{v})$ will overestimate the treatment effect. Second, it increases the part of the variance resulting from variation in the loadings. This holds irrespective of whether the treatment assignment D_i correlates with the loadings $\boldsymbol{\lambda}_i$, as long as they have within group variation $\boldsymbol{\Sigma}_\lambda > 0$.

The properties of the DID estimator follow as the special case of equal weights $\boldsymbol{v} = \bar{\boldsymbol{v}}$. Researchers typically refer to the common trend assumption as condition a for unbiasedness. In the current setting, a trend means a non-zero factor imbalance $\boldsymbol{\xi}_f(\bar{\boldsymbol{v}})$. The trends are common if the factors affect treated and untreated units equally. Hence the DID estimator is unbiased ($b(\bar{\boldsymbol{v}}) = 0$) if either the trends are common ($\boldsymbol{\xi}_\lambda = 0$) or there are no trends ($\boldsymbol{\xi}_f(\bar{\boldsymbol{v}}) = 0$).

Now consider the case where the weights are estimated from the control unit data as per (3). As the number of control units N_0 grows, they converge in probability to the pseudo-true time weights $\boldsymbol{v}^*$ which solve the population equivalent of (3)

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}\in\mathbb{V}}{\arg\min} \left\{ \boldsymbol{\xi}_f(\boldsymbol{v})' \boldsymbol{\Sigma}_\lambda^{(0)} \boldsymbol{\xi}_f(\boldsymbol{v}) + \boldsymbol{v}_a' \boldsymbol{\Sigma}_\varepsilon^{(0)} \boldsymbol{v}_a \right\} \tag{6}$$

The pseudo-true weights minimize an expression close to the variance of $\hat{\tau}(\boldsymbol{v})$ derived in Theorem 1, which is influenced by the factor imbalance $\boldsymbol{\xi}_f(\boldsymbol{v})$ and the error variance $\boldsymbol{\Sigma}_\varepsilon$. As a consequence, the pseudo-true weights do not in general balance the factors completely. The following Theorem establishes asymptotic normality around $\boldsymbol{v}^*$.

Theorem 2. *Suppose Assumptions 1-2 and 4-7 hold. Let* $0 \leq k \leq T_0 - 1$ *be the number of pseudo-true weights equal to zero and denote* $\boldsymbol{v}_k$ *the vector corresponding to the zero weights. As* $N \to \infty$*,*

1. $\hat{\boldsymbol{v}} \xrightarrow{p} \boldsymbol{v}^*$
2. $\Pr(\hat{\boldsymbol{v}}_k = 0) \xrightarrow[N\to\infty]{} 1$
3. $\sqrt{N}(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}_{\hat{v}}]$ *with* $\text{rk}\Sigma_{\hat{v}} = T_0 - 1 - k$, $(\boldsymbol{\Sigma}_{\hat{v}})_{s,t} = 0$ *if* $\min\{v_s^*, v_t^*\} = 0$.

Importantly, we identify the irrelevant pre-treatment periods (those with zero pseudo-true weights) with probability approaching one. For the asymptotic distribution of the weights, we can therefore consider the unrestricted weight estimation after omitting the irrelevant periods. Since this is a least-squares regression problem, asymptotic normality follows from standard arguments.

The estimated weights $\hat{\boldsymbol{v}}$ and the treatment effect estimate $\hat{\tau}(\boldsymbol{v})$ converge at the same rate $\frac{1}{\sqrt{N}}$. Under estimated weights, the treatment effect estimator becomes

$$\hat{\tau}(\hat{\boldsymbol{v}}) = \hat{\tau}(\boldsymbol{v}^*) - (\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda)'(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) + O_p(N^{-1})$$

The weight estimation uncertainty $\hat{\boldsymbol{v}} - \boldsymbol{v}^*$ therefore affects the limiting distribution unless $\boldsymbol{\xi}_\lambda = 0$, which is in line with the properties of two-step estimators (Newey and McFadden, 1994). This leads to the following result.

Theorem 3. *Suppose Assumptions 1,2 and 4-7 hold. Then, as* $N \to \infty$*,*

1. $\hat{\tau}(\hat{\boldsymbol{v}}) \xrightarrow{p} \tau + b(\boldsymbol{v}^*)$
2. $\sqrt{N}(\hat{\tau}(\hat{\boldsymbol{v}}) - \tau - b(\boldsymbol{v}^*)) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}]$

with $b(\boldsymbol{v}^*) = \boldsymbol{\xi}_\lambda' \boldsymbol{\xi}_f(\boldsymbol{v}^*)$ *and*

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{v}^*) + (\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda)'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda - 2\boldsymbol{\Sigma}_{\hat{\tau},\hat{v}}'\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda$$

and $\boldsymbol{\Sigma}_{\hat{\tau},\hat{v}}$ *the* $(T_0 \times 1)$ *covariance between* $\hat{\tau}$ *and* $\hat{\boldsymbol{v}}$*.*

The limit variance consists of three parts. First, $\Sigma_{\hat{\tau}}(\boldsymbol{v}^*)$ is the variance of the treatment effect estimator under pseudo-true weights $\boldsymbol{v}^*$. The second part comes from the variance of the weight estimation noise and the third part from the covariance of the weight estimation and treatment effect estimation.

The magnitude of the bias $b(\boldsymbol{v}^*)$ depends on the remaining factor imbalance under pseudo-true weights. In the simple case of $r = T_0 - 1$, it can be decomposed in two parts

$$\boldsymbol{\xi}_f(\boldsymbol{v}^*) = \boldsymbol{\xi}_f(\boldsymbol{v}_0) + \boldsymbol{F}_{\text{pre}}'\boldsymbol{R}(\boldsymbol{I} + \boldsymbol{A}_{\text{snr}})^{-1}\boldsymbol{R}^-(\boldsymbol{v}_0 - \boldsymbol{v}_\varepsilon) \tag{7}$$

with $\boldsymbol{A}_{\text{snr}}$ a matrix of signal-to-noise ratios in the pre-treatment periods (exact expression in the appendix), $\boldsymbol{v}_\varepsilon := \arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{v}_a'\boldsymbol{\Sigma}_\varepsilon^{(0)}\boldsymbol{v}_a$ the variance minimizing weights and $\boldsymbol{v}_0 := \arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{\xi}_f(\boldsymbol{v})'\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{\xi}_f(\boldsymbol{v})$ the oracle weights, which are unique in this case. The first part, $\boldsymbol{\xi}_f(\boldsymbol{v}_0)$, is the smallest factor imbalance that can be achieved with non-negative weights that sum to one. By construction, the convex hull condition (Assumption 3) implies that $\boldsymbol{\xi}_f(\boldsymbol{v}_0) = 0$. If the convex hull condition does not hold, the bias will be proportional to how far the post-treatment factor is from the convex hull of the pre-treatment factors. The second part comes from the fact that the pseudo true weights minimize the sum of the factor imbalance and error variance. However, the stronger the signal in the data, the more the weights will focus on eliminating the factor imbalance. Consequently, the pseudo-true weights will be closer to the oracle weights and $\boldsymbol{\xi}_f(\boldsymbol{v}^*) \approx \boldsymbol{\xi}_f(\boldsymbol{v}_0)$. Indeed, the Monte Carlo experiments in Section 3 show that the first term is of greater concern.

Comparing the DID estimator $\hat{\tau}(\bar{\boldsymbol{v}})$ to the time-weighted version $\hat{\tau}(\hat{\boldsymbol{v}})$ leads to the following conclusions. The bias of the latter is smaller if the pseudo-true weights decrease the factor imbalance compared to equal weights. This is arguably the case in most relevant scenarios, although, technically, counterexamples can be constructed. Next, weighting has a two-fold effect on the relative variance.

$$\frac{\Sigma}{\Sigma_{\hat{\tau}}(\bar{\boldsymbol{v}})} = \frac{\Sigma_{\hat{\tau}}(\boldsymbol{v}^*)}{\Sigma_{\hat{\tau}}(\bar{\boldsymbol{v}})} + \frac{\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda - 2\boldsymbol{\Sigma}_{\hat{\tau},\hat{v}}'\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda}{\Sigma_{\hat{\tau}}(\bar{\boldsymbol{v}})}$$

First, $\boldsymbol{v}^*$ minimizes the first term by construction. This comes at the cost of weight estimation noise, which is reflected in the second term. The Monte Carlo simulations in Section 3 show that TWDID substantially reduces bias and variance in a setting with one factor.

2.3 Inference with two-step standard errors

A consistent estimator of the limit variance derived in Theorem 3 is

$$\widehat{\Sigma} = \widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{v}}) + \dot{\boldsymbol{\Delta}}'_{\text{pre}}\widehat{\boldsymbol{\Sigma}}_{\hat{v}}\dot{\boldsymbol{\Delta}}_{\text{pre}} - 2\widehat{\boldsymbol{\Sigma}}'_{\hat{\tau},\hat{v}}\dot{\boldsymbol{\Delta}}_{\text{pre}} \tag{8}$$

The first part $\widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{v}})$ is the cluster-covariance robust variance estimator of Arellano (1987), often referred to as clustering the standard errors on the unit level (Bertrand, Duflo, and Mullainathan, 2004), applied to the weighted data. It consistently estimates the variance under pseudo-true weights $\Sigma_{\hat{\tau}}(\boldsymbol{v}^*)$. The second part accounts for the additional variance caused by the weight estimation noise. It consist of the demeaned average pre-treatment differences $\dot{\boldsymbol{\Delta}}_{\text{pre}} = (\Delta_1 - \bar{\Delta}_{\text{pre}}, \dots, \Delta_{T_0} - \bar{\Delta}_{\text{pre}})'$ with $\bar{\Delta}_{\text{pre}} = \frac{1}{T_0}\sum_{t \leq T_0} \Delta_t$ and a consistent estimator $\widehat{\boldsymbol{\Sigma}}_v$ of the weight variance. The third part accounts for correlations of $\hat{\tau}(\boldsymbol{v}^*)$ and $\hat{\boldsymbol{v}}$, which tends to be negligible in practice. In the remainder of this section, I will explain how to construct the different components of $\widehat{\Sigma}$ and associated confidence intervals.

The first part of the estimated variance is

$$\widehat{\Sigma}_{\hat{\tau}}(\boldsymbol{v}) = \boldsymbol{v}'_a \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Delta}} \boldsymbol{v}_a; \quad \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Delta}} = \frac{\widehat{\boldsymbol{\Sigma}}_y^{(0)}}{n_0} + \frac{\widehat{\boldsymbol{\Sigma}}_y^{(1)}}{1 - n_0} \tag{9}$$

with within-group sample variances $\widehat{\boldsymbol{\Sigma}}_y^{(j)} = \frac{1}{N_j}\sum_{i:D_i=j}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})'$ for $j = 0, 1$. It can be obtained in the following way. First estimate the time weights $\hat{\boldsymbol{v}}$. Next, weight only the pre-treatment outcomes and call them $\tilde{y}_{it} = T_0 \hat{v}_t y_{it}$ for $t \leq T_0$ and $\tilde{y}_{it} = y_{it}$ for $t > T_0$. Run a two-way fixed effects regression of the weighted outcomes $\tilde{y}_{it}$ on the treatment indicator D_{it}, which yields $\hat{\tau}(\hat{\boldsymbol{v}})$. Applying the Arellano (1987) cluster-covariance estimator on the weighted data then provides $\widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{v}})$.

Next, consider $\dot{\boldsymbol{\Delta}}'_{\text{pre}}\widehat{\boldsymbol{\Sigma}}_{\hat{v}}\dot{\boldsymbol{\Delta}}_{\text{pre}}$ provides as an estimator of $(\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda)'\boldsymbol{\Sigma}_{\hat{v}}\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda$. First, the demeaned average pre-treatment differences become

$$\dot{\boldsymbol{\Delta}}_{\text{pre}} = \boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda + O_p(N^{-1})$$

implying that they consistently estimate $\boldsymbol{F}_{\text{pre}}\boldsymbol{\xi}_\lambda$ as $N \to \infty$. For the estimator of the weight variance $\widehat{\boldsymbol{\Sigma}}_v$, let $\hat{\boldsymbol{v}}$ be the T_0-dimensional vector of estimated weights as per (3). As Theorem 3 establishes, only the non-zero weights matter for the limiting weight variance $\boldsymbol{\Sigma}_{\hat{v}}$. The non-zero weights can be seen as an unrestricted least-squares estimate and the least-squares type of standard errors can be used to estimate its variance.

Let $\hat{\boldsymbol{v}}_{[+]}$ be the T_+-dimensional vector which contains the strictly positive weights. Because the weights sum to one, I can write

$$\hat{\boldsymbol{v}}_{[+]} = \boldsymbol{e}_1 + \boldsymbol{R}\hat{\boldsymbol{v}}_{-1}$$

with $\boldsymbol{e}_1$ the T_+-dimensional unit vector, $\boldsymbol{R} = \begin{pmatrix} -\boldsymbol{\iota}'_{T_+-1} \\ \boldsymbol{I}_{T_+-1} \end{pmatrix}$ a $T_+ \times (T_+ - 1)$ matrix and $\hat{\boldsymbol{v}}_{-1}$ excludes the first element of $\hat{\boldsymbol{v}}_{[+]}$. The latter can be written as the unrestricted least-squares estimate

$$\hat{\boldsymbol{v}}_{-1} = (\tilde{\boldsymbol{Y}}'_{\text{pre}} \tilde{\boldsymbol{Y}}_{\text{pre}})^{-1} \tilde{\boldsymbol{Y}}'_{\text{pre}} \tilde{\boldsymbol{y}}_T$$

with $\tilde{\boldsymbol{Y}}_{\text{pre}} = \dot{\boldsymbol{Y}}^{(0)}_{[+]} \boldsymbol{R}$, $\tilde{\boldsymbol{y}}_T = \dot{\boldsymbol{y}}^{(0)}_T - \dot{\boldsymbol{y}}^{(0)}_1$ and $\dot{\boldsymbol{Y}}^{(0)}_{[+]}$ the $N_0 \times T_+$ matrix of demeaned outcomes of the control units in the remaining pre-treatment periods. A consistent estimator of the weight variance

$$\widehat{\boldsymbol{\Sigma}}_{\hat{v}_{-1}} = \bar{\boldsymbol{S}}_y^{-1} \bar{\boldsymbol{S}}_{qq'} \bar{\boldsymbol{S}}^{-1}; \quad \bar{\boldsymbol{S}}_{y,q} = \frac{1}{N} \sum_i \dot{\boldsymbol{q}}(\hat{\boldsymbol{v}}, \boldsymbol{y}_i) \dot{\boldsymbol{q}}(\hat{\boldsymbol{v}}, \boldsymbol{y}_i)' = \frac{1}{N} \boldsymbol{R}' \boldsymbol{Y}'_p \widehat{\boldsymbol{\Omega}}_q \boldsymbol{Y}_p \boldsymbol{R}$$

with $\widehat{\boldsymbol{\Omega}}_q = \text{diag}(q_1(\hat{\boldsymbol{v}}), \ldots, q_{N_0}(\hat{\boldsymbol{v}}))$ and $\dot{\boldsymbol{q}} = \frac{\partial q}{\partial \boldsymbol{v}'_{-1}}$ Finally, the estimator of the weight covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_{v,[+]} = \boldsymbol{R} \bar{\boldsymbol{S}}_y^{-1} \bar{\boldsymbol{S}}_{qq'} \bar{\boldsymbol{S}}^{-1} \boldsymbol{R}' \tag{10}$$

which follows from $\text{Var}[\hat{\boldsymbol{v}}_{[+]}] = \boldsymbol{R}\text{Var}[\hat{\boldsymbol{v}}_{-1}]\boldsymbol{R}'$.

The following Theorem summarizes the results.

Theorem 4. *Suppose Assumptions 1,2 and 4-6 hold. Then, as* $N \to \infty$, $\widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{v}}) \xrightarrow{p} \Sigma_{\hat{\tau}}(\boldsymbol{v}^*)$ *with* $\hat{\boldsymbol{v}}$ *the estimated weights as per* (3) *and* $\boldsymbol{v}^*$ *the pseudo-true weights as defined in* (6), $\dot{\boldsymbol{\Delta}}'_{\text{pre}} \widehat{\boldsymbol{\Sigma}}_{\hat{v}} \dot{\boldsymbol{\Delta}}_{\text{pre}} \xrightarrow{p} (\boldsymbol{F}_{\text{pre}} \boldsymbol{\xi}_\lambda)' \boldsymbol{\Sigma}_{\hat{v}} \boldsymbol{F}_{\text{pre}} \boldsymbol{\xi}_\lambda$ *and therefore* $\widehat{\Sigma} \xrightarrow{p} \Sigma$

Inference can be based on the t-statistic

$$T_{\tau_0} = \frac{\hat{\tau}(\hat{\boldsymbol{v}}) - \tau_0 - b(\boldsymbol{v}^*)}{\sqrt{\widehat{\Sigma}/N}}$$

as $\Pr[|T_{\tau_0}| > q_{1-\alpha/2}] \xrightarrow{p} \alpha$, with τ_0 the treatment effect under the null, α the intended size and $q_{1-\alpha/2}$ the $1 - \alpha/2$ quantile of the standard normal distribution. Without any further restrictions, the resulting confidence interval

$$\left[\hat{\tau}(\hat{\boldsymbol{v}}) \pm q_{1-\alpha/2} \sqrt{\widehat{\Sigma}/N}\right]$$

will be centered around $\tau_0 + b(\boldsymbol{v}^*)$. The Monte Carlo experiments in Section 3 it is more reliable and shorter compared to an unweighted DID approach.

2.4 Multiple treated periods

In the case of multiple treated periods $t = T_0 + 1, \ldots, T$, the object of interest becomes $\boldsymbol{\tau} = (\tau_{T_0+1}, \ldots, \tau_T)'$, the vector of ATTs in all $T_1 = T - T_0$ post-treatment periods. To estimate the dynamic effects $\boldsymbol{\tau}$, we can apply

TWDID for each treated period separately. First, estimate a vector of time weights

$$\hat{\boldsymbol{v}}^{(j)} = \underset{\boldsymbol{v} \in \mathbb{V}, \alpha}{\arg\min} \sum_{i:\, D_i = 0} (y_{i,j} - \alpha - \boldsymbol{y}'_{i,pre}\boldsymbol{v})^2 \tag{11}$$

for each post-treatment period $j = T_0{+}1, \dots, T$. Let $\boldsymbol{V} = [\boldsymbol{v}^{(T_0+1)}, \dots, \boldsymbol{v}^{(T)}]$ be the corresponding $T_0 \times T_1$ matrix of time weights. Then estimate the treatment effects

$$\hat{\boldsymbol{\tau}}(\boldsymbol{V}) = \boldsymbol{V}_a'(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(0)}) = \begin{pmatrix} \Delta_{T_0+1} - \sum_{t \le T_0} v_t^{(T_0+1)} \Delta_t \\ \vdots \\ \Delta_T - \sum_{t \le T_0} v_t^{(T)} \Delta_t \end{pmatrix} \tag{12}$$

with $\boldsymbol{V}_a = \begin{pmatrix} -\boldsymbol{V} \\ \boldsymbol{I}_{T_1} \end{pmatrix}$ the $T \times T_1$ matrix of augmented time weights. A consistent estimator of the $T_1 \times T_1$ covariance matrix of $\hat{\boldsymbol{\tau}}(\boldsymbol{V})$ is

$$\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\boldsymbol{V}) = \boldsymbol{V}_a'\widehat{\boldsymbol{\Sigma}}_\Delta \boldsymbol{V}_a \tag{13}$$

with $\widehat{\boldsymbol{\Sigma}}_\Delta = \frac{\widehat{\boldsymbol{\Sigma}}_y^{(0)}}{n_0} + \frac{\widehat{\boldsymbol{\Sigma}}_y^{(1)}}{1-n_0}$ and $\widehat{\boldsymbol{\Sigma}}_y^{(0)} = \frac{1}{N_j}\sum_{i:D_i=j}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})'$ the $T \times T$ sample covariance within the treated and untreated units.

The following theorem formally extends Theorem 3 to the dynamic case.

Theorem 5. *Suppose Assumptions 1-2 and 4-6 hold. Then, as $N \to \infty$,*

1. $\hat{\boldsymbol{\tau}}(\widehat{\boldsymbol{V}}) \xrightarrow{p} \boldsymbol{\tau} + \boldsymbol{b}(\boldsymbol{V}^*)$
2. $\sqrt{N}(\hat{\boldsymbol{\tau}}(\widehat{\boldsymbol{V}}) - \boldsymbol{\tau} - \boldsymbol{b}(\boldsymbol{V}^*)) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_.]$
3. $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\widehat{\boldsymbol{V}}) \xrightarrow{p} \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V}^*)$

The appendix contains the full expression of the limiting variance $\boldsymbol{\Sigma}_.$. In practice one may estimate the variance as

$$[\widehat{\boldsymbol{\Sigma}}]_{i,j} = [\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\widehat{\boldsymbol{V}})]_{i,j} + 1[i=j]\dot{\boldsymbol{\Delta}}'_{\text{pre}}\widehat{\boldsymbol{\Sigma}}_{v^{(T_0+j)}}\dot{\boldsymbol{\Delta}}_{\text{pre}}$$

for $i, j = 1, \dots, T_1$, where only the diagonal elements affected by the weight estimation noise. This is because $\hat{\boldsymbol{v}}^{(j)}$ does not affect $\hat{\tau}_k$ for $j \neq k$ We can then test for an effect in any period $H_0\colon \boldsymbol{\tau} = \boldsymbol{0}$ using the corresponding Wald statistic

$$(\hat{\boldsymbol{\tau}})'[\widehat{\boldsymbol{\Sigma}}]^{-1}\hat{\boldsymbol{\tau}} \xrightarrow{d} \chi^2(T_1)$$

as well as testing whether the effects vary over time ($H_0\colon \tau_j = \tau$ for all $j > T_0$).

In some cases, the object of interest may not be the full vector of ATTs $\boldsymbol{\tau}$, but only the average over the post treatment periods $\tau_{\text{avg}} := \frac{1}{T_1}\sum_{t=1}^{T_1}\tau_t$. One way to estimate τ_{avg} would be to average the estimated ATTs of all post-treatment periods. Alternatively, a pooled estimator would first collapse all post-treatment periods to one average post-treatment period and then estimate one vector of time weights using only the average post-treatment outcomes. The pooled estimator is also discussed by Arkhangelsky et al. (2021).

2.5 Staggered Adoption

The estimator can also be applied in settings with staggered adoption of the treatment, in which different groups of units start the treatment at different points in time. Suppose there is a finite number of groups indicated by $G_i \in \{0, 1, \ldots, G\}$, where units in group $g > 0$ start treatment in $t = T_0 + g$ and units in group $G_i = 0$ are never treated. An objects of interest could be the group-time average treatment effects

$$\tau_{g,t} = \mathrm{E}[y_{it}(g) - y_{it}(0)|G_i = g]$$

with $y_{it}(g)$ the potential outcome when starting treatment in $t = T_0 + g$, see also Callaway and Sant'Anna (2020) and Callaway and Karami (2022). Let $\boldsymbol{\tau}_g = (\tau_{g,T_0+g}, \ldots, \tau_{g,T})'$ the corresponding vector of estimated group-time ATTs for group g. It can be estimated with the TWDID approach for each treated group separately. For each group g, first estimate $T_0 + g - 1$ time weights $\hat{\boldsymbol{v}}_{(g)}^{(j)}$ for each period $j = T_0 + g, \ldots, T$ using the outcomes of the never-treated group. Then obtain $\hat{\boldsymbol{\tau}}_g = \hat{\boldsymbol{\tau}}(\boldsymbol{V}_g)$ and standard errors for the corresponding group as previously described. The results from the case with sharp treatment timing hold in the case of staggered adoption, provided that the number of observations per group $N_g = \sum_i 1[G_i = g]$ is large compared to the number of groups G.

Corollary 1. *Suppose* $\frac{N_g}{N} \to n_g \in (0, 1)$ *as* $N \to \infty$. *Then*

1. $\hat{\boldsymbol{\tau}}_g \xrightarrow{p} \boldsymbol{\tau}_g$

2. $\sqrt{N}(\hat{\boldsymbol{\tau}}_g - \boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}[0, \frac{\boldsymbol{\Sigma}_{\hat{\tau}}^{(g)}(\boldsymbol{V}_g^*)}{n_0+n_g}]$

3. $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}^{(g)}(\widehat{\boldsymbol{V}}_g) \xrightarrow{p} \boldsymbol{\Sigma}_{\hat{\tau}}^{(g)}(\boldsymbol{V}_g^*)$

for all $g = 1, \ldots, G$.

One can then aggregate the group-time ATTs to other statistics of interest, for example the overall ATT or the event-study ATT. These issues are further discussed by Callaway and Sant'Anna (2020) and Callaway and Karami (2022).

3 Monte Carlo Experiments

3.1 Design

In each Monte Carlo replication $r = 1, \dots, R$ I generate data from

$$y_{it}^{(r)} = \tau D_{it} + \sigma_f \lambda_i^{(r)} f_t^{(r)} + \varepsilon_{it}^{(r)}$$

with and $\varepsilon_{it}^{(r)} \sim \mathcal{N}[0,1]$ for $i = 1, \dots, N$ and $t = 1, \dots, T$, mutually independent. The loadings are $\lambda_i^{(r)} = \frac{\xi_\lambda}{\sqrt{N}} D_i + \nu_i^{(r)}$ with $\nu_i^{(r)} \sim \mathcal{N}[0,1]$ and loading imbalance $\xi_\lambda = 2$. The true treatment effect is $\tau = 0$. The number of units is $N = 100$ of which half are treated ($n_0 = \frac{N_0}{N} = 0.5$) The treatment occurs in the last of $T = 7$ periods. I consider two factor processes. For the first one I draw $f_t^{(r)} \sim \mathcal{N}[0,1]$ for $t = 1, \dots, T$, which implies that the convex hull condition is violated in approximately 30% of the draws. For the second case I first draw the pre-treatment factors $f_t^{(r)} \sim \mathcal{N}[0,1]$ for $t = 1, \dots, T_0$. Then I draw post-treatment factor from a truncated normal $f_T^{(r)} \sim \mathcal{TN}[0,1; f_{(1)}, f_{(T_0)}]$, enforcing the convex hull condition. I repeat the simulation exercise along a grid of factor standard deviations $\sigma_f \in \{0, 0.1, \dots, 2\}$. For each combination of parameters and sample size I conduct $R = 10,000$ replications.

In each replication, I compute the pseudo-true time weights $\boldsymbol{v}^*$ as of (6), the estimated time weights $\hat{\boldsymbol{v}}$ as of (3) and the SDID unit weights $\hat{\boldsymbol{\omega}}$. Let $\bar{\boldsymbol{\omega}}$ be the vector of equal weights. I compare the DID estimator $\hat{\tau}_{\text{did}} = \hat{\tau}(\bar{\boldsymbol{\omega}}, \bar{\boldsymbol{v}})$, the TWDID estimator $\hat{\tau}_{\text{twdid}} = \hat{\tau}(\bar{\boldsymbol{\omega}}, \hat{\boldsymbol{v}})$, the SDID estimator $\hat{\tau}_{\text{sdid}} = \hat{\tau}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{v}})$ and the demeaned synthetic control (DSC) estimator of Ferman and Pinto (2021) $\hat{\tau}_{\text{dsc}} = \hat{\tau}(\hat{\boldsymbol{\omega}}, \bar{\boldsymbol{v}})$. I also study the performance of the variance estimators described in Section 2.3. To do so, I compare the following Monte Carlo statistics.

1. The *conditional bias* is $b(\boldsymbol{\omega}, \boldsymbol{v}) = \xi_\lambda(\boldsymbol{\omega})\xi_f(\boldsymbol{v})$ with $\xi_\lambda(\boldsymbol{\omega}) = \bar{\lambda}^{(1)} - \sum_{i \in \mathbb{N}_0} \omega_i \lambda_i$ the loading imbalance after applying the control unit weights. For DID and DSC I use $\boldsymbol{v} = \bar{\boldsymbol{v}}$, and $\boldsymbol{v} = \boldsymbol{v}^*$ for TWDID and SDID. Likewise, $\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}$ for DID and TWDID. For DSC and SDID I use $b(\boldsymbol{v}) = \mathrm{E}_{MC}[\xi_\lambda(\hat{\boldsymbol{\omega}})]\xi_f(\boldsymbol{v})$, with $\mathrm{E}_{MC}[\xi_\lambda(\hat{\boldsymbol{\omega}})]$ the average weighted loading imbalance over all monte-carlo replications. Because I redraw the factors, I measure the magnitude of the conditional bias term by its (unconditional) standard deviation $\mathrm{sd}[b(\boldsymbol{\omega}, \boldsymbol{v})]$ with respect to the distribution of the factors.

2. Next, I look at the simulated *conditional standard deviation* $\mathrm{sd}[\hat{\tau}(\boldsymbol{\omega}, \boldsymbol{v})|\boldsymbol{F}]$ of the point estimates. I compute it as the Monte Carlo standard deviation of the bias corrected estimator $\hat{\tau}(\boldsymbol{\omega}, \boldsymbol{v}) - b(\boldsymbol{\omega}, \boldsymbol{v})$.

3. Finally, I consider the *expected standard error* $\mathrm{E}\left[\sqrt{\widehat{\Sigma}}\right]$ and the *coverage* of the 95% confidence intervals $\left[\hat{\tau} \pm q_{0.975}\sqrt{\widehat{\Sigma}/N}\right]$ obtained from the four estimation approaches. For all approaches I use the CCM estimator applied to the weighted data $\widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{v}})$; for TWDID and SDID I also account for the time weight estimation uncertainty as defined in (8).

3.2 Results

The top panel of Figure 1 plots the magnitude of the bias against the strength of the factors. The bias of DID increases as the factors become stronger. The bias of TWDID is about 50% lower compared to DID when the convex hull condition is frequently violated (left). When the convex hull condition is satisfied (right), the remaining bias of TWDID is bounded and vanishes as the factors become stronger. The bottom panel shows the conditional standard deviation. For sufficiently strong factors, the TWDID estimator has a substantially lower standard deviation than TWDID. For weak factors, however, there is not much to be gained and the weight estimation noise leads to a slightly larger standard deviation.

The DSC and the SDID estimator, both using unit weights, are even more successful and almost eliminate the bias. Their standard deviation is slightly than TWDID, except when the convex hull condition is violated and factors are strong.

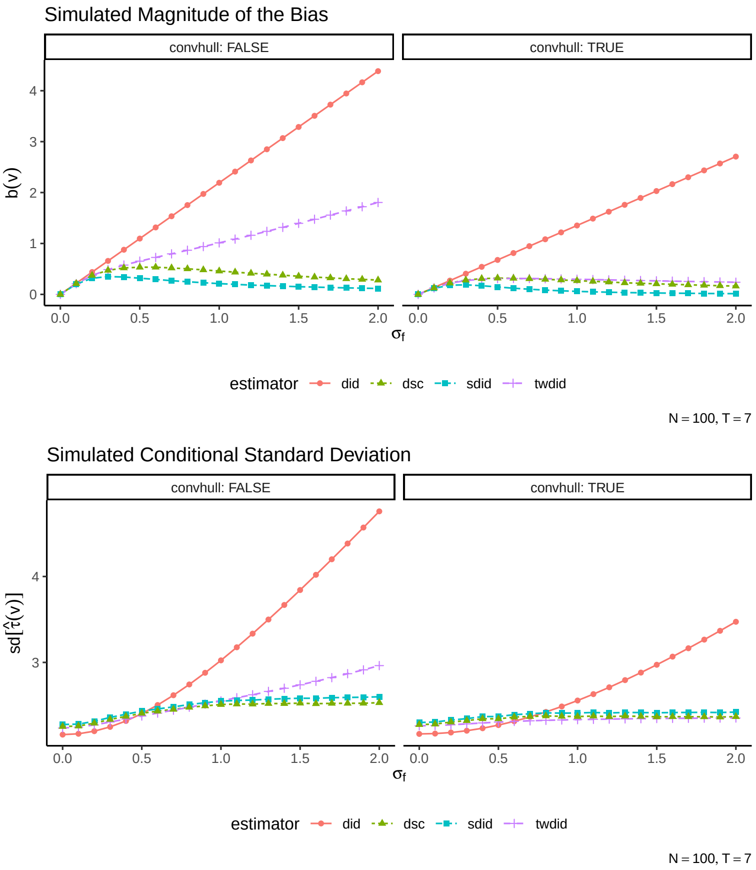
Consider now the properties of inference. The top panel of Figure 2 shows the true coverage of the 95% confidence interval. The remaining bias of the TWDID estimator is negligible, as the coverage remains around the desired 95%. This holds even when the convex hull condition is frequently violated (left). The larger bias of the DID estimator leads its coverage to deteriorate.

Inference based on the SDID and DSC estimators is conservative with coverages above 95%, because the standard errors are too large compared to the true standard deviation. As the bottom panel of Figure 2 shows, SDID and DSC standard errors are larger than those of TWDID, leading to wider confidence intervals.

4 TWDID in practice: revisiting two applications

4.1 The effect of the NOx Budget Trading Program

I revisit Deschenes et al. (2017) studying the effect of the NOx Budget Trading Program (NBP) 2003-2008 on NOx emissions. It entailed a cap and trade program to reduce NOx emissions from power plants. It was only active in the summer months May - September in the years 2003-2008 in 19



Simulated Magnitude of the Bias

Simulated Conditional Standard Deviation

Figure 1: Magnitude of the conditional bias term $b(\boldsymbol{\omega}, \boldsymbol{v})$ (top) and simulated conditional standard deviation $\text{sd}[\hat{\tau}(\boldsymbol{\omega}, \boldsymbol{v})]$ (bottom) of four estimators: difference-in-differences (DID), time-weighted DID (TWDID), synthetic DID (SDID) estimator and demeaned synthetic control (DSC). The horizontal axis depicts different levels of the factor standard deviation σ_f. The factors are drawn such that the convex hull condition is violated in the left panels and holds in the right panels.

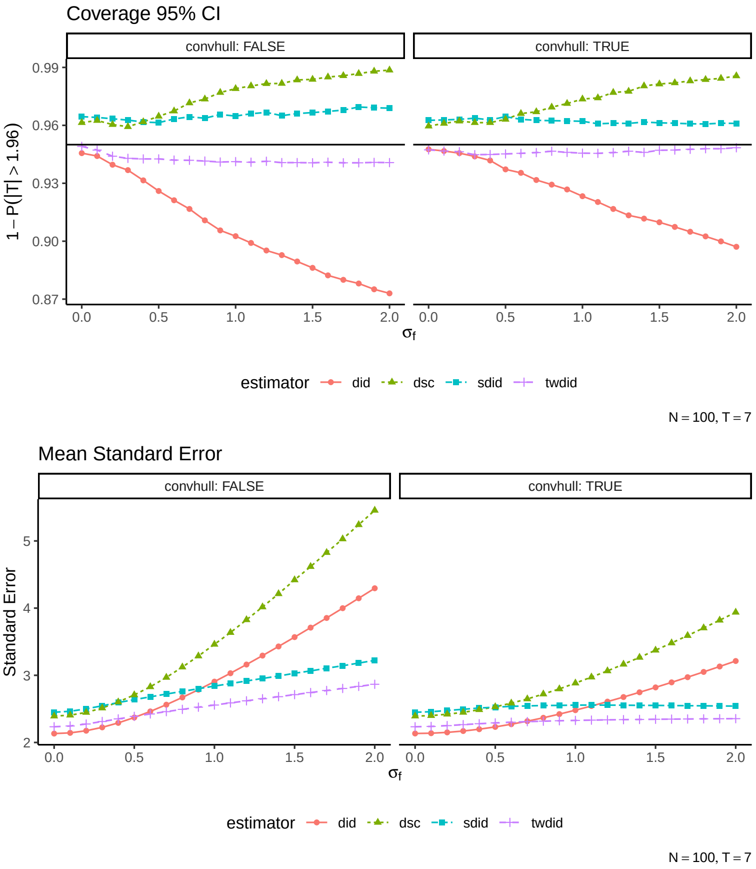


Figure 2: Simulated coverage of the 95% confidence intervals (top) and expected standard errors E$\left[\sqrt{\widehat{\Sigma}}\right]$ depending on the factor strength σ_f. The setup is the same as in Figure 1. For all approaches I use the CCM estimator applied to the weighted data $\widehat{\Sigma}_{\hat{\tau}}(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{v}})$; for TWDID and SDID I also account for the time weight estimation uncertainty as defined in (8)

states in the US. In 2003 the program was active only in a subset of the 19 treated states. States not adjacent to the NBP states remain as non-treated states (22 in total).

Data on NOx emissions is available on county level for $N = 2539$ counties from 1997-2007. We observe $N_1 = 1,354$ counties in the treated states and $N_0 = 1,185$ in the untreated states. Per county and year we observe data for the seasons summer and winter, where summer is defined as May - September.

4.1.1 Econometric Specification

Consider the interactive fixed effect model

$$y_{ist} = \sum_{j=2004}^{2008} \tau_j^{\text{att}} D_{ist}(j) + \mu_{it} + \nu_{is} + \boldsymbol{\lambda}_i' \boldsymbol{f}_{st} + \tilde{\varepsilon}_{ist}$$

with $D_{ist}(j) = \mathrm{I}(i \in \mathcal{N}_1, t = j, s = 1)$ an post-treatment dummy of year j indicating whether NBP is operating in county i in season $s = 0, 1$ (winter, summer). μ_{it}, ν_{is} are county-year and county-season fixed effects, respectively. $\boldsymbol{f}_{st}$ are season-year specific common shocks that affect the emissions of county i with intensity $\boldsymbol{\lambda}_i$. $\tilde{\varepsilon}_{ist}$ is an idiosyncratic error term. The special case $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ resembles the additive fixed effect model that Deschenes et al. (2017) use. In that case the factor structure reduces to a season-year fixed effect.

To identify τ, eliminate μ_{it} by taking the difference between summer and winter observations

$$\check{y}_{it} := y_{i1t} - y_{i0t} = \sum_{j=2004}^{2008} \tau_j^{\text{att}} D_{it}(j) + \beta_i + \boldsymbol{\lambda}_i' \check{\boldsymbol{f}}_t + \varepsilon_{it}$$

with $\beta_i = \nu_{i1} - \nu_{i0}$, $\check{\boldsymbol{f}}_t = \boldsymbol{f}_{1t} - \boldsymbol{f}_{0t}$ and $\varepsilon_{it} = \tilde{\varepsilon}_{i1t} - \tilde{\varepsilon}_{i0t}$. A key assumption hidden in this specification is that the program does not affect emissions in the winter months in the treated years. The identifying assumption is that for all post-treatment periods $j = 2004, \ldots, 2008$ there exists a vector of weights $\boldsymbol{v}_0^{(j)}$ such that

$$\check{\boldsymbol{f}}_j = \sum_{t \le T_0} v_{0,t}^{(j)} \check{\boldsymbol{f}}_t$$

That is, each post-treatment factor can be written as a weighted average of pre-treatment factors.

4.1.2 Evidence for common factors

I first obtain evidence against $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ by considering how the difference in average NOx emissions $\Delta_t = \bar{\check{y}}_t^{(1)} - \bar{\check{y}}_t^{(0)}$ has evolved prior to the intervention. We can write

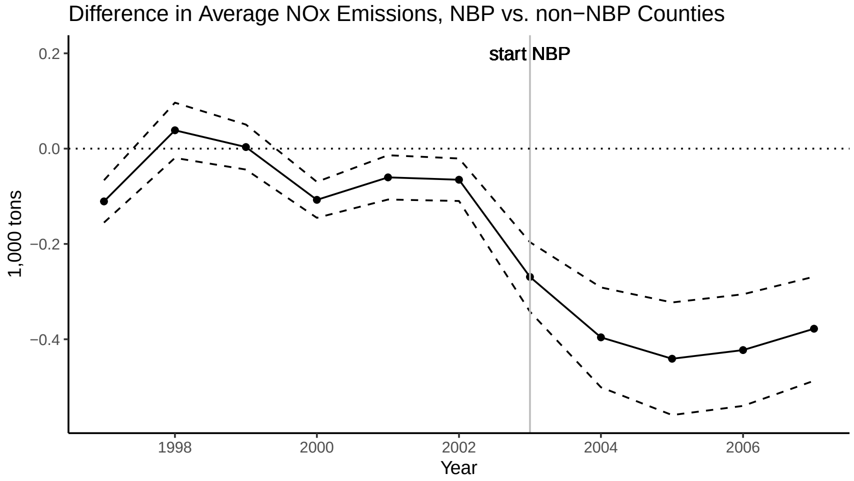


Figure 3: Difference in average NOx emissions $\bar{y}_t^{(1)} - \bar{y}_t^{(0)}$ over time, with 95% confidence band (dashed).

$$\Delta_t = \bar{\beta}^{(1)} - \bar{\beta}^{(0)} + \boldsymbol{\xi}_\lambda' \boldsymbol{f}_t + \sum_{j=2004}^{2008} \tau_j^{\text{att}} \mathrm{I}(t = j) + O_p(\frac{1}{\sqrt{N}})$$

where $\boldsymbol{\xi}_\lambda = 0$ under the equal loading assumption. Then, for large N, Δ_t should be constant prior to the treatment. However, Figure 3 does show variation of Δ_t in periods $t \leq T_0$.

4.1.3 Estimation Results

I estimate the dynamic effects $\boldsymbol{\tau}^{\text{att}}$ with a dynamic TWDID approach presented in Section 2.4. First, I obtain four sets of time weights $\hat{\boldsymbol{v}}^{(j)}$, one for each post-treatment period $j = 2004, \ldots, 2008$, as defined in (11). Then I compute the treatment effect estimate $\hat{\boldsymbol{\tau}}^{\text{att}}$ and the corresponding covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}$. For comparison I also computed the DID estimator, which uses equal time weights. I omit the year 2003 from the estimation because not all treated states had fully implemented the program by then.

The left panel of Figure 4 shows the estimated time weights for each post treatment periods. The last two pre-treatment years (2001 and 2002) receive most of the weight, while the exact distribution of the weights changes over the post treatment periods. The right panel of Figure 4 shows the resulting dynamic treatment effect estimates and their 95% confidence intervals. All estimators suggest a significant negative effect of the NBP program on NOx emissions in all post-treatment years. Time weighting leads, in absolute terms, to slightly lower point estimates. The standard errors of both TWDID estimates are about 10% lower compared to DID, hence the resulting confidence intervals are narrower. This results is in line with the variance reduction property of TWDID estimation. Overall, these findings

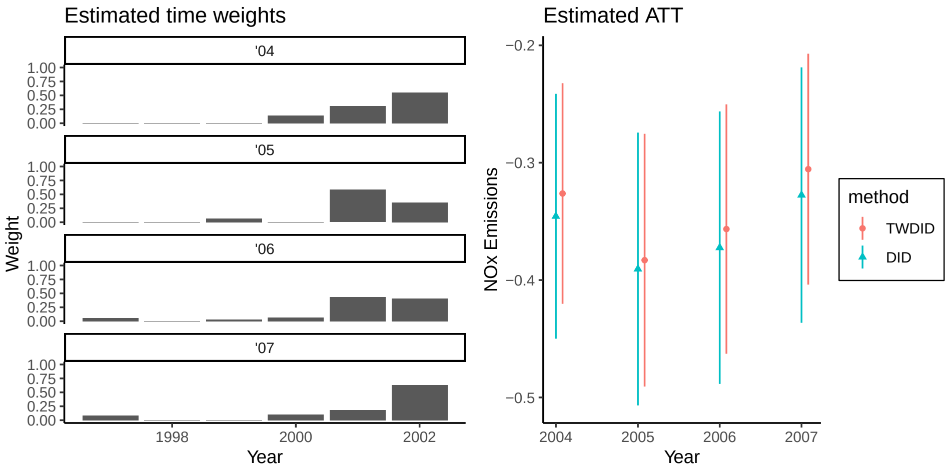


Figure 4: Left: Estimated time weights for each post-treatment period as of (11). Right: Resulting estimates of $\boldsymbol{\tau}^{\text{att}}$ and confidence intervals for both DID and TWDID estimation.

strengthen the results of Deschenes et al. (2017) that the NBP led to a significant reduction of NOx emissions.

4.2 The effect of job displacement on earnings

Consider the application of Callaway and Karami (2022), which estimates effect of early-career job displacement on earnings using data from the 1979 National Longitudinal Study of Youth (NLSY). The dataset contains biannual observations on yearly earnings for 2,850 US residents between 14 and 22 in years $1983, \dots, 1993$. A worker is defined as displaced if they are not longer in the same job as in the previous survey and is considered treated in all years following the first displacement. There are four groups $G = 0, \dots, 3$: those who are never displaced ($N_0 = 2,434$) and those who are first displaced in years 1989 ($N_1 = 129$), 1991 ($N_2 = 154$) and 1993 ($N_3 = 133$). Figure 5 plots the average earnings per group over time.

Interactive fixed effects are likely to be present here, since earnings might depend on unobserved characteristics (e.g. ability) with time-varying effects. To account for this, Callaway and Karami (2022) use Armed Forces Qualification Test (AFQT) test scores as an observed time-invariant covariate to implement their estimator. This serves as an important benchmark for the TWDID estimator, which does not require any additional covariates.

Let $T_{0,g}$ denote the last pre-treatment period for each group. I compare DID and TWDID estimates of the group-time average treatment effects

$$\widehat{ATT}(g,j) = \bar{y}_j^{(g)} - \bar{y}_j^{(0)} - \sum_{t \leq T_{0,g}} v_{t,g}^{(j)} (\bar{y}_t^{(g)} - \bar{y}_t^{(0)})$$

for groups $g = 1, 2, 3$, in the treated periods $j = 1989, 1991, 1993$. Both

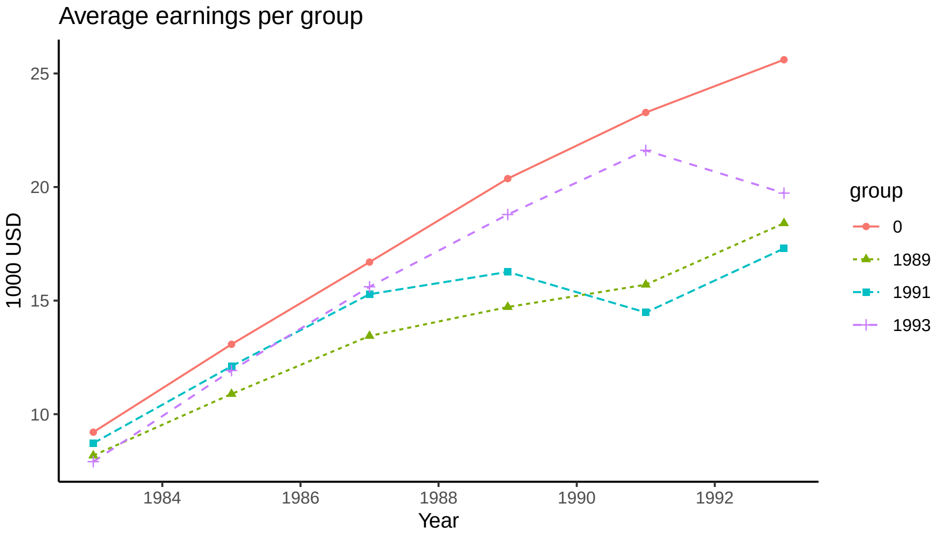


Figure 5: Average earnings by group using NLSY data Callaway and Karami (2022). The group labels correspond to the year in which the worker was first displaced or 0 for those who were never displaced.

approaches use the never-displaced group $g = 0$ as control group and all available pre-treatment observations to estimate the counterfactual. DID estimation uses equal weights $v_{t,g}^{(j)} = \frac{1}{T_{0,g}}$. In TWDID estimation, for each group g I estimate a separate $T_{0,g}$-dimensional vector of time weights for each period in which this group is treated. Table 1 shows the estimated time weights. In all cases, the last two periods before the start of the treatment receive almost all the weight and the last period before the treatment always receives the largest weight.

Table 2 shows the estimates of the group-time ATTs and the corresponding standard errors for both approaches. Compared to DID, in absolute terms, TWDID suggests smaller but still significant effects across all groups and post-treatment periods. However, the extent to which the estimates differ depends on the group. For the last group, this difference is around 10% of the standard error, for the first group around one standard error and for the second group almost two standard errors. This can to some extent be explained by looking at how the average earnings of each group evolve over time. Judging from Figure 5, the difference in average earnings between the never-treated and group 1993 is fairly constant over the pre-treatment periods. Therefore, the estimated ATT is not very sensitive with respect to the pre-treatment weights. The mean differences of groups 1989 and 1991 vis-a-vis the never-treated group are increasing over the pre-treatment periods. Hence, the weighted average of pre-treatment differences, with higher weight on the last two pre-treatment periods, is larger than the simple average. Consequently, TWDID attributes a larger part of the post-treatment

Table 1: Estimated time weights per group and post-treatment period

	1989	1991	1993
	Group 1989		
1983	0.01	0.00	0.00
1985	0.15	0.14	0.08
1987	0.84	0.85	0.92
	Group 1991		
1983		0.01	0.00
1985		0.01	0.00
1987		0.17	0.20
1989		0.81	0.80
	Group 1993		
1983			0.00
1985			0.00
1987			0.08
1989			0.31
1991			0.61

Notes: Estimated weights that were used in the TWDID estimation of $ATT(g,t)$, with $t \in \{1989, 1991, 1993\}$ varying across columns.

Table 2: Estimated Group-Time Average Treatment Effects $\widehat{ATT}_j(g,t)$

	1989	1991	1993
	Group 1989		
DID	-3.50	-5.43	-5.04
	(0.82)	(1.12)	(1.23)
TWDID	-2.59	-4.49	-4.03
	(0.80)	(1.09)	(1.22)
	Group 1991		
DID		-7.06	-6.57
		(0.82)	(0.97)
TWDID		-5.23	-4.75
		(0.83)	(0.99)
	Group 1993		
DID			-4.52
			(1.46)
TWDID			-4.28
			(1.46)

Notes: The outcome is yearly earnings in 1,000USD. Each column corresponds to the year for which the ATT is estimated. The group corresponds to the year in which the worker was first displaced. The time weights are estimated separately for each of the 6 (g,t) combinations. Analytical standard errors in parentheses.

difference in earnings to differences that would have persisted without any displacement and therefore suggest a smaller effect of the displacement.

The estimates of Callaway and Karami (2022) suggest an even smaller effect for groups 1989 and 1991. Assuming their results are unbiased, this would indicate that TWDID reduces the bias compared to DID, yet a non-negligible bias remains. Another indication are the increasing pre-treatment differences, which could be caused by unobserved common factors with a strong trend. In that case, the post-treatment factors are outside the convex hull of the pre-treatment factors, which would explain the remaining bias of TWDID.

5 Conclusion

This paper proposes a time-weighted difference-in-differences (TWDID) estimation approach that is robust against interactive fixed effects in short T panels. Time weighting substantially reduces both bias and variance compared to conventional DID estimation through balancing the pre-treatment and post-treatment unobserved common factors. To conduct valid inference on the average treatment effect, I develop a correction term that adjusts conventional standard errors for weight estimation uncertainty. Revisiting a study on the effect of a cap-and-trade program on NOx emissions, TWDID estimation reduces the standard errors of the estimated treatment effect by 10% compared to a conventional DID approach. In a second application I illustrate how to implement TWDID with staggered adoption.

References

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

Abadie, A. and J. L'hour (2020): "A penalized synthetic control estimator for disaggregated data," *Working Paper.*

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

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

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

Bai, J. (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Ben-Michael, E., A. Feller, and J. Rothstein (2021): "The Augmented Synthetic Control Method," *Journal of the American Statistical Association*, 116, 1789–1803.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Borusyak, K., X. Jaravel, and J. Spiess (2022): "Revisiting Event Study Designs: Robust and Efficient Estimation," ArXiv:2108.12419 [econ].

Brown, N. and K. Butts (2022): "A Unified Framework for Dynamic Treatment Effect Estimation in Interactive Fixed Effect Models," Tech. rep., Working Paper.

Callaway, B. and S. Karami (2022): "Treatment effects in interactive fixed effects models with a small number of time periods," *Journal of Econometrics.*

Callaway, B. and P. H. C. Sant'Anna (2020): "Difference-in-Differences with multiple time periods," *Journal of Econometrics.*

Chan, M. K. and S. S. Kwok (2021): "The PCDID Approach: Difference-in-Differences When Trends Are Potentially Unparallel and Stochastic," *Journal of Business & Economic Statistics*, 0, 1–18.

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

Deschenes, O., M. Greenstone, and J. S. Shapiro (2017): "Defensive investments and the demand for air quality: Evidence from the NOx budget program," *American Economic Review*, 107, 2958–89.

Ferman, B. (2021): "On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls," *Journal of the American Statistical Association*, 116, 1764–1772.

Ferman, B. and C. Pinto (2021): "Synthetic controls with imperfect pretreatment fit," *Quantitative Economics*, 12, 1197–1221.

Gobillon, L. and T. Magnac (2016): "Regional policy evaluation: Interactive fixed effects and synthetic controls," *Review of Economics and Statistics*, 98, 535–551.

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

Imai, K. and I. S. Kim (2021): "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data," *Political Analysis*, 29, 405–415.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Ketz, P. (2018): "Subvector inference when the true parameter vector may be near or at the boundary," *Journal of Econometrics*, 207, 285–306.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74, 967–1012.

Roth, J. (2022): "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends," *American Economic Review: Insights*, 4, 305–322.

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

Xu, Y. (2017): “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models,” *Political Analysis*, 25, 57–76.

A Proofs of Theorems

Assumption 7 (Regularity Conditions). *There are constants $K < \infty$ and $\delta > 0$ such that* $\mathrm{E}[|\varepsilon_{it}|^{4+\delta}] < K$, $\mathrm{E}[|\lambda_i|^{4+\delta}] < K$, $\mathrm{E}[|\tau_{it} - \tau_t|^{2+\delta}] < K$ *for all* i, t.

A.1 Proof of Theorem 1

The proof is provided for the case of multiple treated periods. The versions in the main text follow as a special case of one treated period. Using (2), write

$$\hat{\boldsymbol{\tau}}(\boldsymbol{V}) - \boldsymbol{\tau} = \boldsymbol{V}_a'\boldsymbol{F}(\xi_\lambda + \bar{\boldsymbol{\nu}}_\lambda) + \bar{\boldsymbol{\nu}}_\tau^{(1)} + \boldsymbol{z}_\varepsilon(\boldsymbol{V})$$

with $\bar{\boldsymbol{\nu}}_\lambda = \bar{\boldsymbol{\lambda}}^{(1)} - \bar{\boldsymbol{\lambda}}^{(0)} - \boldsymbol{\xi}_\lambda$, $\bar{\boldsymbol{\nu}}_\tau^{(1)} = \frac{1}{N_1}\sum_{D_i=1}(\boldsymbol{\tau}_i - \boldsymbol{\tau})$ and $\boldsymbol{z}_\varepsilon(\boldsymbol{V}) = \boldsymbol{V}_a'(\bar{\boldsymbol{\varepsilon}}^{(1)} - \bar{\boldsymbol{\varepsilon}}^{(0)})$.

We have $\mathrm{E}\,\bar{\boldsymbol{\nu}}_\lambda = 0$ (Assumption 2), $\mathrm{E}\,\bar{\nu}_\tau^{(1)} = 0$ (Assumption 6) and $\mathrm{E}\,\boldsymbol{z}_\varepsilon(\boldsymbol{V}) = \boldsymbol{0}$ (Assumption 4). Together with standard central limit theorems this implies the following asymptotic results.

Lemma 1. *Suppose Assumptions 1,2 and 4-7 hold. Then, for any fixed* $\boldsymbol{v} \in \mathbb{V}$

1. $\sqrt{N}\boldsymbol{z}_\varepsilon(\boldsymbol{V}) \xrightarrow{d} \mathcal{N}[\boldsymbol{0}, \boldsymbol{V}_a'\boldsymbol{\Sigma}_\varepsilon\boldsymbol{V}_a]$
2. $\sqrt{N}\bar{\boldsymbol{\nu}}_\lambda \xrightarrow{d} \mathcal{N}[\boldsymbol{0}, \boldsymbol{\Sigma}_\lambda]$
3. $\sqrt{N}\bar{\nu}_\tau^{(1)} \xrightarrow{d} \mathcal{N}[\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_\tau}{1-n_0}]$

with $\boldsymbol{\Sigma}_\varepsilon = \frac{\boldsymbol{\Sigma}_\varepsilon^{(0)}}{n_0} + \frac{\boldsymbol{\Sigma}_\varepsilon^{(1)}}{1-n_0}$ *and* $\boldsymbol{\Sigma}_\lambda = \frac{\boldsymbol{\Sigma}_\lambda^{(0)}}{n_0} + \frac{\boldsymbol{\Sigma}_\lambda^{(1)}}{1-n_0}$.

Proof. Liapounov Central Limit Theorem implies $\sqrt{N}\bar{\boldsymbol{\varepsilon}}^{(1)} \xrightarrow{d} \mathcal{N}[0, \frac{\boldsymbol{\Sigma}_\varepsilon^{(1)}}{1-n_0}]$ and $\sqrt{N}\bar{\boldsymbol{\varepsilon}}^{(0)} \xrightarrow{d} \mathcal{N}[0, \frac{\boldsymbol{\Sigma}_\varepsilon^{(0)}}{n_0}]$. □

The last assertion of Theorem 1 follows with independence of loadings and errors

$$\boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V}) = \boldsymbol{\xi}_f(\boldsymbol{V})'\boldsymbol{\Sigma}_\lambda\boldsymbol{\xi}_f(\boldsymbol{V}) + (1-n_0)\boldsymbol{\Sigma}_\tau + \boldsymbol{V}_a'\boldsymbol{\Sigma}_\varepsilon\boldsymbol{V}_a$$

and $\boldsymbol{\xi}_f(\boldsymbol{V}) = \boldsymbol{F}'\boldsymbol{V}_a$ the $r \times T_1$ matrix of factor imbalances.

A.2 Proof of Theorem 2

A subscript indicating the untreated units has been skipped for notational convenience. The estimated time weights $\hat{\boldsymbol{v}}$ as defined in (3) can be rewritten as

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}\in\mathbb{V}} Q_N(\boldsymbol{v}), \qquad Q_N(\boldsymbol{v}) = (\boldsymbol{y}_T - \boldsymbol{Y}_{\text{pre}}\boldsymbol{v})'\boldsymbol{M}_0(\boldsymbol{y}_T - \boldsymbol{Y}_{\text{pre}}\boldsymbol{v})$$

with $\boldsymbol{y}_t$ the vector of the untreated units' outcomes in the period t, $\boldsymbol{Y}_{\text{pre}}$ the $N_0 \times T_0$ matrix of the untreated units' pre-treatment outcomes and $\boldsymbol{M}_0 = \boldsymbol{I}_{N_0} - \frac{\boldsymbol{\iota\iota}'}{N_0}$. We notice that $\hat{\boldsymbol{v}}$ always exists, and is unique if the matrix of demeaned pre-treatment outcomes $\boldsymbol{M}_0\boldsymbol{Y}_{\text{pre}}$ has full column rank.

Consistency follows from the standard conditions for the extremum estimators. First, applying a uniform law of large numbers, the objective function Q_N uniformly convergence to its population equivalent $Q_\infty(\boldsymbol{v}) = \boldsymbol{v}_a'(\boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon^{(0)})\boldsymbol{v}_a$, so

$$\sup_{\boldsymbol{v}\in\mathbb{V}} |Q_N(\boldsymbol{v}) - Q_\infty(\boldsymbol{v})| \xrightarrow{p} 0$$

Second, $\boldsymbol{v}^*$ as defined in (6) is the uniquely identifiable minimizer of $Q_\infty(\boldsymbol{v})$, since the population objective

$$Q_\infty(\boldsymbol{v}) = \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} + 2\boldsymbol{v}'\boldsymbol{A}\boldsymbol{b} + c$$

is a strictly convex second order polynomial with $\boldsymbol{A}$ positive definite as long as $\boldsymbol{\Sigma}_\varepsilon^{(0)}$ is positive definite.

To obtain asymptotic normality, suppose first that $\boldsymbol{v}^*$ lies in the interior of $\mathbb{V}$, i.e. the non-negativity constraints are not binding in the limit. Without the non-negativity constraint we can rewrite the estimation as an unconstrained minimization and obtain an explicit solution. To do so, we transform the condition $\sum_t v_t = 1$ into $v_1 = 1 - \boldsymbol{\iota}_{T_0-1}'\boldsymbol{v}_{-1}$ with $\boldsymbol{v}_{-1} = (v_2, \dots, v_{T_0})$, so $\boldsymbol{v} = \boldsymbol{e}_1 + \boldsymbol{R}\boldsymbol{v}_{-1}$ with $\boldsymbol{R} = \begin{pmatrix} -\boldsymbol{\iota}_{T_0-1}' \\ \boldsymbol{I}_{T_0-1} \end{pmatrix}$. The minimization problem then becomes

$$\min_{\boldsymbol{v}_{-1}\in\mathbb{R}^{T_0-1}} (\tilde{\boldsymbol{y}}_T - \tilde{\boldsymbol{Y}}_{\text{pre}}\boldsymbol{v}_{-1})'(\tilde{\boldsymbol{y}}_T - \tilde{\boldsymbol{Y}}_{\text{pre}}\boldsymbol{v}_{-1})$$

with $\tilde{\boldsymbol{y}}_T = \bar{\boldsymbol{y}}_T - \boldsymbol{y}_1$ and $\tilde{\boldsymbol{Y}}_{\text{pre}} = \boldsymbol{Y}_{\text{pre}}\boldsymbol{R}$. The solution is

$$\hat{\boldsymbol{v}}_{-1} = \bar{\boldsymbol{S}}_y^{-1}\tilde{\boldsymbol{Y}}_{\text{pre}}'\tilde{\boldsymbol{y}}_T$$

with $\bar{\boldsymbol{S}}_y = \boldsymbol{R}'\boldsymbol{Y}_{\text{pre}}'\boldsymbol{Y}_{\text{pre}}\boldsymbol{R}$. The corresponding T_0 vector of time weights will be $\hat{\boldsymbol{v}} = (1 - \sum_{t=2}^{T_0}\hat{v}_t, \hat{\boldsymbol{v}}_{-1}')'$. That is a least-squares regression where the outcomes and regressors are in terms of their difference to the first regressor. The next lemma provides the ingredients for asymptotic normality of $\hat{\boldsymbol{v}}_{-1}$.

Lemma 2. *Let* $\hat{\boldsymbol{v}}_{-1} = \arg\min_v \frac{1}{N}\sum_i q_i(\boldsymbol{v})$ *with* $q_i(\boldsymbol{v}) = q(\boldsymbol{v}, \boldsymbol{y}_i) = \frac{1}{2}(1 - D_i)(\tilde{y}_{i,T} - \tilde{\boldsymbol{y}}_{i,p}\boldsymbol{v})^2$ *and derivatives* $\dot{q} = \frac{\partial q}{\partial v'}$, $\ddot{q} = \frac{\partial q}{\partial^2 v'}$. *Then, as* $N \to \infty$,

1. $\frac{1}{\sqrt{N}}\sum_i \dot{q}(\boldsymbol{v}^*, \boldsymbol{y}_i) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{S}_{qq'}]$ *with* $\boldsymbol{S}_{qq'} = \lim \frac{1}{N}\sum_i \mathrm{E}[\dot{q}_i(\boldsymbol{v}^*)\dot{q}_i(\boldsymbol{v}^*)']$

2. $\frac{1}{N}\sum_i \ddot{q}_i(\boldsymbol{v}^*) = \bar{\boldsymbol{S}}_y \xrightarrow{p} \boldsymbol{S}_y$

Standard m-estimation theory then implies

$$\sqrt{N}(\hat{\boldsymbol{v}}_{-1} - \boldsymbol{v}^*_{-1}) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_{\hat{v}_{-1}}]$$

with $\boldsymbol{\Sigma}_{\hat{v}_{-1}} = \boldsymbol{S}_y^{-1}\boldsymbol{S}_{qq'}\boldsymbol{S}_y^{-1}$. Because of the sum-to-1 condition it will immediately follow for the full vector

$$\sqrt{N}(\hat{\boldsymbol{v}} - \boldsymbol{v}^*) = \sqrt{N}\boldsymbol{R}(\hat{\boldsymbol{v}}_{-1} - \boldsymbol{v}^*_{-1}) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_{\hat{v}}]$$

with $\boldsymbol{\Sigma}_{\hat{v}} = \boldsymbol{R}\boldsymbol{\Sigma}_{\hat{v}_{-1}}\boldsymbol{R}'$ and $\boldsymbol{v}^* = \boldsymbol{e}_1 + \boldsymbol{R}\boldsymbol{v}^*_{-1}$. Note that both $\boldsymbol{\Sigma}_{\hat{v}_{-1}}$ and $\boldsymbol{\Sigma}_{\hat{v}}$ have rank $T_0 - 1$.

Suppose now that $\boldsymbol{v}^*$ lies on the boundary of $\mathbb{V}$. That is, at least one element is exactly zero and $\boldsymbol{v}^*$ is sparse. In general, asymptotic normality of extremum estimators can break down near the boundary of the parameter space (Ketz, 2018). I provide conditions such that the part of $\hat{\boldsymbol{v}}$ belonging to the non-zero elements of $\boldsymbol{v}^*$ will be asymptotically normal and the remaining part is negligible. Suppose without loss of generality that the first k elements $(0 < k < T_0 - 1)$ of $\boldsymbol{v}^*$ are zero and write $\boldsymbol{v} = (\boldsymbol{v}_k', \boldsymbol{v}_{-k}')'$.

Assumption 8. *Let $\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}) = \frac{\partial Q_\infty}{\partial \boldsymbol{v}'}$ the gradient of the population objective. It holds that $\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i \neq 0$ for all $i = 1, \ldots, k$.*

The assumption above excludes cases in which the unconstrained optimal weight in some period (after dropping the corresponding non-negativity constraint) is "coincidentally" exactly zero. I will show that in this case $\Pr(\hat{\boldsymbol{v}}_k = 0) \to 1$. Then the asymptotic distribution of $\hat{\boldsymbol{v}}$ is the same as if we had just set $\hat{\boldsymbol{v}}_k = 0$ in the first place. But this means that the first k periods are irrelevant and we have reduced the exercise to $T_0 - k$ pre-treatment periods, for which the pseudo-true weights $\boldsymbol{v}^*_{-k}$ are not sparse. Then we can apply the results from the previous paragraph to get asymptotic normality for the non-zero elements. Let $C^2(\mathbb{V})$ be the space of twice-differentiable, strictly convex functions and define $\boldsymbol{v}^Q = \arg\min_{\boldsymbol{v}\in\mathbb{V}} Q(\boldsymbol{v})$ for $Q \in C^2(\mathbb{V})$.

Lemma 3. *If there exists a neighborhood $B_0(Q_\infty) \subset C^2(\mathbb{V})$ around Q_∞ such that $\boldsymbol{v}_k^Q = \boldsymbol{0}$ for all $Q \in B_0(Q_\infty)$, Then $\Pr[\hat{\boldsymbol{v}}_k = 0] \to 1$ as $N \to \infty$ and thus $\sqrt{N}\hat{\boldsymbol{v}}_k \xrightarrow{p} 0$.*

Proof. First, $\Pr[\hat{\boldsymbol{v}}_k = 0] = \Pr(Q_N \in B_0(Q_\infty)) \to 1$ since $Q_N \xrightarrow{p} Q_\infty$ uniformly. The last assertion follows from $\Pr(||\sqrt{N}\hat{\boldsymbol{v}}_k|| > \delta) \leq 1 - \Pr(\hat{\boldsymbol{v}}_k = \boldsymbol{0})$ for any $\delta > 0$. □

Lemma 4. *If $\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i \neq 0$ for all $i = 1, \ldots, k$, then there exists a neighborhood $B_0(Q_\infty) \subset C^2(\mathbb{V})$ around Q_∞ such that $\boldsymbol{v}_k^Q = \boldsymbol{0}$ for all $Q \in B_0(Q_\infty)$.*

Proof. Applying the Karush-Kuhn-Tucker Theorem, either $v_i^Q = 0$ or $\boldsymbol{\nabla} Q(\boldsymbol{v}^Q)_i = 0$. Hence $\boldsymbol{\nabla} Q(\boldsymbol{v}^Q)_i \neq 0 \Rightarrow v_i^Q = 0$. By continuity and differentiability, for any $\delta > 0$ we can find $B_\delta(Q_\infty)$ such that $\max_i |\boldsymbol{\nabla} Q(\boldsymbol{v}^Q)_i - \boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i| < \delta$ for all $Q \in B_\delta(Q_\infty)$. Choosing $0 < \delta < \min_i |\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i|$, we have $\min_i |\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i| \geq \min_i |\boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i|| - \max_i |\boldsymbol{\nabla} Q(\boldsymbol{v}^Q)_i - \boldsymbol{\nabla} Q_\infty(\boldsymbol{v}^*)_i| > 0$. □

A.3 Proof of Theorems 3 and 5

Let $\boldsymbol{z}_i = (D_i, \boldsymbol{y}_i')$. Write $\hat{\tau}$ as the solutions to $\frac{1}{N}\sum_i g(\boldsymbol{z}_i, \tau, \hat{\boldsymbol{v}}) = 0$ with

$$g(\boldsymbol{z}, \tau, \boldsymbol{v}) = [Dn_0 - (1-D)(1-n_0)]\boldsymbol{v}_a'\dot{\boldsymbol{y}} - \tau[Dn_0^2 - (1-D)(1-n_0)^2]$$

and $\hat{\boldsymbol{v}}$ as the solution to $\frac{1}{N}\sum_i \boldsymbol{m}(\boldsymbol{z}_i, \boldsymbol{v}) = 0$ with

$$\boldsymbol{m}(\boldsymbol{z}, \boldsymbol{v}) = (1-D)\boldsymbol{v}_a'\dot{\boldsymbol{y}}\boldsymbol{R}'\dot{\boldsymbol{y}}_{p,+}$$

with $\dot{\boldsymbol{y}}_{p,+}$ the T_+ vector containing only observations from pre-treatment periods with non-zero time weights and $\boldsymbol{R} = (\boldsymbol{I}_{T_+-1}, -\boldsymbol{\iota}_{T_+-1})'$. Then, according to Theorem 6.1 of Newey and McFadden (1994)

$$\sqrt{N}(\hat{\tau}(\hat{\boldsymbol{v}}) - \tau) \xrightarrow{d} \mathcal{N}[0, \Sigma_.]$$

with $\Sigma_. = \frac{1}{g_\tau^2}\,\mathrm{E}[\{g(z) + \boldsymbol{g}_v'\boldsymbol{\psi}(z)\}\{g(z) + \boldsymbol{g}_v'\boldsymbol{\psi}(z)\}']$ where

$$\begin{aligned} g(z) &= g(z, \tau_0, \boldsymbol{v}^*) = \boldsymbol{v}_a'[Dn_0(\boldsymbol{y}_i(1) - \bar{\boldsymbol{y}}^{(1)}) - (1-D)(1-n_0)(\boldsymbol{y}_i(0) - \bar{\boldsymbol{y}}^{(0)})] + o_p(1) \\ g_\tau &= \mathrm{E}[\nabla_\tau g(z, \tau_0, \boldsymbol{v}^*)] = -n_0(1-n_0) \\ \boldsymbol{g}_v &= \mathrm{E}[\boldsymbol{\nabla}_v g(z, \tau_0, \boldsymbol{v}^*)] = -g_\tau \boldsymbol{R}'\boldsymbol{F}_{\mathrm{pre}}\boldsymbol{\xi}_\lambda \\ \mathrm{E}[\boldsymbol{\nabla}_v \boldsymbol{m}(z, \boldsymbol{v}^*)] &= -\boldsymbol{S}_y \\ \boldsymbol{\psi}(z) &= \boldsymbol{S}_y^{-1}\boldsymbol{m}(z, \boldsymbol{v}^*) \end{aligned}$$

It also holds that $\frac{1}{g_\tau^2}\,\mathrm{E}[g(z)^2] = \Sigma_{\hat{\tau}}(\boldsymbol{v}^*)$ and $\mathrm{E}[\boldsymbol{\psi}(z)\boldsymbol{\psi}(z)'] = \boldsymbol{\Sigma}_{\hat{v}} = \boldsymbol{S}_y^{-1}\boldsymbol{S}_{qq'}\boldsymbol{S}_y^{-1}$ with $\boldsymbol{S}_{qq'} = \mathrm{E}[\boldsymbol{m}(z, \boldsymbol{v}^*)\boldsymbol{m}(z, \boldsymbol{v}^*)']$. This implies the statements of Theorem 3 with $\boldsymbol{\Sigma}_{\hat{\tau},\hat{v}} := \frac{1}{g_\tau}\,\mathrm{E}[\boldsymbol{\psi}(z)g(z)]$.

This can be extended to the case of multiple treated periods by formulating everything in terms of the T_1 vector $\boldsymbol{\tau}$ and the T_0T_1 stacked weight vector $\boldsymbol{v} = \mathrm{vec}\,\boldsymbol{V} = (\boldsymbol{v}^(T_0 + 1), \ldots, \boldsymbol{v}^(T))$. The joint moment conditions are obtained by stacking the moments for each post-treatment period

$$\boldsymbol{g}(z, \boldsymbol{\tau}, \boldsymbol{v}) = \begin{pmatrix} g_1(z, \tau_{T_0+1}, \boldsymbol{v}^{(T_0+1)}) \\ \vdots \\ g_{T_1}(z, \tau_{T_0+1}, \boldsymbol{v}^{(T)}) \end{pmatrix}; \qquad \boldsymbol{m}(z, \boldsymbol{\tau}, \boldsymbol{v}) = \begin{pmatrix} \boldsymbol{m}_1(z, \boldsymbol{v}^{(T_0+1)}) \\ \vdots \\ \boldsymbol{m}_{T_1}(z, \boldsymbol{v}^{(T)}) \end{pmatrix}$$

We therefore have $\sqrt{N}(\hat{\boldsymbol{\tau}}(\hat{\boldsymbol{V}}) - \boldsymbol{\tau}) \xrightarrow{d} \mathcal{N}[0, \boldsymbol{\Sigma}_.]$ with

$$\boldsymbol{\Sigma}_. = \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V}^*) + \boldsymbol{G}_{\boldsymbol{v}}\,\mathrm{E}[\boldsymbol{\psi}(z)\boldsymbol{\psi}(z)']\boldsymbol{G}_{\boldsymbol{v}}' - 2\boldsymbol{G}_{\boldsymbol{v}}\,\mathrm{E}[\boldsymbol{\psi}(z)\boldsymbol{g}(z)']$$

where $\mathrm{E}[\boldsymbol{g}(z)\boldsymbol{g}(z)'] = \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V}^*)$. Since $\frac{\partial g_j}{\partial \tau_k} = \frac{\partial g_j}{\partial \boldsymbol{v}^{(k)}} = \frac{\partial \boldsymbol{m}_j}{\partial \boldsymbol{v}^{(k)}} = 0$ for $k \neq j$, we have $\boldsymbol{G}_{\boldsymbol{v}} = \mathrm{diag}_{j=1,\ldots,T_1}((\boldsymbol{R}_j'\boldsymbol{F}_{\mathrm{pre}}\xi_\lambda)')$ and $\mathrm{E}[\boldsymbol{\psi}\boldsymbol{\psi}'] = \mathrm{diag}_{j=1,\ldots,T_1}(\boldsymbol{\Sigma}_{v^{T_0+j}})$ Therefore $\boldsymbol{G}_{\boldsymbol{v}}\,\mathrm{E}[\boldsymbol{\psi}\boldsymbol{\psi}']\boldsymbol{G}_{\boldsymbol{v}}'$ is a block-diagonal matrix. However, since all moments use the same pre-treatment observations, in general $\mathrm{E}[\boldsymbol{m}_j(z)\boldsymbol{g}_k(z)'] \neq 0$ for all j, k.

A.4 Proof of Theorem 4

Lemma 5 (Consistency of sample variances). *Suppose Assumptions 2 and 4-7 hold. Then, as* $N \to \infty$

1. $\frac{1}{N_j}\sum_{D_i=j}(\boldsymbol{\lambda}_i - \bar{\boldsymbol{\lambda}}^{(j)})(\boldsymbol{\lambda}_i - \bar{\boldsymbol{\lambda}}^{(j)})' \xrightarrow{p} \boldsymbol{\Sigma}_\lambda^{(j)}$ *for* $j = 0, 1$.
2. $\frac{1}{N_j}\sum_{D_i=j}(\boldsymbol{\varepsilon}_i - \bar{\boldsymbol{\varepsilon}}^{(j)})(\boldsymbol{\varepsilon}_i - \bar{\boldsymbol{\varepsilon}}^{(j)})' \xrightarrow{p} \boldsymbol{\Sigma}_\varepsilon^{(j)}$ *for* $j = 0, 1$.
3. $\frac{1}{N_1}\sum_{D_i=1}(\boldsymbol{\tau}_i - \bar{\boldsymbol{\tau}}^{(1)})(\boldsymbol{\tau}_i - \bar{\boldsymbol{\tau}}^{(1)})' \xrightarrow{p} \boldsymbol{\Sigma}_\tau$

and thus

$$\widehat{\boldsymbol{\Sigma}}_y^{(j)} \xrightarrow{p} \boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{(j)}\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon^{(j)} + 1[j=1]\boldsymbol{B}\boldsymbol{\Sigma}_\tau\boldsymbol{B}'$$

Proof. Write $\boldsymbol{y}_i = \boldsymbol{F}\boldsymbol{\lambda}_i + D_i\boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i$ with the $T \times T_1$ matrix $\boldsymbol{B} = (\boldsymbol{0}, \boldsymbol{I}_{T_1})'$. Since sufficiently many moments are bounded, the sample variances converge to the population equivalents. □

Recall that $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\boldsymbol{V}) = \boldsymbol{V}_a'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Delta}}\boldsymbol{V}_a$ with $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\Delta}} = \frac{\widehat{\boldsymbol{\Sigma}}_y^{(0)}}{n_0} + \frac{\widehat{\boldsymbol{\Sigma}}_y^{(1)}}{1-n_0}$ and $\boldsymbol{V}_a = \begin{pmatrix} -\boldsymbol{V} \\ \boldsymbol{I}_{T_1} \end{pmatrix}$ Lemma 5 thus implies $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\boldsymbol{V}) \xrightarrow{p} \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V})$ for any given $\boldsymbol{V}$. It also follows that $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\widehat{\boldsymbol{V}}) \xrightarrow{p} \boldsymbol{\Sigma}_{\hat{\tau}}(\boldsymbol{V}^*)$ if $\widehat{\boldsymbol{V}} \xrightarrow{p} \boldsymbol{V}^*$. Also note that $\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\widehat{\boldsymbol{V}}) = \frac{1}{N}\sum_i \hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i'$ with $\hat{\boldsymbol{g}}_i = \boldsymbol{g}(\boldsymbol{z}_i, \hat{\tau}, \hat{\boldsymbol{v}})$.

Likewise, a consistent estimator of the weight variance $\boldsymbol{\Sigma}_{\hat{v}_{-1}}$ is

$$\widehat{\boldsymbol{\Sigma}}_{\hat{v}_{-1}} = \bar{\boldsymbol{S}}_y^{-1}\bar{\boldsymbol{S}}_{qq'}\bar{\boldsymbol{S}}^{-1}; \quad \bar{\boldsymbol{S}}_{y,q} = \frac{1}{N}\sum_i \dot{q}(\hat{\boldsymbol{v}}, \boldsymbol{y}_i)\dot{q}(\hat{\boldsymbol{v}}, \boldsymbol{y}_i)' = \frac{1}{N}\boldsymbol{R}'\boldsymbol{Y}_{\mathrm{pre}}'\widehat{\boldsymbol{\Omega}}_q\boldsymbol{Y}_{\mathrm{pre}}\boldsymbol{R}$$

with $\widehat{\boldsymbol{\Omega}}_q = \mathrm{diag}(q_1(\hat{\boldsymbol{v}}), \ldots, q_{N_0}(\hat{\boldsymbol{v}}))$. Ignoring the covariance term, the variance estimators proposed in (8) and (13) Taking into account the covariance, one can use

$$\widetilde{\boldsymbol{\Sigma}}_{\cdot} = \widehat{\boldsymbol{\Sigma}}_{\cdot} - \boldsymbol{G}_{\boldsymbol{v}}\frac{2}{N}\sum_i \hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{g}}_i'$$

The estimators correspond to (6.12) and (6.11) of Newey and McFadden (1994), respectively.

B Additional Derivations

B.1 Proof of selected statements

Lemma 6. *In case of one treated period, the time-weighted DID estimator $\hat{\tau}(\hat{\boldsymbol{v}}) = \boldsymbol{v}_a'(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(0)})$ solves the weighted two-way fixed effect regression problem*

$$\min_{\tau,\boldsymbol{\mu},\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{t=1}^{T} v_t(y_{it} - \tau D_{it} - \mu_i - \gamma_t)^2$$

where $v_T = 1$.

Proof. Follows from Arkhangelsky et al. (2021) using equal unit weights. □

Lemma 7. *For the case of one treated period, $\widehat{\Sigma}_{\hat{\tau}}(\boldsymbol{v})$ (now a scalar) corresponds to the cluster covariance matrix (CCM) estimator (Arellano, 1987) applied to a two-way fixed effects regression of the time-weighted outcomes on the treatment indicator.*

Proof. The CCM estimator is $\widehat{\Sigma}_{ccm} = \frac{N^{-1}\sum \dot{\boldsymbol{D}}_i'\boldsymbol{M}_\iota\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i'\boldsymbol{M}_\iota\dot{\boldsymbol{D}}_i}{(N^{-1}\sum_i \dot{\boldsymbol{D}}_i'\boldsymbol{M}_\iota\boldsymbol{D}_i')^2}$ with residuals $\hat{\boldsymbol{u}}_i = \boldsymbol{\Upsilon}\dot{\boldsymbol{y}}_i - \dot{\boldsymbol{D}}_i\hat{\tau}(\boldsymbol{v})$ with $\hat{\tau}(\boldsymbol{v}) = \boldsymbol{P}_b\boldsymbol{\Delta}_y$ and $\boldsymbol{P}_b = \frac{1}{t_0}\boldsymbol{b}'\boldsymbol{M}_\iota\boldsymbol{\Upsilon} = \boldsymbol{v}_a'$. With the binary treatment structure we get

$$\begin{aligned}
\dot{\boldsymbol{D}}_i &= \dot{D}_i\boldsymbol{b}\\
\dot{D}_i &= D_i - \frac{1}{N}\sum_i D_i = n_0 D_i - (1-n_0)(1-D_i)\\
N^{-1}\sum_i \dot{D}_i^2 &= n_0(1-n_0)^2 + (1-n_0)n_0^2 = n_0(1-n_0)\\
\boldsymbol{b}'\boldsymbol{M}_\iota\boldsymbol{b} = N^{-1}\sum_i \dot{\boldsymbol{D}}_i'\boldsymbol{M}_\iota\dot{\boldsymbol{D}}_i' &= N^{-1}\sum_i \dot{D}_i^2\boldsymbol{b}\boldsymbol{M}_\iota\boldsymbol{b} = t_0 n_0(1-n_0)
\end{aligned}$$

Next, using $\dot{\boldsymbol{y}}_i = \boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(1)} + n_0\boldsymbol{\Delta}_y = \boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(0)} + (1-n_0)\boldsymbol{\Delta}_y$ and $\boldsymbol{\Upsilon}\boldsymbol{b} = \boldsymbol{b}$,

$$\begin{aligned}
\hat{\boldsymbol{u}}_i &= \boldsymbol{\Upsilon}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(1)} + n_0\boldsymbol{\Delta}_y) - n_0\boldsymbol{b}\boldsymbol{P}_b\boldsymbol{\Delta}_y\\
&= \boldsymbol{\Upsilon}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(1)} + n_0\boldsymbol{M}_b\boldsymbol{\Delta}_y)
\end{aligned}$$

for $i \in \mathbb{N}_1$ with $\boldsymbol{M}_b = 1 - \boldsymbol{b}\boldsymbol{P}_b$ and

$$\hat{\boldsymbol{u}}_i = \boldsymbol{\Upsilon}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(0)} - (1-n_0)\boldsymbol{M}_b\boldsymbol{\Delta}_y), \quad i \in \mathbb{N}_0$$

Using that $\boldsymbol{b}'\boldsymbol{M}\boldsymbol{\Upsilon}\boldsymbol{M}_b = \boldsymbol{0}$, the numerator becomes

$$\begin{aligned}
\sum_i \dot{\boldsymbol{D}}_i'\boldsymbol{M}\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i'\boldsymbol{M}\dot{\boldsymbol{D}}_i &= \boldsymbol{b}'\boldsymbol{M}\left[(1-n_0)^2\sum_{i\in\mathbb{N}_0}\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i' + n_0^2\sum_{i\in\mathbb{N}_1}\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i'\right]\boldsymbol{M}\boldsymbol{b}\\
&= N n_0^2(1-n_0)^2 t_0^2\boldsymbol{v}_a'\left[\frac{\widehat{\boldsymbol{\Sigma}}_y^{(0)}}{n_0} + \frac{\widehat{\boldsymbol{\Sigma}}_y^{(1)}}{1-n_0}\right]\boldsymbol{v}_a
\end{aligned}$$

Consequently, $\widehat{\Sigma}_{ccm} = \boldsymbol{v}_a'\widehat{\boldsymbol{\Sigma}}_\Delta\boldsymbol{v}_a = \widehat{\Sigma}_{\hat{\tau}}(\boldsymbol{v})$ □

B.2 Bias decomposition

Let $\mathbb{V}_0 = \arg\min_{\boldsymbol{v}\in\mathbb{V}} \boldsymbol{\xi}_f(\boldsymbol{v})'\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{\xi}_f(\boldsymbol{v})$ be the set of weights that minimizes the factor imbalance. Let $\boldsymbol{R} = (-\boldsymbol{\iota}, \boldsymbol{I}_{T_0-1})'$ and $\boldsymbol{R}_a = (\boldsymbol{R}', \boldsymbol{0})'$.

Lemma 8. *Suppose $r = T_0 - 1$ and $\boldsymbol{v}^*, \boldsymbol{v}_0, \boldsymbol{v}_\varepsilon$ all have strictly positive elements. Then $\mathbb{V}_0$ contains of one element $\boldsymbol{v}_0$ and*

$$\boldsymbol{\xi}_f(\boldsymbol{v}^*) = \boldsymbol{\xi}_f(\boldsymbol{v}_0) + \boldsymbol{F}_0'\boldsymbol{R}(\boldsymbol{I} + \boldsymbol{A}_{\text{snr}})^{-1}\boldsymbol{R}^-(\boldsymbol{v}_0 - \boldsymbol{v}_\varepsilon)$$

with $\boldsymbol{A}_{\text{snr}} = \boldsymbol{A}_f\boldsymbol{A}_\varepsilon^{-1}$, $\boldsymbol{A}_f = \boldsymbol{R}_a'\boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{F}'\boldsymbol{R}_a$ and $\boldsymbol{A}_\varepsilon = \boldsymbol{R}_a'\boldsymbol{\Sigma}_\varepsilon^{(0)}\boldsymbol{R}_a$.

Proof. Uniqueness of $\boldsymbol{v}_0$ follows from full rank of $\boldsymbol{R}'\boldsymbol{F}_0\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{F}_0'\boldsymbol{R}$ for $r = T_0 - 1$. Write $\boldsymbol{\xi}_f(\boldsymbol{v}^*) = \boldsymbol{\xi}_f(\boldsymbol{v}_0) + \boldsymbol{F}_0'\boldsymbol{R}\boldsymbol{R}^-(\boldsymbol{v}^* - \boldsymbol{v}_0)$. We have $\boldsymbol{R}^-\boldsymbol{v}^* = (\boldsymbol{A}_f + \boldsymbol{A}_\varepsilon)^{-1}\boldsymbol{R}_a'(\boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{F}' + \boldsymbol{\Sigma}_\varepsilon^{(0)})\boldsymbol{e}_{1,T}$, $\boldsymbol{R}^-\boldsymbol{v}_0 = \boldsymbol{A}_f^{-1}\boldsymbol{R}_a'\boldsymbol{F}\boldsymbol{\Sigma}_\lambda^{(0)}\boldsymbol{F}'\boldsymbol{e}_{1,T}$ and $\boldsymbol{R}^-\boldsymbol{v}_\varepsilon = \boldsymbol{A}_\varepsilon^{-1}\boldsymbol{R}_a'\Sigma_\varepsilon^{(0)}\boldsymbol{e}_{1,T}$. Therefore $\boldsymbol{R}^-(\boldsymbol{v}^* - \boldsymbol{v}_0) = (\boldsymbol{I} + \boldsymbol{A}_{\text{snr}})^{-1}\boldsymbol{R}^-(\boldsymbol{v}_0 - \boldsymbol{v}_\varepsilon)$ □

Under the convex hull condition we additionally have $\boldsymbol{\xi}_f(\boldsymbol{v}_0) = 0$.

Example. Consider the following example with $r = 1$, $T_0 = 2$, one treated period and $\boldsymbol{\Sigma}_\varepsilon^{(0)} = \sigma_\varepsilon^2 \cdot \text{diag}(1-\omega, \omega, 1)$ for some $\omega \in [0,1]$ (no time dependence in ε_{it}). Let $\boldsymbol{f} = \sigma_f \cdot (0, 1, \alpha_0)'$ for some $\alpha_0 \in [0,1]$ and $\sigma_f > 0$. Hence the oracle weights are $\boldsymbol{v}_0 = (1-\alpha_0, \alpha_0)'$. The pseudo true weights $\boldsymbol{v}^* = (1-\alpha^*, \alpha^*)'$ solve

$$\min_{\alpha\in[0,1]} \sigma_f^2(\alpha - \alpha_0)^2 + (1-\omega)(1-\alpha)^2 + \omega\alpha^2$$

where I set $\boldsymbol{\Sigma}_\lambda^{(0)} = 1$ and $\sigma_\varepsilon^2 = 1$ without loss of generality. Hence σ_f^2 should be interpreted as the signal-to-noise ratio. The solution is $\alpha^* = (1-\omega)\frac{1}{1+\sigma_f^2} + \alpha_0\frac{\sigma_f^2}{1+\sigma_f^2}$, which is a convex combination of the oracle weights α_0 and weights that minimize the error variance $\alpha_\varepsilon := \arg\min_{\alpha\in[0,1]}(1-\omega)(1-\alpha)^2 + \omega\alpha^2 = 1-\omega$. Intuitively, $\alpha^* \to \alpha_\varepsilon$ as $\sigma_f \to 0$ and $\alpha^* \to \alpha_0$ as $\sigma_f \to \infty$. The remaining (squared) factor imbalance is

$$\xi_f(\alpha^*, \sigma_f)^2 = \sigma_f^2(\alpha^* - \alpha_0)^2 = \frac{\sigma_f^2}{(1+\sigma_f^2)^2}(\alpha_0 - \alpha_\varepsilon)^2 \tag{14}$$

We see that $\xi_f(\alpha^*, \sigma_f)^2 \to 0$ as $\sigma_f \to \infty$ – the remaining bias disappears as factors get stronger. Figure 6 plots the remaining bias $\xi_f(\alpha, \sigma_f)$ for the TWDID and DID estimator as a function of the factor strength σ_f. For DID the weights are equal ($\alpha = 0.5$). They correspond to the variance-minimizing weights $\alpha_\varepsilon = 0.5$, but differ from the oracle weights, which I set to $\alpha_0 = 1$. Hence the bias is proportional to σ_f. For TWDID the weights are $\alpha^* = \frac{\frac{1}{2}+\sigma_f^2}{1+\sigma_f^2}$ and converge to α_0 as σ_f increases.

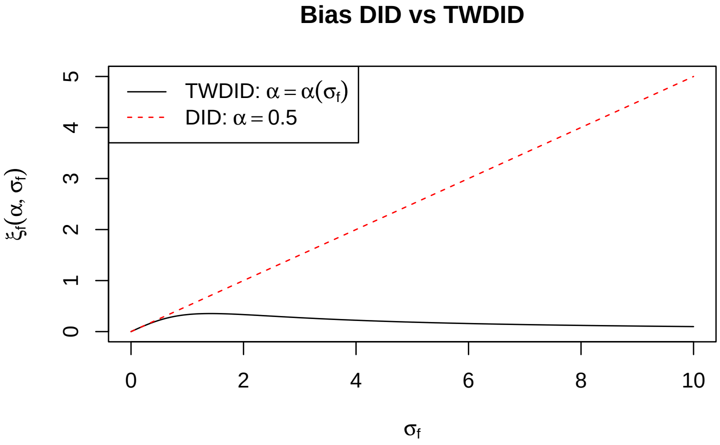


Figure 6: Analytical bias of DID (equal weights $\alpha = 0.5$) vs. Bias of TWDID ($\alpha^*(\sigma_f)$) depending on the factor strength σ_f.

B.3 Controlling for weights representation

Remark 3. The TWDID estimator $\hat{\tau}(\boldsymbol{v}) = \bar{\Delta}_{[1]} - \bar{\Delta}^v_{[0]}$ with $\bar{\Delta}^v_{[0]} = \sum_{t \leq T_0} v_t \Delta_t$ can be written as

$$\hat{\tau}(\boldsymbol{v}) = \frac{\boldsymbol{b}'\boldsymbol{M}_v\boldsymbol{\Delta}}{\boldsymbol{b}'\boldsymbol{M}_v\boldsymbol{b}}$$

with $\boldsymbol{M}_v = \boldsymbol{I}_T - \frac{\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}'}{\tilde{\boldsymbol{v}}'\tilde{\boldsymbol{v}}}$ and $\tilde{\boldsymbol{v}} = (\frac{\boldsymbol{v}'}{\boldsymbol{v}'\boldsymbol{v}}, \boldsymbol{\iota}'_{T_1})'$ the T vector of normalized weights. Hence, instead of the time-weighted 2wfe regression representation (4) we may present $\hat{\tau}(\boldsymbol{v})$ as a 2wfe regression controlling for the normalized time weights $\tilde{\boldsymbol{v}}$. This opens up a new interpretation of time weights. Instead of projecting out the entire factor structure ($\boldsymbol{M}_F$), the time weights make the factors orthogonal to the treatment. $\boldsymbol{M}_v\boldsymbol{F} \neq 0$, but $\boldsymbol{b}'\boldsymbol{M}_v\boldsymbol{F} \approx 0$.

Proof. Start with $\frac{\tilde{\boldsymbol{v}}'\boldsymbol{b}}{\tilde{\boldsymbol{v}}'\tilde{\boldsymbol{v}}} = \frac{T_1}{\frac{1}{\boldsymbol{v}'\boldsymbol{v}}+T_1} = \frac{1-t_0}{\frac{1}{T\boldsymbol{v}'\boldsymbol{v}}+1-t_0} =: c$ and thus $(1-c)T(1-t_0) = \frac{c}{\boldsymbol{v}'\boldsymbol{v}}$. Next, $\boldsymbol{M}_v\boldsymbol{b} = (-\frac{c}{\boldsymbol{v}'\boldsymbol{v}}\boldsymbol{v}', (1-c)\boldsymbol{\iota}'_{T_1})'$ and thus $(\boldsymbol{M}_v\boldsymbol{b})'\boldsymbol{x} = T(1-c)(1-t_0)(\bar{x}_{[1]} - x^v_{[0]})$. Apply it to $\boldsymbol{x} = \boldsymbol{\Delta}$ for the numerator and $\boldsymbol{x} = \boldsymbol{b}$ for the denominator. □

The multivariate case follows accordingly. Estimate the dynamic treatment effects as

$$\hat{\boldsymbol{\tau}}^{\text{att}}(\boldsymbol{V}) = (\boldsymbol{B}'\boldsymbol{M}_v\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{M}_v\boldsymbol{\Delta} = \begin{pmatrix} \Delta_{T_0+1} - \sum_{t \leq T_0} v_t^{(T_0+1)}\Delta_t \\ \vdots \\ \Delta_T - \sum_{t \leq T_0} v_t^{(T)}\Delta_t \end{pmatrix} \tag{15}$$

with $\boldsymbol{B} = [\boldsymbol{b}^{(T_0+1)}, \ldots, \boldsymbol{b}^{(T)}]$ the $T \times T_1$ matrix with dummies for each post-treatment period, $\boldsymbol{M}_v = \boldsymbol{I}_T - \tilde{\boldsymbol{V}}(\tilde{\boldsymbol{V}}'\tilde{\boldsymbol{V}})^{-}\tilde{\boldsymbol{V}}'$ and $\tilde{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-} \\ \boldsymbol{I}_{T_1} \end{pmatrix}$ the $T \times T_1$ matrix of standardized time weights. This is similar to a two-way fixed effects regression of y_{it} on the treatment interacted with dummies for each post treatment period. However, instead of controlling for unit fixed effects $\boldsymbol{\iota}_T$ one controls for the standardized time weights $\tilde{\boldsymbol{V}}$. The CCM variance estimator now becomes the $T_1 \times T_1$ matrix

$$\widehat{\boldsymbol{\Sigma}}_{\hat{\tau}}(\boldsymbol{V}) = (\boldsymbol{B}'\boldsymbol{M}_v\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{M}_v\widehat{\boldsymbol{\Sigma}}_y\boldsymbol{M}_v\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{M}_v\boldsymbol{B})^{-1}$$

with $\widehat{\boldsymbol{\Sigma}}_y = \frac{\widehat{\boldsymbol{\Sigma}}_y^{(0)}}{n_0} + \frac{\widehat{\boldsymbol{\Sigma}}_y^{(1)}}{1-n_0}$ and $\widehat{\boldsymbol{\Sigma}}_y^{(0)} = \frac{1}{N_j}\sum_{i:D_i=j}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})(\boldsymbol{y}_i - \bar{\boldsymbol{y}}^{(j)})'$ the $T \times T$ sample covariance within the treated and untreated units.