What Do We Get from Two-Way Fixed Effects Regressions? Implications From Numerical Equivalence

Shoya Ishimaru*

October 19, 2022

Abstract

In any multiperiod panel, a two-way fixed effects (TWFE) regression is numerically equivalent to a first-difference (FD) regression that pools all possible between-period gaps. Building on this observation, this paper develops numerical and causal interpretations of the TWFE coefficient. At the sample level, the TWFE coefficient is a weighted average of FD coefficients with different between-period gaps. This decomposition is useful for assessing the source of identifying variation for the TWFE coefficient. At the population level, a causal interpretation of the TWFE coefficient requires a common trends assumption for any between-period gap, and the assumption has to be conditional on changes in time-varying covariates. I show that these requirements can be naturally relaxed by modifying the estimator using a pooled FD regression.

*Hitotsubashi University, Department of Economics (email: shoya.ishimaru@hit-u.ac.jp).
1 Introduction

Linear regression methods are favored in contemporary empirical practice for their transparency about how they map data into the estimators even when the linear models that motivate them fail (Angrist and Pischke, 2010). The growing popularity of two-way fixed effects (TWFE) regressions is no exception to this trend. In panel data with units $i = 1, \ldots, N$ and periods $t = 1, \ldots, T$, a TWFE regression is motivated by an equation

$$Y_{it} = X_{it}' \beta + \alpha_i + C_i' \gamma_t + \varepsilon_{it},$$

where $Y_{it}$ is a scalar outcome, $X_{it}$ is a vector of explanatory variables, and $C_i$ is a vector of time-invariant covariates that includes unity. The inclusion of unit effects $\alpha_i$ and time effects $\gamma_t$ arises from a difference-in-differences (DID) identification logic in a two-period panel with a binary treatment. However, many practical applications use TWFE regressions for analyzing multiperiod panels with nonbinary treatment variables. Relatively little has been known about how a TWFE regression maps data into the coefficient in these general settings.

This paper develops general numerical and causal interpretations of the TWFE estimator from a contemporary viewpoint that does not regard the linear regression equation (1) as the true model. My investigation builds on an observation that the following two least-squares problems give algebraically identical estimates of the coefficient $\beta$.

$$\min_{\beta, \{\alpha_i\}_{i=1}^N, \{\gamma_t\}_{t=1}^T} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - X_{it}' \beta - \alpha_i - C_i' \gamma_t)^2,$$

(2)

$$\min_{\beta, \{\gamma_t\}_{t=1}^T} \sum_{i=1}^N \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \{(Y_{i,t+k} - Y_{it}) - (X_{i,t+k} - X_{it})' \beta - C_i' (\gamma_{t+k} - \gamma_t)\}^2.$$

(3)

(2) is a TWFE regression that is directly motivated by the equation (1), whereas (3) is a pooled first difference (FD) regression that is motivated by differences of the equation (1) for all possible between-period gaps $k = 1, \ldots, T - 1$. This is a natural extension of the well-known equivalence between TWFE and FD regressions in a two-period panel ($T = 2$) to any multiperiod panel ($T \geq 2$).

This observation, which implies that TWFE regressions capture both short-run and long-
run associations between changes in \( Y_{it} \) and changes in \( X_{it} \), is not an intuitively unexpected result. Nor is it a theoretically novel discovery on its own; using a model without period-specific effects in (1), Han and Lee (2017, 2022) suggest an equivalence result that corresponds to the equivalence between (2) and (3). However, this general numerical property and its implications for contemporary empirical practice appear to have been largely overlooked in the literature on TWFE regressions. Han and Lee (2017, 2022) discuss its implications only in the context of the within estimator from a traditional standpoint that regards a linear regression equation as the true model. While Strezhnev (2018) and Goodman-Bacon (2021) uncover that the TWFE estimator aggregates all possible pairwise DID comparisons, their results apply only to binary \( X_{it} \).

The equivalence between TWFE and pooled FD regressions has several implications. First, at the most fundamental level, this equivalence reveals the mechanics of a TWFE regression that is not immediately evident from its original least-squares objective; it captures the association among changes in variables of interest, just like a cross-sectional regression picking up the relationship among levels of variables of interest.

Second, it implies that the TWFE estimator can be expressed as a weighted average of the FD estimators with all possible between-period gaps. This decomposition is possible because the pooled FD objective (3) aggregates least-squares objectives of FD regressions across \( k = 1, \ldots, T - 1 \). For example, suppose \( X_{it} \) is univariate and let \( \hat{\beta}_{FE} \) be the TWFE coefficient given by (2) or (3). Defining \( \hat{\beta}_{FD,k} \) to be the coefficient from an FD regression motivated by \( k \)-period difference of (1), \( \hat{\beta}_{FE} \) can be expressed as

\[
\hat{\beta}_{FE} = \sum_{k=1}^{T-1} \hat{w}_k \hat{\beta}_{FD,k},
\]

where \( \hat{w}_k \geq 0 \) and \( \sum_{k=1}^{K-1} \hat{w}_k = 1 \). This decomposition provides a diagnostic tool for uncovering how a TWFE regression aggregates short-run and long-run associations between changes in \( Y_{it} \) and changes in \( X_{it} \) into a single coefficient.

Third, the objective function (3) indicates that a causal interpretation of the TWFE coefficient can be developed solely from assumptions on the relationship among changes in relevant variables, without viewing (1) as the true causal model. Exploiting this insight, I explore a causal interpretation of the TWFE coefficient under a potential outcome framework when \( X_{it} \) consists of a scalar treatment and a vector of time-varying covariates. As can be inferred from (3), the causal interpretation requires a common trends assumption to be satisfied for all possible \( k \)-period differences. Furthermore, the assumption has to be conditional on changes, not levels, of time-varying covariates. This assumption diverges
from an identification strategy for a two-period DID with covariates (Heckman et al., 1997, 1998; Abadie, 2005), which often employs pre-treatment observables.

Finally, the objective function (3) can be modified to overcome these limitations and to allow for greater flexibility in mapping the data into the coefficient, which I call a generalized TWFE regression. For example, “too short” or “too long” differences can be excluded from (3), to improve the credibility of the common trends assumption or to focus on relevant time frames in measuring the effects. In addition, any time-dependent variables can be added to (3) as regressors, which enables a causal interpretation under a common trends assumption conditional on levels of these variables, instead of their changes.

Using the TWFE estimates of the minimum wage effect on employment outcomes in the U.S. state–year panel, I demonstrate the practical importance of these insights. The decomposition of the TWFE coefficient reveals that constituent short-run and long-run FD estimates disagree with one another. This result suggests the importance of being aware of the mechanics of how the TWFE regression maps the data into the coefficient, as well as the practical relevance of the new estimator that enables the selective use of between-period differences. For example, while the standard TWFE estimate suggests a 10% minimum wage increase is associated with a statistically insignificant 0.04 percentage point decline in the net job creation rate, a generalized TWFE regression that uses only 1–5 year differences suggests the impact that is ten times larger. I also demonstrate that a generalized TWFE regression enables a common trends assumption conditional on pre-existing dynamics of employment and minimum wage levels. This is not possible for the standard TWFE regression, which requires covariate changes to be the basis for the conditional common trends assumption.

This paper is related to the recent econometric literature on TWFE regressions. The literature focuses on DID and event-study settings with a binary treatment, and intensively investigates properties of the TWFE estimator with univariate or multivariate binary regressors. While the literature offers many important insights in binary treatment settings, providing a general result that applies to any TWFE regression is the focus and the contribution of this paper.

More broadly, this paper is a part of continuing efforts of the econometric literature to improve the transparency of linear regression estimators. The literature has explored how

---

Athey and Imbens (2022), Callaway and Sant’Anna (2020), Goodman-Bacon (2021), and Strezhnev (2018) study a setting with a binary treatment and a staggered adoption design, while and de Chaisemartin and D’Haultfoeuille (2020a) consider a binary treatment with a general path of adoption. Borusyak et al. (2021), Lin and Zhang (2022), Schmidheiny and Siegloch (2020), and Sun and Abraham (2020) explore event-study settings. de Chaisemartin and D’Haultfoeuille (2020b) and Goldsmith-Pinkham et al. (2022) consider a setting with multiple binary treatment variables. Callaway et al. (2021) and Wooldridge (2021) are notable exceptions studying nonbinary treatment settings.

---
researchers can interpret the linear regression estimators without assuming that the linear models that motivate them describe the true causal relationships. Similarly to discussions in the literature, this paper provides both optimistic and cautionary views on the linear regression estimator: the TWFE estimator may have a general causal interpretation beyond the linear model, but it requires careful consideration of what it identifies under what assumptions.

The rest of this paper is organized as follows. Section 2 investigates numerical properties of a TWFE regression, including its equivalence with a pooled FD regression and a weighted-average interpretation of the TWFE coefficient. Section 3 offers a causal interpretation of the TWFE coefficient under a potential outcome framework. Section 4 discusses how a TWFE regression can be extended using the equivalent FD objective. Section 5 illustrates the practical relevance of these results. Section 6 concludes.

2 Numerical Properties

This section develops a numerical interpretation of the TWFE estimator. I focus on a balanced panel for expositional convenience; Appendix B discusses how the results in the main paper extend to or differ in an unbalanced panel. To simplify exposition, I hereafter use $\Delta_k a_t \equiv a_{t+k} - a_t$ to denote a $k$-period difference and $\bar{a} \equiv \frac{1}{T} \sum_{t=1}^T a_t$ to denote the average of a time series $\{a_t\}_{t=1}^T$.

2.1 Equivalence of Least-Square Objectives

In a two-period panel, the numerical equivalence between TWFE and FD regressions is well known. In a multiperiod panel with a binary treatment and a staggered adoption design, Strezhnev (2018) and Goodman-Bacon (2021) show that the TWFE estimator is equivalent to a weighted average of all possible pairwise DID comparisons. The following result reveals that the equivalence among TWFE, FD, and DID approaches holds in more general settings.

**Theorem 1.** Let $\widehat{\beta}_{FE}$ be the coefficient on $X_{it}$ given by (2). $\beta = \widehat{\beta}_{FE}$ is the solution to (3).

---

4See Angrist and Krueger (1999) and Angrist and Pischke (2008) for the overview. Mogstad and Wiswall (2010), Løken et al. (2012), Lochner and Moretti (2015), Słoczyński (2020a,b), and Ishimaru (2022) are examples of recent takes.
and is also the solution to the following least-squares problem\footnote{In all these least-squares objectives, one of the time effects \(\{\gamma_t\}_{t=1}^{T}\) is indeterminate due to linear dependence. Moreover, since one of elements of \(C_i\) is unity, the corresponding element of \(\gamma_t\) is indeterminate in \(\ref{eq:least-squares-problem}\) for all \(t = 1, \ldots, T\). To simplify exposition, I do not normalize these parameters.}. \vspace{0.5em}

\[
\min_{\beta, \{\gamma_t\}_{t=1}^{T}} \sum_{1 \leq i < j \leq N} \sum_{k=1}^{T-k} \sum_{t=1}^{T-1} ((\Delta_k Y_{jt} - \Delta_k Y_{it}) - (\Delta_k X_{jt} - \Delta_k X_{it})' \beta - (C_j - C_i)' \Delta_k \gamma_t) \quad (5)
\]

The equivalence between \(\ref{eq:twfe-regression}\) and \(\ref{eq:fd-regression}\) implies the equivalence between a TWFE regression and a FD regression that pools all possible \(k\)-period differences. This is not a novel result from a pure theoretical perspective. Using a model without period-specific effects in \(\ref{eq:fd-regression}\), Han and Lee \cite{han2017, han2022} shows that the within estimator can be recovered by a regression of \(\Delta_k Y_{it}\) on \(\Delta_k X_{it}\) with no intercept. Since \(X_{it}\) can include interactions between \(C_i\) and time dummies, their model nests the model \(\ref{eq:fd-regression}\). Han and Lee \cite{han2017, han2022} use this observation for bias correction of the within estimator, treating a linear regression equation as the causal model. My contribution is primarily to point out its various implications for the TWFE estimator without viewing a linear regression equation as the causal model.

The least-squares problem \(\ref{eq:least-squares-problem}\) regresses a DID in \(Y_{it}\) across units and periods on the corresponding DID in \(X_{it}\), controlling for a difference in covariates \(C_i\). If \(C_i\) consists only of unity, then \(\ref{eq:least-squares-problem}\) suggests that \(\hat{\beta}_{FE}\) is the sum of \(\Delta_k Y_{jt} - \Delta_k Y_{it}\) weighted by \(\Delta_k X_{jt} - \Delta_k X_{it}\). Further assuming that \(X_{it}\) is a binary and staggered treatment, \(\hat{\beta}_{FE}\) aggregates two-unit, two-period comparisons in which only one of the two units starts getting treated.\footnote{\((\Delta_k Y_{jt} - \Delta_k Y_{it})(\Delta_k X_{jt} - \Delta_k X_{it})\) is equal to \(\Delta_k Y_{jt} - \Delta_k Y_{it}\) if only \(j\) starts getting treated between periods \(t\) and \(t + k\), to \(\Delta_k Y_{it} - \Delta_k Y_{jt}\) if only \(i\) starts getting treated, and to 0 otherwise.} This implication is closely related to a decomposition suggested by Strezhnev \cite{strezhev2018}. In addition, aggregation by treatment timing groups instead of individual units results in a decomposition suggested by Goodman-Bacon \cite{goodman-bacon2021}. The equivalence between \(\ref{eq:twfe-regression}\) and \(\ref{eq:least-squares-problem}\) generalizes these implications beyond a binary and staggered treatment setting.

\section{2.2 Weighted-Average Relationship}

The objective function \(\ref{eq:fd-regression}\), which is now shown to reproduce the TWFE estimator, pools least-square objectives of FD estimators with different between-period gaps. Taking a \(k\)-period difference of the model \(\ref{eq:fd-regression}\) motivates a least-squares method

\[
\min_{\beta, \{\gamma_t\}_{t=1}^{T}} \sum_{1 \leq i < j \leq N} \sum_{k=1}^{T-k} \sum_{t=1}^{T-1} (\Delta_k Y_{jt} - \Delta_k X_{jt}' \beta - C_j' \Delta_k \gamma_t) \quad (6)
\]
where the period-specific effect parameter $\gamma_{k,t}^*$ originates from a $k$-period difference of the period-specific effect parameter $\gamma_t$ in (1). The pooled FD objective (3) aggregates (6) with the joint restriction $\beta = \beta_k$ and $\gamma_{k,t}^* = \gamma_{t+k} - \gamma_t$ across different $k = 1, \ldots, T - 1$.

Given the resemblance of the objective functions, it is natural to expect a numerical connection between the TWFE coefficient $\hat{\beta}_{FE}$ on $X_{it}$ given by (2) and the FD coefficients $\hat{\beta}_{FD,k}$ on $\Delta_kX_{it}$ given by (6) for $k = 1, \ldots, T - 1$. Define

$$\bar{Y}_{it} \equiv Y_{it} - C_i' \left( \sum_{i=1}^{N} C_i C_i' \right)^{-1} \left( \sum_{i=1}^{N} C_i Y_{it} \right)$$

to be a residual from a regression of $Y_{it}$ on $C_i$ independently performed for each $t = 1, \ldots, T$. Applying the Frisch–Waugh–Lovell (FWL) theorem to (6), $\hat{\beta}_{FD,k}$ is the coefficient from a regression of $\Delta_k\bar{Y}_{it}$ on $\Delta_k\bar{X}_{it}$ independently performed for each $k = 1, \ldots, T - 1$, that is,

$$\hat{\beta}_{FD,k} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \bar{X}_{it} \Delta_k \bar{X}_{it}' \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \bar{X}_{it} \Delta_k \bar{Y}_{it} \right).$$

The restriction $\gamma_{k,t}^* = \gamma_{t+k} - \gamma_t$ imposed on (3) compared with (6) makes it difficult to apply the FWL theorem directly to (3), but it can be shown that $\hat{\beta}_{FE}$ is the coefficient from a regression of $\Delta_k\bar{Y}_{it}$ on $\Delta_k\bar{X}_{it}$ pooled for all $k = 1, \ldots, T - 1$. As a result, the weighted-average interpretation of the TWFE coefficient arises.

**Theorem 2.** Let $\hat{\beta}_{FE}$ be the TWFE coefficient on $X_{it}$ in (2). Then, a regression of $\Delta_k\bar{Y}_{it}$ on $\Delta_k\bar{X}_{it}$ pooled for all $k = 1, \ldots, T - 1$ yields $\hat{\beta}_{FE}$, that is,

$$\hat{\beta}_{FE} = \left( \sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \bar{X}_{it} \Delta_k \bar{X}_{it}' \right)^{-1} \left( \sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \bar{X}_{it} \Delta_k \bar{Y}_{it} \right).$$

(7)

Suppose the FD coefficients $\hat{\beta}_{FD,k}$ on $\Delta_kX_{it}$ in (6) are well-defined for all $k = 1, \ldots, T - 1$. Then,

$$\hat{\beta}_{FE} = \sum_{k=1}^{T-1} \frac{\hat{\beta}_{k,FD}}{\hat{\beta}_{FD,k}},$$

where each weight matrix for $k = 1, \ldots, T - 1$ is defined as

$$\hat{\Omega}_k = \left( \sum_{\ell=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-\ell} \Delta_\ell \bar{X}_{it} \Delta_\ell \bar{X}_{it}' \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \bar{X}_{it} \Delta_k \bar{X}_{it}' \right),$$

\[7\]The proof of Theorem 2 shows that the restriction $\gamma_{k,t}^* = \gamma_{t+k} - \gamma_t$ does not bind in (3).
which are all positive semi-definite and sum to the identity matrix.

When \(X_{it}\) is univariate, Theorem 2 implies that \(\hat{\beta}_{FE}\) is a convex combination of \(\hat{\beta}_{FD,k}\), as presented in (4). If the model (1) is “true” in a textbook sense\(^8\) then the FD coefficients for \(k = 1, \ldots, T - 1\) are all asymptotically equivalent. However, the difference among these coefficients can arise in practice. For example, short-run and long-run effects of changes in \(X_{it}\) may differ. In addition, severity of endogeneity problems may differ across short-run and long-run comparisons. The decomposition provided by Theorem 2 can reveal how the FD coefficients that may disagree with each other are aggregated into a single coefficient.

**Interpretation with a Scalar Treatment and Time-Varying Covariates**

When \(X_{it}\) is multivariate, Theorem 2 suggests a matrix-weighted average interpretation, which may not be very intuitive. To build a more practical insight from this result, I assume that \(X_{it}\) consistsof a scalar treatment variable \(D_{it}\) that produces a causal effect of interest and a vector of time-varying covariates \(W_{it}\), following a typical design-based view on regression equations.\(^9\) Then, the regression equation (1) can be rewritten as

\[
Y_{it} = \beta^D D_{it} + W_{it}'\beta^W + \alpha_i + C_i'\gamma_t + \varepsilon_{it}. \tag{8}
\]

Let \((\hat{\beta}_{FE}^D, \hat{\beta}_{FE}^W)\) be the coefficients of \((D_{it}, W_{it})\) from a TWFE regression motivated by (8), and let \((\hat{\beta}_{FD,k}^D, \hat{\beta}_{FD,k}^W)\) be the coefficients of \((\Delta_k D_{it}, \Delta_k W_{it})\) from a \(k\)-period FD regression motivated by (8). Theorem 2 suggests that \((\hat{\beta}_{FE}^D, \hat{\beta}_{FE}^W)\) is a matrix-weighted average of \((\hat{\beta}_{FD,k}^D, \hat{\beta}_{FD,k}^W)\). This in turn implies that \(\hat{\beta}_{FE}^D\) can be expressed as a weighted average of \(\hat{\beta}_{FD,k}^D\) alone only when all covariates are time-invariant. In the presence of time-varying covariates, a weighted-average expression of \(\hat{\beta}_{FE}^D\) is contaminated by \(\hat{\beta}_{FD,k}^W\).

Now, I examine the mechanics of how this contamination takes place. Using Theorem 2 and the Frisch–Waugh–Lovell (FWL) theorem, \(\hat{\beta}_{FE}^D\) is given by

\[
\hat{\beta}_{FE}^D = \frac{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k Y_{it} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FE}^W \right)}{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k D_{it} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FE}^W \right)}, \tag{9}
\]

where \(\hat{\delta}_{FE}^W\) is the coefficient from a regression of \(\Delta_k \tilde{D}_{it}\) on \(\Delta_k \tilde{W}_{it}\) pooled for all \(k = 1, \ldots, T-1\).

---

\(^8\)This means that the model (1) describes a causal relationship and that a strict exogeneity condition \(E[\varepsilon_{it}|C_i, X_{i1}, \ldots, X_{iT}] = 0\) is satisfied.

\(^9\)Angrist and Pischke (2017, p. 129) remark, “In the modern paradigm, regressors are not all created equal. Rather, only one variable at a time is seen as having causal effects. All others are controls included in service of this focused causal agenda.”
On the other hand, each $\hat{\beta}_{FD,k}^D$ is given by

$$
\hat{\beta}_{FD,k}^D = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k Y_{it} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FD,k}^W \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k D_{it} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FD,k}^W \right)},
$$

(10)

where $\hat{\delta}_{FD,k}^W$ is the coefficient from a regression of $\Delta_k \tilde{D}_{it}$ on $\Delta_k \tilde{W}_{it}'$ independently performed for each $k = 1, \ldots, T - 1$. Apparently, the two expressions (9) and (10) are not directly comparable unless $\hat{\delta}_{FE}^W = \hat{\delta}_{FD,k}^W$ for all $k = 1, \ldots, T - 1$.

The following theorem formally shows that $\hat{\beta}_{FE}^D$ cannot be expressed as a weighted average of $\hat{\beta}_{FD,k}^D$ alone in general, and demonstrates how the weighted-average formula is contaminated by $\hat{\beta}_{FD,k}^W$.

**Theorem 3.** The two-way FE coefficient given in (9) is equivalent to

$$
\hat{\gamma}_{D}^D = \sum_{k=1}^{T-1} \hat{w}_k^D \hat{\gamma}_{FD,k}^D + \hat{B},
$$

(11)

where the weights are defined as

$$
\hat{w}_k^D \equiv \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-k} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FE}^W \right)^2}{\sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-t} \left( \Delta_t \tilde{D}_{it} - \Delta_t \tilde{W}_{it}' \hat{\delta}_{FE}^W \right)^2},
$$

(12)

and the bias term $\hat{B}$ is defined as

$$
\hat{B} = \frac{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \left( \hat{\gamma}_{FD,k}^W - \hat{\gamma}_{FE}^W \right) \left( \Delta_k \tilde{W}_{it}' \Delta_k \tilde{W}_{it}' \hat{\delta}_{FD,k}^W - \hat{\delta}_{FE}^W \right)}{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{W}_{it}' \hat{\delta}_{FD,k}^W \right)^2}.
$$

(13)

Theorem 3 shows that $\hat{\gamma}_{FE}^D$ is equivalent to a weighted average of $\hat{\gamma}_{FD,k}^W$ plus the bias term that depends on a discrepancy between $\hat{\delta}_{FE}^W$ and $\hat{\delta}_{FD,k}^W$ or between $\hat{\beta}_{FE}^W$ and $\hat{\beta}_{FD,k}^W$. In essence, a TWFE regression accounts for the effect of covariate changes $\Delta_k W_{it}$ less flexibly than FD regressions, which collectively allow the coefficients of $\Delta_k W_{it}$ to be heterogeneous across $k = 1, \ldots, T - 1$. This inflexibility results in a deviation from an exact weighted-average interpretation, due to the bias term $\hat{B}$ that depends on heterogeneity of the coefficients on $\Delta_k W_{it}$. While the bias term $\hat{B}$ impedes a clean weighted-average interpretation of the TWFE coefficient $\hat{\gamma}_{FE}^D$, it should not be perceived as a mere residual or decomposition error; Section 3 reveals that it has implications for a causal interpretation of the TWFE coefficient.
3 Causal Interpretation

This section develops a causal interpretation of the population coefficient on $D_{it}$ that arises from a TWFE regression motivated by the equation (8). The equivalence between TWFE and pooled FD regressions implies that the TWFE coefficient on $D_{it}$ is given by a regression of $\Delta_k Y_{it}$ on $\Delta_k D_{it}$ controlling for $\Delta_k W_{it}$ and $C_i$. A causal interpretation of the coefficient can be developed without viewing (8) as the true causal model, in the same manner as how a potential outcome framework with conditional independence assumptions can produce causal interpretations of cross-sectional regression coefficients.

3.1 Setup

To consider a causal interpretation of the population TWFE coefficient with $N \to \infty$ and $T$ being fixed, let $(Y_{it}, D_{it}, W_{it})_{t=1}^T$ and $C_i$ be observations of a scalar outcome $Y_{it}$, a scalar treatment $D_{it}$, a vector of time-varying covariates $W_{it}$, and a vector of time-invariant covariates $C_i$, which are i.i.d. across $i = 1, \ldots, N$. Since it is a large $N$ setting, $E[D_{it}]$ represents the cross-sectional average, while the time-series average is denoted by $\overline{D}_i$. In addition, I use $L_{C_i}(D_{it}) \equiv C_i' (E[C_i C_i']^{-1} E[C_i D_{it}])$ to denote a linear projection of $D_{it}$ onto $C_i$.

I define

$$\ddot{D}_{it} \equiv D_{it} - \overline{D}_i - L_{C_i} (D_{it} - \overline{D}_i).$$

(14)

Defining $\ddot{W}_{it}$ analogously to (14),

$$\delta_{FE}^W \equiv \left( \sum_{t=1}^T E[\ddot{W}_{it} \ddot{W}_{it}'] \right)^{-1} \left( \sum_{t=1}^T E[\ddot{W}_{it} \ddot{D}_{it}] \right)$$

(15)

represents the population coefficient from a TWFE regression of $D_{it}$ on $W_{it}$, and

$$R_{it} \equiv \ddot{D}_{it} - \ddot{W}_{it}' \delta_{FE}^W$$

(16)

represents a residual from the regression.

Using these definitions, the population version of $\hat{\beta}_{FE}^D$ is given by

$$\beta_{FE}^D \equiv \frac{\sum_{t=1}^T E[Y_{it} R_{it}]}{\sum_{t=1}^T E[D_{it} R_{it}]}.$$  

(17)

\footnote{In this definition and hereafter, I implicitly assume that the first and second moments of $(Y_{it}, D_{it}, W_{it})_{t=1}^T$ and $C_i$ exist.}
To provide a causal interpretation of $\beta_{FE}^D$, I make the following assumptions throughout Section 3.

**Assumption PO.** (Potential Outcome) For each $t = 1, \ldots, T$, \{\(Y_{it}(d) : d \in (\underline{d}, \overline{d})\}\} is a stochastic process that defines a potential outcome associated with each possible treatment level $d \in (\underline{d}, \overline{d})$, where $-\infty \leq \underline{d} < \overline{d} \leq \infty$. The observed outcome is given by $Y_{it} = Y_{it}(D_{it})$.

**Assumption V.** (Variation) $\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[\text{Var}(\Delta_k D_{it}|\Delta_k W_{it}, C_i)] > 0$.

**Assumption L.** (Linearity) For any $(k, t)$ with $1 \leq t < t+k \leq T$, the conditional expectation of $\Delta_k D_{it}$ given $(\Delta_k W_{it}, C_i)$ can be expressed as

$$E[\Delta_k D_{it}|\Delta_k W_{it}, C_i] = \Delta_k W_{it}' \delta^W_k + C_i' \delta^C_k.$$

Assumption PO rules out dynamic treatment effects by restricting a potential outcome to be a function only of the current treatment status. Dynamic effects are typically allowed in a staggered adoption design setting with a binary treatment (see, e.g., Callaway and Sant’Anna, 2020; Goodman-Bacon, 2021). In a more general setting considered here, Assumption PO is necessary to avoid making a potential outcome an extremely high-dimensional object. Assumption V requires that $\Delta_k D_{it}$ have some variation conditional on $(\Delta_k W_{it}, C_i)$. Assumption L requires that the conditional mean $E[\Delta_k D_{it}|\Delta_k W_{it}, C_i]$ be linear in $(\Delta_k W_{it}, C_i)$. Similar linearity assumptions have been commonly used in interpreting linear regression coefficients in the presence of covariates. This assumption simplifies the analysis by ruling out omitted variable bias associated with unaccounted nonlinear effects of covariates.

It will turn out that Assumption L is not sufficient for enabling a causal interpretation of $\beta_{FE}^D$ without any omitted variable bias. Thus, I make the following assumption to eliminate the bias.

**Assumption LH.** (Linearity and Time Homogeneity) For any $(k, t)$ with $1 \leq t < t+k \leq T$, the conditional expectation of $\Delta_k D_{it}$ given $(\Delta_k W_{it}, C_i)$ can be expressed as

$$E[\Delta_k D_{it}|\Delta_k W_{it}, C_i] = \Delta_k W_{it}' \delta^W_k + C_i' \delta^C_k,$$

where the coefficient $\delta^W$ does not vary across $(k, t)$.

Assumption LH adds an additional condition to Assumption L requiring that the relationship between $\Delta_k D_{it}$ and $\Delta_k W_{it}$ be homogeneous across periods. Assumption L can

---

11See, for example, Angrist and Krueger (1999), Lochner and Moretti (2015), Słoczyński (2020b), Ishimaru (2022), and Goldsmith-Pinkham et al. (2022).
be made as nonrestrictive as possible by choosing a flexible specification of covariates (e.g., adding polynomial and interaction terms). In contrast, Assumption LH puts a fundamental restriction on the relationship between treatment changes and covariate changes, which cannot be eased by a flexible specification within a paradigm of the equation (8).

Assumption LH is closely related to an imperfect weighted-average formula in Theorem 3. The following lemma clarifies the relationship.

Lemma 1. Under Assumption LH, \( \delta^W_{FE} = \delta^W = \delta^W_{FD,k} \) for all \( k = 1, \ldots, T - 1 \), where \( \delta^W_{FE} \) is as defined in (15) and \( \delta^W_{FD,k} \) is a coefficient from a regression of \( \Delta_k D_{it} - L_{Ci} (\Delta_k D_{it}) \) on \( \Delta_k W_{it} - L_{Ci} (\Delta_k W_{it}) \).

Lemma 1 suggests that Assumption LH is a sufficient condition for \( \hat{\delta}^W_{FE} \) and \( \hat{\delta}^W_{FD,k} \) to be equivalent in the population. Therefore, under Assumption LH the bias term \( \hat{B} \) in Theorem 3 vanishes in the population.

The remaining, and the most important assumption for the causal interpretation is a common trends assumption for potential outcome changes, \( \Delta_k Y_{it}(d) \). Typically, this assumption is made only for \( d = 0 \) in the literature, which focuses on a binary treatment setting. However, such a natural baseline treatment level may not exist in a more general setting. For example, if a treatment variable is the minimum wage, it would be unnatural to assume common trends only at a particular minimum wage level. Therefore, I consider two versions of the causal interpretation with different common trends assumption. In Section 3.2 I make a common trends assumption only for a particular baseline treatment level, following a typical assumption in the literature. In Section 3.3 I make a common trends assumption for every treatment level, which can be more appropriate when a treatment variable does not have a natural baseline level.

3.2 Causal Interpretation With a Baseline Treatment Level

I make the following common trends assumption, presuming that a natural baseline treatment level exists, such as “no treatment” in a binary treatment setting.

Assumption CT. (Conditional Common Trends) There exists \( d_0 \in (d, \overline{d}) \) such that

\[
E [\Delta_k Y_{it}(d_0) | \Delta_k D_{it}, \Delta_k W_{it}, C_i] = E [\Delta_k Y_{it}(d_0) | \Delta_k W_{it}, C_i]
\]

for any \( (k, t) \) with \( 1 \leq t < t + k \leq T \).

Assumption CT requires that trends in potential outcomes \( \Delta_k Y_{it}(d_0) \) at a baseline treatment level \( d_0 \) be mean independent from treatment changes \( \Delta_k D_{it} \), conditional on time-
invariant covariates $C_i$ and the concurrent changes $\Delta_k W_{it}$ in time-varying covariates. Combining Assumption [CT] with the other assumptions yields the following result.

**Theorem 4.** Under Assumptions [PO] [V] [L] and [CT],

$$
\beta_{DFE}^{D} = \sum_{t=1}^{T} E[\tau_{it}\omega_{it}] + B,
$$

where

$$
\tau_{it} \equiv \begin{cases} 
\frac{Y_{it}(D_{it}) - Y_{it}(d_{0})}{D_{it} - d_{0}} & D_{it} \neq d_{0} \\
0 & D_{it} = d_{0} 
\end{cases}
$$

is a per-unit effect of a deviation of the treatment $D_{it}$ from the baseline level $d_{0}$. The weights are defined as

$$
\omega_{it} \equiv \frac{(D_{it} - d_{0}) R_{it}}{\sum_{t'=1}^{T} E[D_{it'} R_{it'}]},
$$

which satisfy $\sum_{t=1}^{T} E[\omega_{it}] = 1$. The bias term $B$ is given by

$$
B = \frac{\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[\Delta_k Y_{it}(d_{0}) (\Delta_k W_{it} - L_{C_i} (\Delta_k W_{it})) (\delta_{W}^{k} - \delta_{W}^{FE})]}{\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E[\Delta_k D_{it} \Delta_k R_{it}]}.
$$

With Assumption [LH] added, $B = 0$ and the weights satisfy

$$
\omega_{it} \propto (D_{it} - d_{0}) \sum_{s=1}^{T} (D_{it} - D_{is} - E[D_{it} - D_{is}|W_{it} - W_{is}, C_i]) .
$$

(18)

Theorem 4 demonstrates that the TWFE coefficient $\beta_{DFE}^{D}$ can be interpreted as a weighted average of per-unit effects of the treatment under a certain set of assumptions, without relying on a linear parametric model such as (8). The results on the bias term $B$ indicates that Assumption [LH] is necessary for the exact weighted-average interpretation. The weight function $\omega_{it}$ interacts a deviation of $D_{it}$ from the baseline level $d_{0}$ with a deviation of $D_{it}$ from the TWFE predicted value. As seen in (18), the weight can also be interpreted as an interaction between $D_{it} - d_{0}$ and the sum of treatment changes starting from or ending at a period $t$ that cannot be predicted by covariates.

Theorem 4 is in line with an optimistic view about linear regressions by Angrist and Pischke (2010), who suggest that regression estimators produce average effects even without assuming a linear model (see the quote in the footnote 1). However, Theorem 4 also provides a cautionary view on the TWFE estimator by highlighting the following two potential problems.
The first problem is the necessity of several strong assumptions for the causal interpretation. In Section 4, I discuss how these assumptions may fail in practice and propose a generalized TWFE estimator that can have a valid causal interpretation under a more plausible set of assumptions.

The second problem, as similarly highlighted in a binary treatment case (see, e.g., Borusyak et al., 2021; de Chaisemartin and D’Haultfoeuille, 2020a; Goodman-Bacon, 2021), is that the weights \( \omega_{it} \) can be negative. If some weights are negative and weights are strongly correlated with per-unit treatment effects, then \( \beta_{FE}^D \) can even lie outside the support of per-unit treatment effects \( \tau_{it} \). While this problem is suggested to be severe in a binary treatment setting with a staggered adoption design, how serious the problem is would depend on how systematically treatment effects are related with weights. Section 3.3 shows that the weight function \( \omega_{it} \) contains redundant variation that is orthogonal to treatment effects under a more strict common trends assumption, which may alleviate the concern about negative weights in some cases.

Remark

de Chaisemartin and D’Haultfoeuille (2020a, Appendix 3.3) provide a weighted-average interpretation of a TWFE coefficient that resembles Theorem 4. While they consider a binary treatment case with time-varying covariates, it is straightforward to extend their proof to allow for a non-binary treatment and time-invariant covariates. They assume

\[
E [\Delta_k Y_{it}(d_0)|D_{i1}, \ldots, D_{iT}, W_{i1}, \ldots, W_{iT}] = E [\Delta_k Y_{it}(d_0)|\Delta_k W_{it}], \tag{19}
\]

which is a sufficient but not necessary condition for Assumption CT (ignoring time-invariant covariates \( C_i \))\(^{12}\). Given the existing result in de Chaisemartin and D’Haultfoeuille (2020a), primary contribution of Theorem 4 is that it shows Assumption CT which builds only on changes in variables of interest, is a fundamental requirement for a causal interpretation of the TWFE coefficient. Theorem 4 also serves as a stepping stone toward the analysis in Section 3.3, which has not been considered in de Chaisemartin and D’Haultfoeuille (2020a) or other studies in the literature.

\(^{12}\) Another difference is that they assume linearity and time homogeneity of \( E [\Delta_k Y_{it}(d_0)|\Delta_k W_{it}] \) instead of Assumption LH. The bias term \( B \) in Theorem 4 disappears under this assumption, too.
3.3 Causal Interpretation Without a Baseline Treatment Level

Even though a natural baseline treatment level such as “no treatment” exists in some applications, in other applications making Assumption CT only for a particular treatment level \(d_0\) can be too arbitrary. I use the following assumptions to consider these cases.

**Assumption CTE.** (Conditional Common Trends at Any Level of Treatment) For any \(d \in (d_d, d_0)\) and any \((k, t)\) with \(1 \leq t < t + k \leq T\),

\[
E \left[ \Delta_k Y_{it}(d) | \Delta_k D_{it}, \Delta_k W_{it}, C_i \right] = E \left[ \Delta_k Y_{it}(d) | \Delta_k W_{it}, C_i \right].
\]

**Assumption LC.** (Lipschitz Continuity) For each \(t = 1, \ldots, T\), there exists a random variable \(M_{it}\) that satisfies \(|Y_{it}(d') - Y_{it}(d)| \leq M_{it}|d' - d|\) for any \(d', d \in (d_d, d_0)\), \(E[M_{it}^2] < \infty\), and \(E[M_{it}^2 D_{it}^2] < \infty\).

Assumption CTE extends Assumption CT to every possible treatment level. Even though Assumption CTE is technically stronger than Assumption CT, it can be practically more reasonable since it does not require an arbitrary baseline treatment level. Assumption LC is a regularity condition which ensures that a potential outcome process \(Y_{it}(d)\) is differentiable almost everywhere and is sufficiently smooth.

I maintain Assumptions PO, V, and LH, which are also used in Section 3.2. I focus on a setting under Assumption LH for simplicity; the bias term \(B\) does not differ from Theorem 4 when Assumption LH is replaced by Assumption L.

It turns out that these assumptions are sufficient for providing a weighted-average interpretation of a TWFE coefficient, but not sufficient for an unambiguous weighted-average interpretation.

**Theorem 5.** Suppose Assumptions PO, V, LH, CTE and LC hold. Then, for any \(d_0 \in (d_d, d_0)\) and any \(g : (d_d, d_0) \mapsto \mathbb{R}\) with \(\int_d^{d_0} |g(x)|dx < \infty\),

\[
\beta_{FE}^D = \int_d^{d_0} \sum_{t=1}^T E \left[ Y_{it}'(x) (\psi_{it}(x; d_0) + g(x)R_{it}) \right] dx,
\]

where the weight function

\[
\psi_{it}(d; d_0) \equiv \frac{1_{D_{it} \geq d} - 1_{d_0 \geq d}}{\sum_{t'=1}^T E[D_{it'}R_{it'}]} R_{it}
\]

satisfy \(\int_d^{d_0} \sum_{t=1}^T E \left[ \psi_{it}(x; d_0) \right] dx = 1\).
Theorem 5 expresses the TWFE coefficient $\beta_{PE}^D$ as a weighted-average of marginal effects $Y'_{it}(d)$. However, there are infinitely many weight functions that can yield this weighted-average expression. Multiplicity of the weight function arises because a common trends assumption essentially requires that a potential outcome $Y_{it}(d)$ be additive in unit and period effects. Due to the assumption, $Y_{it}(d)$ is orthogonal to a TWFE residual $R_{it}$, that is, $\sum_{t=1}^{T} E[Y_{it}(d)R_{it}] = 0$ for any $d \in (d, \bar{d})$. As a result, any function proportional to $R_{it}$ can be added to the weight function without affecting the weighted average.

To provide an unambiguous weighted-average interpretation, it is necessary to eliminate the weight variation orthogonal to $Y_{it}(d)$ by fully utilizing the identification conditions. For this purpose, I make additional conditions that characterizes the exogeneity requirements more explicitly than Assumption CTE.

**Assumption SE.** (Strict Exogeneity) For any $d \in (d, \bar{d})$ and $(k, t)$ with $1 \leq t < t + k \leq T$,

$$E[\Delta_k Y_{it}(d) | D_{i,t+k}, D_{it}, \overline{D}_i, W_{i,t+k}, W_{it}, \overline{W}_i, C_i] = E[\Delta_k Y_{it}(d) | W_{i,t+k}, W_{it}, \overline{W}_i, C_i].$$

**Assumption S.** (Sufficiency) For any $d \in (d, \bar{d})$ and $(k, t)$ with $1 \leq t < t + k \leq T$,

$$E[\Delta_k Y_{it}(d) | W_{i,t+k}, W_{it}, \overline{W}_i, C_i] = E[\Delta_k Y_{it}(d) | \Delta_k W_{it}, C_i].$$

Assumption SE is a nonparametric version of a strict exogeneity condition that is common in the panel data models (see, e.g., Arellano and Honoré, 2001). A typical strict exogeneity condition requires an entire treatment series $(D_{i1}, \ldots, D_{iT})$ to be exogenous. It turns out that only the time-series average $\overline{D}_i$ has be exogenous in addition to $(D_{i,t+k}, D_{it})$ in this setting, although the exogeneity of the time-series average may virtually require the exogeneity of an entire treatment series in practice. Essentially, Assumption SE rules out the possibility that the potential outcome changes $\Delta_k Y_{it}(d)$ influence future treatment assignments or is correlated with past treatment assignments. While Assumption CTE implicitly makes such a requirement as well, Assumption SE does so more explicitly.

Assumption S requires that using the concurrent covariate changes $\Delta_k W_{it}$ is sufficient for predicting the potential outcome changes $\Delta_k Y_{it}(d)$. This assumption is necessary because Assumption SE alone does not imply Assumption CTE. Assumption SE allows the potential outcome changes $\Delta_k Y_{it}(d)$ to depend on $(W_{i,t+k}, W_{it}, \overline{W}_i)$, while a TWFE regression tries

---

13Even though Assumption CTE individually requires nothing more than zero correlation $\Delta_k Y_{it}(d)$ between $\Delta_k D_{it}$ for each $(k, t)$, it collectively imposes a stronger exogeneity condition for $(D_{i1}, \ldots, D_{iT})$. For example, if the outcome change $Y_{i2}(d) - Y_{i1}(d)$ influences the future treatment change $D_{i3} - D_{i2}$, then it is likely that $Y_{i3}(d) - Y_{i1}(d)$ and $D_{i3} - D_{i1}$ are correlated even if there is not any correlation between $Y_{i2}(d) - Y_{i1}(d)$ and $D_{i2} - D_{i1}$ or between $Y_{i3}(d) - Y_{i2}(d)$ and $D_{i3} - D_{i2}$.
to predict the outcome changes only through $\Delta_k W_{it}$. Assumption $S$ rules out any omitted variable bias that may arise from this discrepancy. Assumptions $SE$ and $S$ jointly imply Assumption $CTE$. A combination of Assumptions $SE$ and $S$ is closely related to the condition (19) imposed by de Chaisemartin and D'Haultfoeuille (2020a), except that Assumptions $SE$ and $S$ are required for every possible treatment level.

**Theorem 6.** Under Assumptions $PO$, $V$, $LH$, $LC$, $SE$, and $S$,

$$
\beta_{FE}^D = \int_\mathcal{d} \sum_{t=1}^T E \left[ Y_{it}'(x) \psi_{it}^*(x) \right] dx,
$$

where the weight function is given by

$$
\psi_{it}^*(d) = \frac{1}{T} \sum_{s=1}^T \mathbb{1}_{D_{is} \geq d} R_{is} + \frac{1}{T} \sum_{s=1}^T E \left[ \mathbb{1}_{D_{it} \geq d} R_{it} - \mathbb{1}_{D_{is} \geq d} R_{is} \right] W_{it} - W_{is}, C_i \right] \sum_{t'=1}^T E [D_{it'} R_{it'}]
$$

and satisfies $\int_\mathcal{d} \sum_{t=1}^T E \left[ \psi_{it}^*(x) \right] dx = 1$. In addition,

$$
\int_\mathcal{d} \sum_{t=1}^T E \left[ \psi_{it}^*(x)^2 \right] dx \leq \int_\mathcal{d} \sum_{t=1}^T E \left[ (\psi_{it}^*(x) + g(x) R_{it})^2 \right] dx
$$

holds for any $g : (d, \overline{d}) \mapsto \mathbb{R}$ with $\int_\mathcal{d} |g(x)| dx < \infty$. $\psi_{it}^*(d)$ can also be expressed as

$$
\psi_{it}^*(d) = \frac{1}{T} \sum_{s=1}^T \psi_{is}(d; d_0) + \frac{1}{T} \sum_{s=1}^T E \left[ \psi_{is}(d; d_0) - \psi_{is}(d; d_0) \right] W_{it} - W_{is}, C_i \right],
$$

where $\psi_{it}(d; d_0)$ is the weight function defined in Theorem 5 for any $d_0 \in (d, \overline{d})$.

Theorem 6 uses a weight function $\psi_{it}^*(d)$ that does not depend on an arbitrary chosen $d_0$, unlike in Theorem 5. While any function proportional to $R_{it}$ can still be added to $\psi_{it}^*(d)$ without affecting the weighted average, doing so always results in a larger square norm of the weight. In that sense, $\psi_{it}^*(d)$ can be considered as the most appropriate weight function.

As seen in (20), the weight function $\psi_{it}^*(d)$ is additive in two components, where the first component is constant within the same unit $i$ and the second component is constant within the same period $t$ conditional on covariates. Moreover, as expressed in (22), $\psi_{it}^*(d)$ can be obtained by summarizing any possible weight function $\psi_{it}(d; d_0)$ into the unit-specific and the period-specific terms in exactly the same manner.

To offer an intuition for the weight function, consider the simplest case with no covariates.
In this case, the weight function is given by

$$\psi_{it}^*(d) = \frac{\frac{1}{T} \sum_{s=1}^{T} \left( \mathbb{1}_{D_{is} \geq d} R_{is} - E \left[ \mathbb{1}_{D_{is} \geq d} R_{is} \right] \right) + E \left[ \mathbb{1}_{D_{it} \geq d} R_{it} \right]}{\sum_{t'=1}^{T} E \left[ D_{it'} R_{it} \right]},$$

which is a predicted value from a regression of $\psi_{it}(d; d_0)$ on additive unit and period fixed effects. Since a common trends assumption requires a potential outcome $Y_{it}(d)$ to be additive in unit and period effects, any weight variation orthogonal to unit and period effects is orthogonal to $Y_{it}(d)$. Projecting the weight function onto unit and period effects eliminates this redundant weight variation.

While the set of assumptions in Theorem 4 holds in Theorem 6 too, using the weight function $\omega_{it}$ in Theorem 4 can be misleading for defining the weight on each observation. In particular, the total weight on an observation $(i, t)$ aggregated over $d \in (d, \bar{d})$ in Theorem 6 is given by

$$\int_{d}^{\bar{d}} \psi_{it}^*(x) dx = \frac{1}{T} \sum_{s=1}^{T} D_{is} R_{is} + \frac{1}{T} \sum_{s=1}^{T} E \left[ D_{it} R_{it} - D_{is} R_{is} \right] W_{it} - W_{is}, C_i] \sum_{t'=1}^{T} E \left[ D_{it'} R_{it} \right]$$

$$= \frac{1}{T} \sum_{s=1}^{T} \omega_{is} + \frac{1}{T} \sum_{s=1}^{T} E \left[ \omega_{it} - \omega_{is} \right] W_{it} - W_{is}, C_i].$$

Redundant variation in $\omega_{it}$ is eliminated exactly in the same manner as in (22).

4 Generalizing TWFE Regressions

4.1 Problems of the Standard Approach

Econometric results in Sections 2 and 3 highlight three potential problems of TWFE Regressions. First, empirical researchers may not always want to exploit all possible $k$-period differences in estimating the effect of their interest, despite that a TWFE regression aggregates all of them into the coefficient. In some cases in which the effect materializes only in the long run due to frictional adjustment, they may not be interested in “short” differences. In other cases in which Assumption CT is not plausible for large $k$, they may not want to exploit “long” differences. Assumption CT requires that the change in potential outcomes, $\Delta_k Y_{it}(d_0)$, is uncorrelated with the change in treatment status, $\Delta_k D_{it}$, between periods $t$ and $t + k$. With larger $k$, it may become more likely that a difference in unobservables causes long-run changes in treatment levels and potential outcome levels at the same time. In ad-

14See Ishimaru (2022, Section 3.2) for a discussion in a related context.
dition, a reverse causality problem may arise, where, for example, the change in treatment status between periods $t + k/2$ and $t + k$ may occur in response to the change in potential outcomes between periods $t$ and $t + k/2$.\footnote{One of many well-known examples in which treatment assignments are influenced by past outcomes is Ashenfelter’s dip in the job-training literature (Ashenfelter, 1978).}

The second potential problem is that the role of time-varying covariates in a TWFE regression differs entirely from a two-period DID setting that allows for covariates. When a TWFE regression includes time-varying covariates, a common trends assumption has to be made conditional on changes in covariates, not their initial levels, as illustrated in Section 3. In contrast, a common trends assumption is typically made conditional on pre-treatment observables in a two-period DID setting (Heckman et al., 1997, 1998; Abadie, 2005).\footnote{As a seminal example of this identification strategy, Card and Krueger (1994) implement a two-period DID as a FD regression and control for time-invariant observables in some specifications.}

As pointed out by Caetano et al. (2022), the former type of common trends assumption is less theoretically favored than the latter one. In particular, if the treatment change $\Delta k D_{it}$ influences the covariate change $\Delta k W_{it}$, then variation in $\Delta k D_{it}$ with $\Delta k W_{it}$ being fixed does not correspond to \textit{ceteris paribus} variation that defines the causal effect of the treatment.\footnote{Angrist and Pischke (2008, Section 3.2.3) characterize this type of problems as “bad control” problems, suggesting that occupation is a \textit{“bad control”} when a researcher is interested in the effects of a college degree on earnings.}

Thirdly and finally, a TWFE regression accounts for the effects of time-varying covariates less flexibly than FD regressions, even if a conditional common trends assumption can be justified. This problem results in an imperfect weighted-average interpretation both in a finite sample and in the population, as illustrated in Theorems 3 and 4.

### 4.2 An Alternative Approach

I propose a generalized TWFE regression that overcomes these limitations of the standard TWFE regression. Modifying the pooled FD version of the least-squares objective (3), I define the following least-squares problem as a generalized TWFE regression.

\[
\min_{\beta^D, \{\beta^W_k, \beta^V_k\}_{k=K}^K} \sum_{k=K}^K \sum_{i=1}^N \sum_{t=1}^{T-k} (\Delta_k Y_{it} - \beta^D \Delta_k D_{it} - \Delta_k W'_{it} \beta^W_k - V'_{it} \beta^V_k - C'_{i, \gamma^*_{k,t}})^2. \tag{23}
\]

This objective function naturally extends (3), increasing flexibility in four ways. First, possible between-period gaps are restricted to $K \leq k \leq K$, where the lower bound $K \geq 1$ and the upper bound $K \leq T - 1$ can be chosen by a researcher. Second, an additional
covariate vector $V_{it}$ enters a regression. $V_{it}$ should consist of variables observed until period $t$ that are expected to make the conditional common trends assumption more plausible. $V_{it}$ may include some elements of $W_{it}$, past outcomes ($Y_{it}, Y_{i,t-1}, \ldots$) or past treatment status ($D_{it}, D_{i,t-1}, \ldots$). This specification brings a regression closer to a two-period DID that controls for pre-treatment observables. Third, covariate changes $\Delta_k W_{it}$ can be still controlled for when necessary, and the specification allows for heterogeneous coefficients on $\Delta_k W_{it}$ across $k$. Finally, the additive separability restriction $\gamma^*_{k,t} = \gamma_{t+k} - \gamma_t$ on the coefficient on $C_i$ is removed. While this restriction is not binding in (3), it is necessary to remove this restriction in (23) to enable natural numerical and causal interpretations when coefficients on $\Delta_k W_{it}$ and $V_{it}$ are allow to differ across $k$.

An advantage of this generalized approach is that it does not depart significantly from the standard TWFE regression while allowing for greater flexibility. It can be easily implemented and be compared with results from the standard TWFE regression just as one of alternative specifications. It has a valid causal interpretation under a conditional common trends assumption and other auxiliary assumptions, which can be made less restrictive than those required for the standard TWFE regression.

I make the following assumptions for providing the causal interpretation of the population coefficient from the regression (23), modifying Assumptions CT, V, and LH. The following discussion focuses on the setting with a baseline treatment level as in Section 3.2.

**Assumption CT’.** (Conditional Common Trends) There exists $d_0 \in (d, \overline{d})$ such that

$$E[\Delta_k Y_{it}(d_0)|\Delta_k D_{it}, \Delta_k W_{it}, V_{it}, C_i] = E[\Delta_k Y_{it}(d_0)|\Delta_k W_{it}, V_{it}, C_i]$$

for any $(k, t)$ with $K \leq k \leq \overline{K}$ and $1 \leq t < t + k \leq T$.

**Assumption V’.** (Variation) $\sum_{k=K}^{\overline{K}} \sum_{t=1}^{T-k} E[Var(\Delta_k D_{it}|\Delta_k W_{it}, V_{it}, C_i)] > 0$.

**Assumption LH’.** (Linearity and Time Homogeneity) For any $(k, t)$ with $K \leq k \leq \overline{K}$ and $1 \leq t < t + k \leq T$,

$$E[\Delta_k D_{it}|\Delta_k W_{it}, V_{it}, C_i] = \Delta_k W_{it}'\delta^W_k + V_{it}'\delta^V_k + C_i\delta^C_k,$$

where the coefficients $\delta^W_k$ and $\delta^V_k$ do not vary across $t$.

Unlike in Assumptions CT, V, and LH, identification conditions are now conditional on the initial covariate levels, $V_{it}$, as well as the concurrent covariate changes, $\Delta_k W_{it}$. Since the coefficients $(\beta^V_k, \beta^W_k)$ of these covariates can depend on $k$ in (23), the coefficients $(\delta^V_k, \delta^W_k)$ in

\footnote{Note that using lagged outcomes as covariates generally requires a caution (Chabé-Ferret, 2015, 2017).}
Assumption LH can now depend on \( k \) unlike in Assumption LH. This assumption can be further relaxed by allowing the coefficients \((\beta^V, \beta^W)\) in (23) to depend on \( t \) as well.\(^{19}\)

Using the FWL theorem, the population coefficient on \( \Delta_k D_{it} \) from the regression in (23) is

\[
\beta_{GFE}^D = \frac{\sum_{k=K}^K \sum_{t=1}^{T-k} E \left[ r_{it}^{(k)} \cdot \Delta_k Y_{it} \right]}{\sum_{k=K}^K \sum_{t'=0}^{T-k} E \left[ r_{it}^{(k)} \cdot \Delta_k D_{it} \right]},
\]

(24)

where \( r_{it}^{(k)} \) is the population residual from a regression of \( \Delta_k D_{it} \) on \( \Delta_k W_{it}, V_{it}, \) and \( C_i \) independently performed for \( k = K, \ldots, K \). The following theorem presents the causal interpretation of \( \beta_{GFE}^D \).

**Theorem 7.** Under Assumptions PO, CT\(^{\prime} \), V\(^{\prime} \), and LH\(^{\prime} \),

\[
\beta_{GFE}^D = \sum_{t=1}^T E \left[ \tau_{it} \omega_{it}^G \right],
\]

where \( \tau_{it} \) is a per-unit effect defined in Theorem 4 and the weights are defined as

\[
\omega_{it}^G \equiv \frac{(D_{it} - d_0) \left( \sum_{k=K}^K r_{it-k}^{(k)} 1_{t-k \geq 1} - \sum_{k=K}^K r_{it}^{(k)} 1_{t-k \leq T} \right)}{\sum_{t'=1}^T E \left[ (D_{it'} - d_0) \left( \sum_{k=K}^K r_{it'-k}^{(k)} 1_{t'-k \geq 1} - \sum_{k=K}^K r_{it'}^{(k)} 1_{t'+k \leq T} \right) \right]},
\]

In addition, the weights satisfy

\[
\omega_{it}^G \propto (D_{it} - x_0) \sum_{s=1}^T 1_{K \leq t-s \leq K} \left( D_{it} - D_{is} - E \left[ D_{it} - D_{is} | V_{i,\min\{s,t\}}, W_{it} - W_{is}, C_i \right] \right)
\]

(25)

and \( \sum_{t=1}^T E \left[ \omega_{it}^G \right] = 1 \).

Theorem 7 suggests that the generalized TWFE coefficient \( \beta_{GFE}^D \) can be interpreted as a weighted average of per-unit treatment effects, just like Theorem 4 providing a weighted-average interpretation of the standard TWFE coefficient. The difference from Theorem 4 is that the identification conditions are now required for only a subset of \( k \) and are conditional on the initial levels of covariates as well as their changes.

The weight function \( \omega_{it} \) in Theorem 4 and the weight function \( \omega_{it}^G \) in Theorem 7 are comparable using (18) and (25). In (18), the weight function \( \omega_{it} \) interacts \( D_{it} - d_0 \) with the

\(^{19}\)Such a too flexible specification may require caution in practice, as it may cause a finite sample problem by reducing degree of freedom of a regression.
sum of treatment changes since or until a period $t$ that are not forecastable by covariate changes. In (25), the weight function $\omega_{it}^G$ consists of similar terms, except that treatment changes are restricted to the window of $K$ to $K$ periods and the forecast is based on initial levels of covariates as well as their changes.

In terms of comparability with the standard approach, similarity of the weight function is a nice feature. At the same time, the weight function $\omega_{it}^G$ may suffer from the problem of negative weights just like the standard approach. Several recent approaches (Callaway and Sant’Anna, 2020; de Chaisemartin and D’Haultfoeuille, 2020a; Goodman-Bacon, 2021, among others) are designed to solve the negative weights problem, but these approaches do not apply well to a general setting considered here, which allows for nonbinary treatment as well as both time-invariant and time-varying covariates. A solution to this problem is left for future work.

Appendix C discusses two additional features of the generalized TWFE regression. First, at the sample level, its coefficient on $\Delta_k D_{it}$ can be expressed as a convex combination of FD coefficients. It restores the exact weighted-average interpretation that fails for the standard TWFE coefficient in the presence of time-varying covariates as seen in Theorem 3. Second, while the causal interpretation of the generalized TWFE coefficient does not require a linear model as demonstrated in Theorem 7, a linear regression equation that directly corresponds to the least-squares problem (23) can arise from assumptions of constant causal effects, conditional common trends, and linearity. The derivation of the linear equation closely follows an argument in Angrist and Pischke (2017), which produces a linear regression equation for cross-sectional data from a design-based perspective assuming constant causal effects, conditional independence, and linearity.

5 Empirical Illustration

This section illustrates the econometric results and the proposed estimator by studying the impact of minimum wages on employment using a TWFE regression. The TWFE regression analysis based on state-level panel data in the United States has attracted much attention for decades. The econometric framework of this paper, which works for any multiperiod panel, is suitable for exploring this empirical setting. The treatment variable, the state minimum wages, is provided an overview of the state of the literature. See, for example, Neumark et al. (2008) for a comprehensive review. I do not attempt to offer the “best evidence” for this contentious topic through my empirical analysis. However, I demonstrate that my econometric results provide a foundation for future empirical research and debate on this topic by clarifying how a TWFE regression maps the data into the coefficient.
wage, is continuous and changes many times in a given state. The existing insights from the TWFE regression literature that focuses on a binary treatment and a staggered adoption design cannot be directly applied to this setting.

To present the estimates, I use a state–year panel of employment outcomes and minimum wage laws in 50 U.S. states and D.C. in 1990–2018. As a dependent variable in a TWFE regression, I use the log employment rate of teens (age 16–19) from the Current Population Survey (CPS). Another dependent variable is the net job creation rate from the Business Dynamics Statistics (BDS), using four major sectors with high shares of workers without college education.21 The focus on age groups or sectors more likely affected by the minimum wage is practical, since the majority of workers in the United States are not directly affected by the minimum wage. The minimum wage data is from Vaghul and Zipperer (2019), and I use the log of the annual average of the state minimum wage as the treatment variable in each regression.

I start by illustrating the decomposition suggested in Section 2.2. For a TWFE regression of the log teen employment rate on the log minimum wage, Figure 1 presents the FD coefficients \( \hat{\beta}_{FD,k} \) as dots, the associated weights as bars, and the TWFE coefficient \( \hat{\beta}_{FE} \) as a horizontal dotted line. The TWFE coefficient is \( \hat{\beta}_{FE} = -0.133 \). The FD coefficient is \( \hat{\beta}_{FD,1} = -0.025 \) with a one-year gap and decreases as the gap gets longer.22 A TWFE regression aggregates these heterogeneous FD coefficients into a single coefficient. The weighted average of all 28 FD coefficients is indeed \(-0.133\). Figure 2 presents analogous plots for a regression of the net job creation rate. Even though the TWFE coefficient is \( \hat{\beta}_{FD} = -0.345 \), short-run and long-run FD coefficients have different signs and the order of magnitude of the FD coefficients are much larger than that of the TWFE coefficient.

As illustrated in Figures 1 and 2, a TWFE regression aggregates potentially heterogeneous FD coefficients into a single TWFE coefficient. If the model (1) is “true” in a textbook sense (i.e., (1) describing the causal relationship and \( E[\varepsilon_{it}|X_{i1}, \ldots, X_{iT}] = 0 \)), the FD coefficients should be all identical in the population. Providing a coherent interpretation of the results or finding a better alternative specification is unlikely to succeed within a paradigm of the linear model, since the linear model is not plausible in this empirical example in the first place. The required set of assumptions for the causal interpretation suggested in Section 3 provides a foundation for an objective discussion of the regression results without relying on

---

21 The four sectors are: Construction, Manufacturing, Retail Trade, and Accommodation and Food Services. Using the data from all sectors does not qualitatively change the estimates presented below, but the estimated effects become smaller in magnitude.

22 The difference among “short” FD, “long” FD, and TWFE estimates itself is not a novel empirical finding. It has been long known in the minimum wage effects literature (see, e.g., Neumark and Wascher, 1992), even though an explicit numerical relationship among these estimates has not been known.
the linear model.

There are two possible reasons why short-run and long-run FD coefficients differ for a regression of the log teen employment rate. First, the employment rate, being a stock variable, may not quickly respond to the minimum wage change, as argued by Baker et al. (1999), Sorkin (2015), and Meer and West (2016), among many others. Second, Assumption CT may not hold in the long run. For example, there can be a confounding factor that gives rise to both minimum wage growth and teen employment rate decline in a given state. The standard TWFE coefficient, being a weighted average of all FD coefficients, would be difficult to interpret due to the long-run endogeneity bias.

For a regression of the net job creation rate, short-run FD coefficients are less likely to be influenced by a frictional response, since the dependent variable is now a flow variable. However, Assumption CT may not be plausible for long-run FD coefficients. In addition to a possibility of confounding factor as in the discussion above, a reverse causality is potentially an important issue. For example, an FD coefficient with a 20-year gap would find the positive “effect” even if a decline in the job creation rate in a given decade caused a slower minimum wage growth in the subsequent decade.

A generalized TWFE regression proposed in Section 4.2 provides a more flexible way of mapping the data into the coefficient than the standard TWFE regression. Table 1 presents generalized TWFE coefficients with various restrictions on the gap years. Columns (1) and (3) use no covariates. By construction, the first row in each column with no restriction on the gap years corresponds to the standard TWFE coefficient. The subsequent rows present the generalized TWFE coefficients with restrictions on between-period gaps. Given the discussions above, a generalized TWFE coefficient using suitable gap years may provide a more reasonable estimate of the minimum wage effect than the standard TWFE coefficient. For example, a generalized TWFE regression that uses only 1–5 year gaps suggests that a 10% increase in the minimum wage is associated with about a 0.4 percentage point decrease in the net job creation rate, an effect ten times larger than what is suggested by the standard TWFE regression.

One of the potential concerns of using a TWFE regression approach in estimating the minimum wage effect is that states with better economic prospects may increase their minimum wage levels faster. Intending to mitigate this concern, including state-specific linear time trends in the regression equation is a common practice in the minimum wage literature. However, Meer and West (2016) point out that such a specification may induce bias by fit-

---

23 This argument does not suggest that the short-run FD coefficients are “biased” in a statistical sense. What it suggests is the discrepancy between the effect these FD coefficients identify and the effect we are interested in, which is not a pure statistical problem.
ting post-treatment dynamics as well as pre-treatment trends. In addition, a specification with linear time trends has to assume that the underlying linear regression model is true. An objective discussion of a suitable functional form is extremely difficult.\footnote{See, for example, a disagreement on linear trends between Allegretto et al. (2011) and Neumark et al. (2014).}

I highlight another advantage of a generalized TWFE regression by addressing this potential concern. Using a generalized TWFE regression, I directly control for pre-treatment trends without being influenced by post-treatment dynamics. In particular, for each state $i$ and year $t$, I perform a regression of the employment outcomes $Y_{i,t-12}, \ldots, Y_{i,t-3}$ on calendar years $t-12, \ldots, t-3$, and use the predicted pre-trend as covariates in a generalized TWFE regression.\footnote{I use the data in 1978–1989, which are not used in the main analysis, to estimate state-specific pre-trends for earlier years in the panel. For example, the pre-trends for observations in 1990 are based on the data in 1978–1987. I also try shorter pre-trends as controls and confirm that estimates are not sensitive to the choice of pre-trend lengths.} I exclude immediately preceding two observations, $Y_{i,t-2}$ and $Y_{i,t-1}$, to avoid introducing bias associated with the mechanical correlation of covariates with the current outcome $Y_{it}$. The identification condition in this specification, as implied by Assumption $CT'$, is that the potential employment outcome changes and the minimum wage changes are uncorrelated conditional on pre-existing outcome trends. In addition to outcome pre-trends, I also include pre-trends of the log minimum wage between years $t-5$ and $t-1$, as well as the log minimum wage in year $t$. As a result, this regression counts on a comparison between states with similar pre-existing minimum wage levels, an apple-to-apple comparison that is not feasible for the standard TWFE regression.\footnote{Note that including lagged variables in the standard TWFE regression cannot do the same trick, since the identification condition has to rely on changes in these lags, not their levels.} Columns (2) and (4) of Table 1 present the regression results with these covariates. The estimated coefficients reveal that regression results are not sensitive to pre-existing dynamics of employment outcomes or minimum wage levels.

6 Conclusion

This paper revisits a long-overlooked numerical equivalence between TWFE and pooled FD regressions to uncover the process of how a TWFE regression maps the data into the coefficient of interest. A TWFE regression essentially captures the relationship among changes in variables of interest, pooling all possible short-run and long-run comparisons. Given this property, researchers using a TWFE regression should carefully investigate where the identifying variation of their coefficient of interest originates. As a diagnostic tool for this
For considering whether the standard TWFE estimator has a valid causal interpretation, my analysis highlights its two important limitations. First, it requires a common trends assumption for any between-period gap, including both short-run and long-run comparisons. Second, the common trends assumption has to be conditional on changes in time-varying covariates, not their levels. I propose a generalized TWFE estimator that overcomes these limitations while requiring only a small divergence from the standard approach. This new estimator can produce the coefficient of interest using specific between-period gaps, excluding, for example, “too short” or “too long” ones. In addition, a common trends assumption can be conditioned initial levels of time-varying covariates, like in a typical identification strategy for a two-period DID with covariates. In these ways, the new estimator allows researchers to use a set of identification conditions which they think is more plausible and to use the data variation which they think is more relevant.

References

Abadie, A. (2005): “Semiparametric difference-in-differences estimators,” *The Review of Economic Studies*, 72, 1–19.

Allegretto, S. A., A. Dube, and M. Reich (2011): “Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data,” *Industrial Relations: A Journal of Economy and Society*, 50, 205–240.

Angrist, J. D., and A. B. Krueger (1999): “Empirical strategies in labor economics,” *Handbook of Labor Economics*, 3, 1277–1366.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist’s Companion*: Princeton University Press.

——— (2010): “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics,” *Journal of Economic Perspectives*, 24, 3–30.

——— (2017): “Undergraduate econometrics instruction: through our classes, darkly,” *Journal of Economic Perspectives*, 31, 125–144.

Arellano, M., and B. Honoré (2001): “Panel data models: some recent developments,” in *Handbook of econometrics* Volume 5: Elsevier, 3229–3296.

Ashenfelter, O. (1978): “Estimating the effect of training programs on earnings,” *The Review of Economics and Statistics*, 47–57.
Athey, S., and G. W. Imbens (2022): “Design-based analysis in difference-in-differences settings with staggered adoption,” *Journal of Econometrics*, 226, 62–79.

Baker, M., D. Benjamin, and S. Stanger (1999): “The Highs and Lows of the Minimum Wage Effect: A Time-Series Cross-Section Study of the Canadian Law,” *Journal of Labor Economics*, 17, 318 – 350.

Borusyk, K., X. Jaravel, and J. Spiess (2021): “Revisiting event study designs: Robust and efficient estimation,” *Working Paper*.

Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues (2022): “Difference in Differences with Time-Varying Covariates,” *Working Paper*.

Callaway, B., A. Goodman-Bacon, and P. H. Sant’Anna (2021): “Difference-in-differences with a continuous treatment,” *Working Paper*.

Callaway, B., and P. H. Sant’Anna (2020): “Difference-in-differences with multiple time periods,” *Journal of Econometrics*.

Card, D., and A. B. Krueger (1994): “Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania,” *The American Economic Review*, 84, 772.

Chabé-Ferret, S. (2015): “Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes,” *Journal of Econometrics*, 185, 110–123.

—— (2017): “Should we combine difference in differences with conditioning on pre-treatment outcomes?” *Working Paper*.

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

—— (2020b): “Two-way Fixed Effects Regressions with Several Treatments,” *Working Paper*.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022): “Contamination Bias in Linear Regressions,” *Working Paper*.

Goodman-Bacon, A. (2021): “Difference-in-differences with variation in treatment timing,” *Journal of Econometrics*.
Han, C., and G. Lee (2017): “Efficient Estimation of Linear Panel Data Models with Sample Selection and Fixed Effects,” Discussion Paper Series 1707, Institute of Economic Research, Korea University.

——— (2022): “Bias Correction for Within-Group Estimation of Panel Data Models with Fixed Effects and Sample Selection,” Economic Letters.

Heckman, J. J., H. Ichimura, and P. Todd (1998): “Matching as an econometric evaluation estimator,” The Review of Economic Studies, 65, 261–294.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): “Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme,” The Review of Economic Studies, 64, 605–654.

Ishimaru, S. (2022): “Empirical Decomposition of the IV-OLS Gap with Heterogeneous and Nonlinear Effects,” Review of Economics and Statistics, Forthcoming.

Lin, L., and Z. Zhang (2022): “Interpreting the coefficients in dynamic two-way fixed effects regressions with time-varying covariates,” Economics Letters, 110604.

Lochner, L., and E. Moretti (2015): “Estimating and testing models with many treatment levels and limited instruments,” Review of Economics and Statistics, 97, 387–397.

Løken, K. V., M. Mogstad, and M. Wiswall (2012): “What Linear Estimators Miss: The Effects of Family Income on Child Outcomes,” American Economic Journal: Applied Economics, 4, 1–35.

Manning, A. (2021): “The Elusive Employment Effect of the Minimum Wage,” Journal of Economic Perspectives, 35, 3–26.

Meer, J., and J. West (2016): “Effects of the minimum wage on employment dynamics,” Journal of Human Resources, 51, 500–522.

Mogstad, M., and M. Wiswall (2010): “Linearity in instrumental variables estimation: Problems and solutions,” IZA Discussion Paper No. 5216.

Neumark, D., J. I. Salas, and W. Wascher (2014): “Revisiting the minimum wage—Employment debate: Throwing out the baby with the bathwater?” ILR Review, 67, 608–648.

Neumark, D., and W. Wascher (1992): “Employment effects of minimum and subminimum wages: panel data on state minimum wage laws,” ILR Review, 46, 55–81.
Neumark, D., W. L. Wascher, W. L. Wascher et al. (2008): *Minimum wages*: MIT press.

Schmidheiny, K., and S. Siegloch (2020): “On event studies and distributed-lags in two-way fixed effects models: Identification, equivalence, and generalization,” *ZEW-Centre for European Economic Research Discussion Paper*.

Słoczyński, T. (2020a): “Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights,” *Review of Economics and Statistics*, Forthcoming.

——— (2020b): “When Should We (Not) Interpret Linear IV Estimands as LATE?” *Working Paper*.

Sorkin, I. (2015): “Are there long-run effects of the minimum wage?” *Review of Economic Dynamics*, 18, 306–333.

Strezhnev, A. (2018): “Semiparametric weighting estimators for multi-period difference-in-differences designs,” *Working Paper*.

Sun, L., and S. Abraham (2020): “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*.

Vaghul, K., and B. Zipperer (2019): “Historical State and Sub-state Minimum Wages,” https://github.com/benzipperer/historicalminwage/releases/tag/v1.2.0, Version 1.2.0.

Wooldridge, J. (2021): “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*.
Figure 1: The TWFE and FD Coefficients (Log Employment Rate of Teens)

Figure 2: The TWFE and FD Coefficients (Net Job Creation Rate)
### Table 1: Generalized TWFE Coefficients

| Restrictions on gap (years) | Log employment rate, age 16–19 | Net job creation rate (%) |
|-----------------------------|---------------------------------|---------------------------|
|                             | (1)    | (2)    | (4)    | (5)    |
| None                        | -0.13  | -0.18  | -0.35  | 0.01   |
|                             | (0.06) | (0.06) | (1.38) | (1.21) |
| 1–5                         | -0.01  | -0.04  | -4.03  | -3.85  |
|                             | (0.06) | (0.05) | (1.11) | (1.08) |
| 6–10                        | -0.05  | -0.11  | -1.81  | -0.88  |
|                             | (0.06) | (0.06) | (1.28) | (1.26) |
| 11–15                       | -0.17  | -0.23  | 0.93   | 1.37   |
|                             | (0.07) | (0.08) | (1.92) | (1.95) |
| 16–20                       | -0.28  | -0.31  | -1.35  | -1.50  |
|                             | (0.12) | (0.12) | (2.20) | (1.89) |
| 21–28                       | -0.24  | -0.28  | 5.97   | 5.66   |
|                             | (0.11) | (0.12) | (2.19) | (2.33) |

| Covariates | No | Yes | No | Yes |
|------------|----|-----|----|-----|

Notes: The table presents coefficients of the log minimum wage effect based on a generalized TWFE estimator defined in Section 3. Standard errors are in parentheses and they are robust to heteroskedasticity and correlation across observations on the same state. Covariates are outcome pre-trends between periods $t - 12$ and $t - 3$, pre-trends of the log minimum wage effect between periods $t - 5$ and $t - 1$, and the log minimum wage in the period $t$. 
A Proofs

Some results use the following lemma, which is a well-known property of U-statistics regarding the sample covariance.

**Lemma A1.** For any two real sequences \( \{a_t\}_{t=1}^T \) and \( \{b_t\}_{t=1}^T \),

\[
\sum_{t=1}^{T} \left( a_t - \frac{1}{T} \sum_{s=1}^{T} a_s \right) \left( b_t - \frac{1}{T} \sum_{s=1}^{T} b_s \right) = \frac{1}{T} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} (a_{t+k} - a_t) (b_{t+k} - b_t).
\]

**Proof.** Define \( \bar{v} = \frac{1}{T} \sum_{t=1}^{T} v_t \) for \( v = a, b \). Then, basic algebra yields

\[
\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} (a_{t+k} - a_t) (b_{t+k} - b_t) = \frac{1}{2} \sum_{s=1}^{T} \sum_{t=1}^{T} (a_s - a_t) (b_s - b_t)
\]

\[
= \frac{1}{2} \sum_{s=1}^{T} \sum_{t=1}^{T} (a_s - \bar{a} + \bar{a} - a_t) (b_s - \bar{b} + \bar{b} - b_t)
\]

\[
= \frac{1}{2} \sum_{s=1}^{T} \sum_{t=1}^{T} \left\{ (a_s - \bar{a}) (b_s - \bar{b}) + (a_t - \bar{a}) (b_t - \bar{b}) + (a_s - \bar{a}) (b_t - \bar{b}) + (a_t - \bar{a}) (b_s - \bar{b}) \right\}
\]

\[
= T \sum_{t=1}^{T} (a_t - \bar{a}) (b_t - \bar{b}).
\]

Proof of Theorem 1

Defining \( e_{it} = Y_{it} - X_{it}' \beta - \alpha_i - C_i' \gamma_t \), the TWFE objective \( (2) \) can be rewritten as

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2 = \sum_{i=1}^{N} \left\{ \sum_{t=1}^{T} (e_{it} - \bar{e}_i)^2 + T \bar{e}_i^2 \right\} = \sum_{i=1}^{N} \left\{ \frac{1}{T} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} (\Delta_k e_{it})^2 + T \bar{e}_i^2 \right\}
\]

\[
= \frac{1}{T} \sum_{i=1}^{N} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} (\Delta_k Y_{it} - \Delta_k X_{it}' \beta - C_i' \Delta_k \gamma_t)^2 + T \sum_{i=1}^{N} \left( \bar{Y}_i - \bar{X}_i' \beta - C_i' \bar{\gamma} - \alpha_i \right)^2,
\]

(A.1)

where the second equality uses Lemma A1. In the last line of (A.1), the first term is identical to (3). For any \( (\beta, \{\gamma_t\}_{t=1}^{T}) \), letting \( \alpha_i = \bar{Y}_i - \bar{X}_i' \beta - C_i' \bar{\gamma} \) for each \( i = 1, \ldots, N \) sets the second term of (A.1) to zero. Therefore, \( (\beta, \{\gamma_t\}_{t=1}^{T}) \) that minimizes (3) also constitutes the solution to (2).

Recall that one of the elements of \( C_i \) is unity. Therefore, \( e_{it} \) can be alternatively defined
as

\[ e_{it} \equiv Y_{it} - X_{it}'\beta - \gamma_{(1),t} - C_{(-1),i}'\gamma_{(-1),t}, \]  

(A.2)

where \( C_{(-1),i} \) corresponds to the elements of \( C_i \) other than unity. Then, the FD objective \( (3) \) can be expressed as

\[
\sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \sum_{i=1}^{N} (\Delta_k e_{it})^2 = \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \left\{ \sum_{i=1}^{N} \left( \Delta_k e_{it} - \frac{1}{N} \sum_{j=1}^{N} \Delta_k e_{jt} \right)^2 \right\} + N \left( \frac{1}{N} \sum_{i=1}^{N} \Delta_k e_{it} \right)^2 \]

\[
= \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \sum_{1 \leq i < j \leq N} \left( (\Delta_k Y_{it} - \Delta_k Y_{jt}) - (\Delta_k X_{jt} - \Delta_k X_{it})'\beta - (C_j - C_i)'\Delta_k \gamma_t \right)^2 
+ N \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \left( \frac{1}{N} \sum_{i=1}^{N} (\Delta_k Y_{it} - \Delta_k X_{it}'\beta - C_{(-1),i}'\Delta_k \gamma_{(-1),t}) - \Delta_k \gamma_{(1),t} \right)^2,
\]  

(A.3)

where the second equality applies Lemma \( \text{A1} \) to cross-sectional summation across \( i = 1, \ldots, N \), and the third equality uses the definition \( (A.2) \). For any \( (\beta, \{\gamma_{(-1),t}\}_{t=1}^{T}) \), letting \( \gamma_{(1),t} = Y_{it} - X_{it}'\beta - C_{(-1),i}'\gamma_{(-1),t} \) for each \( t = 1, \ldots, T \) can set the second term of \( (A.3) \) to zero. In addition, the first term of \( (A.3) \) does not depend on \( \{\gamma_{(1),t}\}_{i=1}^{N} \). Therefore, minimizing the first term of \( (A.3) \), which is the DID objective in defined in \( (5) \), and minimizing the FD objective \( (3) \) are equivalent.

**Proof of Theorem 2**

Define

\[ \widetilde{Y}_{it} = Y_{it} - C_i' \left( \sum_{i=1}^{N} C_i C_i' \right)^{-1} \left( \sum_{i=1}^{N} C_i Y_{it} \right) \]

to be a residual from a regression of \( Y_{it} \) on \( C_i \) separately performed for each \( t = 1, \ldots, T \), and define \( \widetilde{X}_{it} \) in the same manner. The FWL theorem implies that each \( \widehat{\beta}_{FD,k} \) is given by

\[
\widehat{\beta}_{FD,k} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \widetilde{X}_{it} \Delta_k \widetilde{X}_{it}' \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \widetilde{X}_{it} \Delta_k \widetilde{Y}_{it} \right).
\]  

(A.4)

On the other hand, solving the least-squares problem \( (2) \), \( \widehat{\beta}_{FE} \) is given by

33
\[
\tilde{\beta}_{FE} = \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \tilde{X}_{it} - \frac{1}{T} \sum_{s=1}^{T} \tilde{X}_{is} \right) \left( \tilde{X}_{it} - \frac{1}{T} \sum_{s=1}^{T} \tilde{X}_{is} \right)' \right\}^{-1} 
\times \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \tilde{X}_{it} - \frac{1}{T} \sum_{s=1}^{T} \tilde{X}_{is} \right) \left( \tilde{Y}_{it} - \frac{1}{T} \sum_{s=1}^{T} \tilde{Y}_{is} \right) \right\}. 
\] (A.5)

Applying Lemma A1 to each element of matrices in (A.5),

\[
\tilde{\beta}_{FE} = \left( \sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \tilde{X}_{it} \Delta_k \tilde{X}_{it}' \right)^{-1} \left( \sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \tilde{X}_{it} \Delta_k \tilde{Y}_{it} \right). 
\] (A.6)

Note that (A.6) is a solution not only to (2) but also to

\[
\min_{\beta, (\gamma^*, k, t)} \sum_{i=1}^{N} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \left( \Delta_k \tilde{Y}_{it} - \Delta_k \tilde{X}_{it}' \beta - C_k \gamma^* \right)^2, 
\] (A.7)

since the FWL theorem implies a regression of \( \Delta_k \tilde{Y}_{it} \) on \( \Delta_k \tilde{X}_{it} \) yields the coefficient on \( \Delta_k \tilde{X}_{it} \) in (A.7). Therefore, the separability restriction \( \gamma^* = \gamma_{k+t} - \gamma_t \) imposed on (3) compared with (A.7) does not bind.

Combining (A.4) and (A.6), the TWFE estimator can be expressed a matrix-weighted average

\[
\hat{\beta}_{FE} = \sum_{k=1}^{T-1} \tilde{\Omega}_k \hat{\beta}_{FD,k},
\]

\[
\hat{\Omega}_k = \left( \sum_{\ell=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-\ell} \Delta_{\ell} \tilde{X}_{it} \Delta_{\ell} \tilde{X}_{it}' \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \tilde{X}_{it} \Delta_k \tilde{X}_{it}' \right),
\]

where the weight matrices \( \hat{\Omega}_k \) are, by construction, all positive semi-definite and sum to the identity matrix.

**Proof of Theorem 3**

Algebraic transformation of (9) yields

\[
\tilde{\beta}_{FE} = \frac{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \tilde{Y}_{it} \left( \Delta_k \tilde{W}_{it} - \Delta_k \tilde{W}_{it}' \tilde{\delta}_{FE} \right)}{\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \tilde{D}_{it} \left( \Delta_k \tilde{D}_{it} - \Delta_k \tilde{D}_{it}' \tilde{\delta}_{FE} \right)}
\]

34
Proof of Lemma 1

The FD coefficient is given by

$$
\hat{\delta}_{FE}^W = \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \hat{W}_{it} \Delta_k \hat{W}'_{it} \right)^{-1} \left( \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \hat{W}_{it} \left( \Delta_k Y_{it} - \hat{\beta}_{FD,k} \Delta_k D_{it} \right) \right),
$$

which follows from FWL theorem. The last equality uses

$$
\sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \Delta_k \hat{W}_{it} \Delta_k \hat{W}'_{it} = 0,
$$

which follows from Theorem 2.
\[ \times \left( \sum_{t=1}^{T-k} E \left[ (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it})) (\Delta_k D_{it} - L_{Ci} (\Delta_k D_{it})) \right] \right) \). \] (A.8)

On the other hand, applying Lemma A1 to (15) yields

\[ \delta_{FE}^W = \left( \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E \left[ (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it})) (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it})) \right] \right)^{-1} \times \left( \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E \left[ (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it})) (\Delta_k D_{it} - L_{Ci} (\Delta_k D_{it})) \right] \right). \] (A.9)

In addition,

\[ E [\Delta_k D_{it}|\Delta_k W_{it}, C_i] - L_{Ci} (\Delta_k D_{it}) = E [\Delta_k D_{it}|\Delta_k W_{it}, C_i] - L_{Ci} \left( E [\Delta_k D_{it}|\Delta_k W_{it}, C_i] \right) = (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it}))' \delta_{FE}^W, \] (A.10)

where the last equality uses Assumption LH. Plugging (A.10) into (A.8) and (A.9) yields \( \delta_{FE}^W = \delta_{FE} = \delta_{FD,k}^W \).

**Proof of Theorem 4**

Using the definition (16) and Assumption L,

\[ \Delta_k R_{it} = \Delta_k D_{it} - L_{Ci} (\Delta_k D_{it}) - (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it}))' \delta_{FE}^W. \]

\[ = (\Delta_k D_{it} - E [\Delta_k D_{it}|\Delta_k W_{it}, C_i]) + (E [\Delta_k D_{it}|\Delta_k W_{it}, C_i] - L_{Ci} (\Delta_k D_{it})) - (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it}))' \delta_{FE}^W \]

\[ = (\Delta_k D_{it} - E [\Delta_k D_{it}|\Delta_k W_{it}, C_i]) + (\Delta_k W_{it} - L_{Ci} (\Delta_k W_{it}))' (\delta_{k,t}^W - \delta_{FE}^W). \] (A.11)

In addition, it follows from Lemma 1 that (A.11) simplifies to

\[ \Delta_k R_{it} = \Delta_k D_{it} - E [\Delta_k D_{it}|\Delta_k W_{it}, C_i] \] (A.12)

under Assumption LH.

Suppose \( d_0 \in (d, \bar{d}) \) satisfies Assumption CT. Then,

\[ \sum_{t=1}^{T} E [Y_{it}(d_0)R_{it}] = \frac{1}{T} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} E [\Delta_k Y_{it}(d_0)\Delta_k R_{it}] \]
\[= \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E [E [\Delta_k Y_{it}(d_0) | \Delta_k D_{it}, \Delta_k W_{it}, C_i] \Delta_k R_{it}] \]
\[= \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E [\Delta_k Y_{it}(d_0) | \Delta_k W_{it}, C_i] \Delta_k R_{it}] \]
\[= \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E [\Delta_k Y_{it}(d_0)] E [\Delta_k R_{it} | \Delta_k W_{it}, C_i] \]
\[= \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E [\Delta_k Y_{it}(d_0) (\Delta_k W_{it} - L_{C_i} (\Delta_k W_{it}))'] (\delta_W^k - \delta_W^W'] . \ (A.13) \]

The first equality follows from Lemma [A1] and \(\sum_{t=1}^{T} R_{it} = 0\). The second equality uses the law of iterated expectations, given that \(\Delta_k R_{it}\) is a function of \((\Delta_k D_{it}, \Delta_k W_{it}, C_i)\). The third equality uses Assumption [CT] and the fourth equality relies on the law of iterated expectations. The last equality uses (A.11).

Using (A.11), the denominator in (17) is given by
\[= \sum_{t=1}^{T} E \left[ D_{it} R_{it} \right] = \frac{1}{T} \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E [\Delta_k D_{it} \Delta_k R_{it}] \]
\[= \frac{1}{T} \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E \left[ (\Delta_k D_{it} - E [\Delta_k D_{it} | \Delta_k W_{it}, C_i])^2 + \left\{ (\Delta_k W_{it} - L_{C_i} (\Delta_k W_{it}))' (\delta_W^k - \delta_W^W) \right\}^2 \right] , \]
which is strictly positive under Assumption [V].

The numerator in (17) is given by
\[= \sum_{t=1}^{T} E \left[ Y_{it} R_{it} \right] = \sum_{t=1}^{T} E \left[ (Y_{it}(D_{it}) - Y_{it}(d_0)) R_{it} + Y_{it}(d_0) R_{it} \right] \]
\[= \sum_{t=1}^{T} E \left[ (Y_{it}(D_{it}) - Y_{it}(d_0)) R_{it} \right] + \frac{1}{T} \sum_{t=1}^{T-1} T-k \sum_{t=1}^{T-k} E \left[ \Delta_k Y_{it}(d_0) (\Delta_k W_{it} - L_{C_i} (\Delta_k W_{it}))' (\delta_W^k - \delta_W^W) \right] \]. \ (A.14) \]

where the first equality uses Assumption [PO] and the second equality uses (A.13). Moreover, Lemma [I] implies that (A.14) can be rewritten as
\[= \sum_{t=1}^{T} E \left[ Y_{it} R_{it} \right] = \sum_{t=1}^{T} E \left[ (Y_{it}(D_{it}) - Y_{it}(d_0)) R_{it} \right] \]
under Assumption [LH]. The weighted-average expression in Theorem [I] immediately follows

37
from this result.

Finally, using (A.12) and \( \sum_{t=1}^{T} R_{it} = 0 \),

\[
(D_{it} - d_0) R_{it} = (D_{it} - d_0) \sum_{s=1}^{T} (R_{it} - R_{is})
\]

\[
= (D_{it} - d_0) \sum_{s=1}^{T} \{(D_{it} - D_{is}) - E[D_{it} - D_{is}|W_{it} - W_{is}, C_i]\},
\]

which implies (18).

**Proof of Theorem 5**

For any \( \bar{d} \in (\underline{d}, d) \),

\[
\sum_{t=1}^{T} E[Y_{it}(\bar{d}) R_{it}] = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{T-k} E[\Delta_k Y_{it}(\bar{d}) (\Delta_k W_{it} - L_{C_i} (\Delta_k W_{it}))' (\delta^W_{k,t} - \delta^W_{FE})]
\]

\[
= 0.
\]

(A.15)

The first equality uses (A.13), which relies on Assumptions CTE and C. The second equality uses (A.12), which relies on Assumption LH.

Then, the numerator of (17) is given by

\[
\sum_{t=1}^{T} E[Y_{it} R_{it}] = \sum_{t=1}^{T} E[(Y_{it}(D_{it}) - Y_{it}(d_0)) R_{it}]
\]

\[
= \sum_{t=1}^{T} E \left[ \int_{\underline{d}}^{\overline{d}} Y_{it}'(x) (\mathbb{1}_{D_{it} \geq x} - \mathbb{1}_{d_0 \geq x}) d x R_{it} \right]
\]

\[
= \int_{\underline{d}}^{\overline{d}} \sum_{t=1}^{T} E[Y_{it}'(x) (\mathbb{1}_{D_{it} \geq x} - \mathbb{1}_{d_0 \geq x}) R_{it}] d x
\]

\[
= \int_{\underline{d}}^{\overline{d}} \sum_{t=1}^{T} E[Y_{it}'(x) (\mathbb{1}_{D_{it} \geq x} - \mathbb{1}_{d_0 \geq x} + g(x)) R_{it}] d x.
\]

(A.16)

The first equality uses (A.15) and Assumption PO. The second equality follows from the fundamental theorem of calculus and an identity that \( \mathbb{1}_{D_{it} \geq x} - \mathbb{1}_{d_0 \geq x} \) is equal to 1 if \( d_0 < x \leq D_{it} \), to -1 if \( D_{it} < x \leq d_0 \), and to 0 otherwise. The third equality follows from Fubini’s theorem. Note that the absolute integrability condition required for using the theorem is
satisfied because Assumption LC and the Cauchy–Schwarz inequality imply
\[
E \left[ \int \overline{d} |Y'_{it}(x) \left( \mathbb{1}_{D_{it} \geq x} - \mathbb{1}_{d_0 \geq x} \right) R_{it} | \, dx \right] \leq E \left[ M_{it} \cdot |D_{it} - d_0| \cdot |R_{it}| \right] \\
\leq \sqrt{E \left[ M_{it}^2 \cdot (D_{it} - d_0)^2 \right]} \cdot E \left[ R_{it}^2 \right] \\
< \infty.
\]

Finally, the fourth equality in (A.16) follows from (A.15) and Assumption LC.

Using (A.16) immediately yields the weighted average expression in Theorem 5. In addition,
\[
\int \overline{d} \sum_{t=1}^{T} E \left[ \psi_{it}(x; d_0) \right] \, dx = \sum_{t=1}^{T} E \left[ \int \overline{d} \psi_{it}(x; d_0) \, dx \right] \\
= \sum_{t=1}^{T} E \left[ (D_{it} - d_0) R_{it} \right] \\
\sum_{t=1}^{T} E \left[ D_{it} R_{it} \right] \\
= 1,
\]
where the first equality uses Fubini’s theorem and the last equality follows from \( E[R_{it}] = 0 \).

**Proof of Theorem 6**

Since the set of assumptions in Theorem 5 is still satisfied,
\[
\beta_{FE}^{D} = \int \overline{d} \sum_{t=1}^{T} E \left[ Y'_{it}(x) \psi_{it}(x; d_0) \right] \, dx.
\]

For any \( d \in (d, \overline{d}) \),
\[
\sum_{t=1}^{T} E \left[ Y'_{it}(d) \left( \psi_{it}(d; d_0) - \frac{1}{T} \sum_{s=1}^{T} \psi_{is}(d; d_0) \right) \right] \\
= \frac{1}{T} \sum_{s > t} E \left[ (Y'_{is}(d) - Y'_{it}(d)) \left( \psi_{is}(d; d_0) - \psi_{it}(d; d_0) \right) \right] \\
= \frac{1}{T} \sum_{s > t} E \left[ E \left[ Y'_{is}(d) - Y'_{it}(d) \mid D_{is}, D_{it}, D_{i}, W_{is}, W_{it}, W, C_i \right] \left( \psi_{is}(d; d_0) - \psi_{it}(d; d_0) \right) \right] \\
= \frac{1}{T} \sum_{s > t} E \left[ E \left[ Y'_{is}(d) - Y'_{it}(d) \mid W_{is} - W_{it}, C_i \right] \left( \psi_{is}(d; d_0) - \psi_{it}(d; d_0) \right) \right]
\]
Letting $A(t, s)$ follows from (A.12), which relies on Assumption LH. This proves (20).

Assumptions SE and S. The last equality follows from the law of iterated expectations.

Let $A(s, t)$ be any real-valued function that satisfies $A(s, t) = -A(t, s)$ and $A(t, t) = 0$ and let $\{b_t\}_{t=1}^T$ be any real sequence

$$
\sum_{s>t} (b_s - b_t) A(s, t) = \sum_{s>t} b_s A(s, t) - \sum_{s>t} b_t A(s, t)
$$

$$
= \sum_{t>s} b_t A(t, s) + \sum_{s>t} b_t A(t, s) = \sum_{t=1}^T \sum_{s=1}^T b_t A(t, s). \tag{A.18}
$$

Letting $A(s, t) = E[\psi_{is}(d; d_0) - \psi_{it}(d; d_0)|W_{is} - W_{it}, C_i]$ and $b_t = Y_{it}(d)$ in (A.17) yields

$$
\sum_{t=1}^T E[Y_{it}'(d)\psi_{it}(d; d_0)] = \sum_{t=1}^T E[Y_{it}'(d)\frac{1}{T}\sum_{s=1}^T (\psi_{is}(d; d_0) + E[\psi_{it}(d; d_0) - \psi_{is}(d; d_0)|W_{it} - W_{is}, C_i])].
$$

This proves (22).

Then,

$$
\frac{1}{T} \sum_{s=1}^T \psi_{is}(d; d_0) = \frac{1}{T} \sum_{s=1}^T (\mathbb{1}_{D_{is} \geq d} - \mathbb{1}_{D_{is} \geq d} R_{is}) \frac{R_{is}}{E[D_{it'} R_{it}]} = \frac{1}{T} \sum_{s=1}^T \mathbb{1}_{D_{is} \geq d} \sum_{t'=1}^T E[D_{it'} R_{it}]
$$

follows from $\frac{1}{T} \sum_{s=1}^T R_{is} = 0$ and

$$
E[\psi_{it}(d; d_0) - \psi_{is}(d; d_0)|W_{it} - W_{is}, C_i]
$$

$$
= E[\mathbb{1}_{D_{is} \geq d} R_{it} - \mathbb{1}_{D_{is} \geq d} R_{is} - (R_{it} - R_{is}) \mathbb{1}_{d_0 \geq d}|W_{it} - W_{is}, C_i]
$$

$$
= \frac{1}{T} \sum_{t'=1}^T E[D_{it'} R_{it}]
$$

follows from (A.12), which relies on Assumption LH. This proves (20).

Finally,

$$
\sum_{t=1}^T E[\psi_{it}'(d) R_{it}] = \frac{1}{T} \sum_{s=1}^T \sum_{t=1}^T E[R_{it} (\psi_{is}(d; d_0) + E[\psi_{it}(d; d_0) - \psi_{is}(d; d_0)|W_{it} - W_{is}, C_i])]
$$
\[
\frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} E \left[ R_{it} E \left[ \psi_{it}(d; d_0) - \psi_{is}(d; d_0) | W_{it} - W_{is}, C_i \right] \right] \\
= \frac{1}{T} \sum_{s > t} E \left[ (R_{is} - R_{it}) E \left[ \psi_{is}(d; d_0) - \psi_{it}(d; d_0) | W_{is} - W_{it}, C_i \right] \right] \\
= \frac{1}{T} \sum_{s > t} E \left[ E \left[ R_{is} - R_{it} | W_{is} - W_{it}, C_i \right] (\psi_{is}(d; d_0) - \psi_{it}(d; d_0)) \right] \\
= 0,
\]

where the first equality uses (22), the second equality follows from \( \sum_{t=1}^{T} R_{it} = 0 \), the third equality uses (A.18) by letting \( A(s, t) = E \left[ \psi_{is}(d; d_0) - \psi_{it}(d; d_0) | W_{is} - W_{it}, C_i \right] \) and \( b_t = R_{it} \), the fourth equality uses the law of iterated expectations, and the last equality uses (A.12), which relies on Assumption LH. This proves (21).

**Proof of Theorem 7**

Using Assumption LH', a residual \( r_{it}^{(k)} \) is given by

\[
r_{it}^{(k)} = \Delta_k D_{it} - E \left[ \Delta_k D_{it} | \Delta_k W_{it}, V_{it}, C_i \right]. \tag{A.19}
\]

Then,

\[
E \left[ r_{it}^{(k)} \Delta_k Y_{it}(d_0) \right] = E \left[ r_{it}^{(k)} E \left[ \Delta_k Y_{it}(d_0) | \Delta_k D_{it}, \Delta_k W_{it}, V_{it}, C_i \right] \right] \\
= E \left[ r_{it}^{(k)} E \left[ \Delta_k Y_{it}(d_0) | \Delta_k W_{it}, V_{it}, C_i \right] \right] \\
= E \left[ \Delta_k Y_{it}(d_0) E \left[ r_{it}^{(k)} | \Delta_k W_{it}, V_{it}, C_i \right] \right] \\
= 0, \tag{A.20}
\]

where the first and third equalities use the law of iterated expectations, the second equality follows from Assumption CT', and the last equality uses (A.19).

For any random variable \( A_{it} \), a basic algebra yields

\[
\sum_{k=K}^{K} \sum_{t=1}^{T-k} E \left[ r_{it}^{(k)} \Delta_k A_{it} \right] = \sum_{s > t} \mathbb{1}_{K \leq |s-t| \leq K} E \left[ r_{it}^{(s-t)} (A_{is} - A_{it}) \right] \\
= \sum_{s > t} \mathbb{1}_{K \leq |s-t| \leq K} E \left[ r_{it}^{(s-t)} A_{is} \right] - \sum_{s > t} \mathbb{1}_{K \leq |s-t| \leq K} E \left[ r_{it}^{(s-t)} A_{it} \right] \\
= \sum_{t > s} \mathbb{1}_{K \leq |s-t| \leq K} E \left[ r_{is}^{(t-s)} A_{it} \right] - \sum_{s > t} \mathbb{1}_{K \leq |s-t| \leq K} E \left[ r_{it}^{(s-t)} A_{it} \right]
\]

41
\[
\sum_{t=1}^{T} E \left[ A_{it} \sum_{s=1}^{T} \left( r_{is}^{(t-s)} \mathbb{1}_{K \leq t-s \leq R} - r_{it}^{(s-t)} \mathbb{1}_{K \leq s-t \leq R} \right) \right]. \tag{A.21}
\]

Letting \( A_{it} = Y_{it}(D_{it}) - Y_{it}(d_0) \) in (A.21) and combining it with (A.20), the numerator of (24) is given by
\[
\sum_{k=K}^{K} \sum_{s=1}^{T} E \left[ r_{it}^{(k)} \Delta_k Y_{it} \right] = \sum_{t=1}^{T} E \left[ (Y_{it}(D_{it}) - Y_{it}(d_0)) \sum_{s=1}^{T} \left( r_{is}^{(t-s)} \mathbb{1}_{K \leq t-s \leq R} - r_{it}^{(s-t)} \mathbb{1}_{K \leq s-t \leq R} \right) \right]. \tag{A.22}
\]

In addition, it follows from (A.19) that
\[
\sum_{s=1}^{T} \left( r_{is}^{(t-s)} \mathbb{1}_{K \leq t-s \leq R} - r_{it}^{(s-t)} \mathbb{1}_{K \leq s-t \leq R} \right) = \sum_{s=1}^{T} \mathbb{1}_{K \leq |t-s| \leq R} \left( D_{it} - D_{is} - E \left[ D_{it} - D_{is} | V_{i,\min(s,t)}, W_{it} - W_{is}, C_i \right] \right).
\]

Each element of the denominator in (24) is given by
\[
E \left[ r_{it}^{(k)} \Delta_k D_{it} \right] = E \left[ \Delta_k D_{it} (\Delta_k D_{it} - E [\Delta_k D_{it} | W_{it}]) \right] = E \left[ Var (\Delta_k D_{it} | W_{it}) \right],
\]
which sums to a nonnegative value due to Assumption \( V \). In addition, by letting \( A_{it} = D_{it} \) in (A.21), the denominator of (24) is given by
\[
\sum_{k=K}^{K} \sum_{t=1}^{T-k} E \left[ r_{it}^{(k)} \Delta_k D_{it} \right] = \sum_{t=1}^{T} E \left[ (D_{it} - d_0) \sum_{s=1}^{T} \left( r_{is}^{(t-s)} \mathbb{1}_{K \leq t-s \leq R} - r_{it}^{(s-t)} \mathbb{1}_{K \leq s-t \leq R} \right) \right].
\]

Combining the numerator and the denominator gives the weighted-average expression in Theorem 7.

**B Results in Unbalanced Panels**

The main paper assumes a balanced panel setting, in which each unit–period combination has exactly one observation. This section discusses how the results in the main paper extends to or differs in a more general setting.
B.1 Equivalence of Least-Square Objectives

Without assuming a balanced panel, the least-squares objective in \((2)\) should be rewritten as

\[
\min_{\beta,(\alpha_i)_{i=1}^N,(\gamma_t)_{t=1}^T} \sum_{i=1}^{N} \sum_{t=1}^{T} B_{it} (Y_{it} - X_{it}'\beta - \alpha_i - \gamma_t)^2 . \tag{B.1}
\]

In an unbalanced panel, \(B_{it}\) is an observability indicator that takes a value of 1 if an observation \((i, t)\) is available and a value of 0 if the observation is missing. More generally, \(B_{it}\) can be any real number. For example, a TWFE regression may use repeated cross-section pooled FD regressions naturally extends to these cases. In particular, the coefficient on \(X_{it}\) given by \((B.1)\) is algebraically identical to the coefficient on \(\Delta_k X_{it}\) given by the following least-squares problem.

\[
\min_{\beta,(\gamma_t)_{t=1}^T} \sum_{k=1}^{T-1} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \frac{B_{it} B_{i,t+k}}{\bar{B}_i} \left( \Delta_k Y_{it} - \Delta_k X_{it}'\beta - C_i'(\gamma_{t+k} - \gamma_t) \right)^2 . \tag{B.2}
\]

The proof is as follows. Suppose a sequence \(\{w_t\}_{t=1}^{T}\) satisfies \(\sum_{t=1}^{T} w_t = T\) and define \(\bar{a}^w \equiv \frac{1}{T} \sum_{t=1}^{T} w_t a_t\) for another sequence \(\{a_t\}_{t=1}^{T}\). Basic algebra yields

\[
\frac{1}{T} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} w_{t+k} w_t (\Delta_k a_t)^2 = \frac{1}{2T} \sum_{s=1}^{T} \sum_{t=1}^{T} w_s w_t (a_s - a_t)^2 = \frac{1}{2T} \sum_{s=1}^{T} \sum_{t=1}^{T} w_s w_t (a_s - \bar{a}^w + \bar{a}^w - a_t)^2
\]

\[
= \frac{1}{2T} \sum_{s=1}^{T} \sum_{t=1}^{T} w_s w_t \left\{ (a_s - \bar{a}^w)^2 + (a_t - \bar{a}^w)^2 - 2 (a_s - \bar{a}^w)(a_t - \bar{a}^w) \right\} = \sum_{t=1}^{T} w_t (a_t - \bar{a}^w)^2
\]

\[
= \sum_{t=1}^{T} w_t a_t^2 - (\bar{a}^w)^2 . \tag{B.3}
\]

Exploiting the algebraic identity \((B.3)\) by letting \(w_t = \frac{B_{it}}{\bar{B}_i}\) and \(a_t = Y_{it} - X_{it}'\beta - \alpha_i - \gamma_t\) for each \(i = 1, \ldots, N\), the objective function \((B.1)\) can be expressed as

\[
\sum_{i=1}^{N} \sum_{t=1}^{T} B_{it} (Y_{it} - X_{it}'\beta - \alpha_i - C_i'\gamma_t)^2 = \frac{1}{T} \sum_{i=1}^{N} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \frac{B_{it} B_{i,t+k}}{\bar{B}_i} \left( \Delta_k Y_{it} - \Delta_k X_{it}'\beta - C_i'(\gamma_{t+k} - \gamma_t) \right)^2
\]

\[
+ \sum_{i=1}^{N} \bar{B}_i \left( \sum_{t=1}^{T} \frac{B_{it}}{TB_i} (Y_{it} - X_{it}'\beta - C_i'\gamma_t) - \alpha_i \right)^2 . \tag{B.4}
\]

Since choosing \(\beta\) and \(\{\gamma_t\}_{t=1}^{T}\) that solves \((B.2)\) and letting \(\alpha_i = \sum_{t=1}^{T} \frac{B_{it}}{TB_i} (Y_{it} - X_{it}'\beta - C_i'\gamma_t)\)
for each \( i = 1, \ldots, N \) results in minimizing (B.4), the two objectives (B.1) and (B.2) yields the same estimates of \( \beta \).

### B.2 Weighted-Average Relationship

Note that (B.2) is distinct from

\[
\min_{\beta, \{\gamma^*_k, t\}} \sum_{i=1}^{N} \sum_{k=1}^{T-1} \sum_{t=1}^{T-k} \frac{B_{it}B_{i,t+k}}{B^2} \left( \Delta_k Y_{it} - \Delta_k X_i' \beta - C_i' \gamma^*_k \right)^2. \tag{B.5}
\]

While (B.2) includes for fixed effects for starting and ending periods in an additive manner, (B.5) accounts for fixed effects associated with all possible combinations of starting and ending periods. Only in a balanced panel with \( B_{it} = 1 \) for every \((i, t)\), the two problems (B.2) and (B.5) yield identical estimates of \( \beta \) (see the proof of Theorem 2).

As a result, even when \( X_{it} \) is univariate, unlike in Theorem 2, the TWFE coefficient \( \hat{\beta}_{FE} \) given by (B.1) or (B.2) is not exactly identical to a convex combination of FD coefficients \( \hat{\beta}_{FD,k} \) that are given by solving

\[
\min_{\beta_k, \{\gamma^*_k, t\}} \sum_{i=1}^{N} \sum_{t=1}^{T-k} \frac{B_{it}B_{i,t+k}}{B^2} \left( \Delta_k Y_{it} - \Delta_k X_i' \beta_k - C_i' \gamma^*_k \right)^2 \tag{B.6}
\]

for \( k = 1, \ldots, T - 1 \). One possible way to restore a weighted-average relationship is to use (B.5) instead of (B.1) or (B.2) for producing the TWFE coefficient.

### C Additional Features of a Generalized TWFE Regression

#### C.1 Numerical Interpretation

The following result provides a numerical interpretation of a generalized TWFE estimator.

**Theorem C1.** Let \( \hat{\beta}^D_{GFE} \) be the coefficient on \( \Delta_k D_{it} \) given by a generalized TWFE regression in (23). Then,

\[
\hat{\beta}^D_{GFE} = \sum_{k=K}^{K} \hat{\omega}^D_{k} \hat{\beta}^D_{GFD,k} \tag{C.1}
\]
where each $\hat{\beta}_{GFD,k}^D$ is the coefficient on $\Delta_k D_{it}$ from an FD regression

$$\min_{\beta_k^D, \delta_k^W, \delta_k^V, \{\gamma_{k,t}\}} \sum_{i=1}^N \sum_{t=1}^{T-k} \left( \Delta_k Y_{it} - \hat{\beta}_{GFE}^D \Delta_k D_{it} - \Delta_k W'_{it} \beta_k^W - V'_{it} \delta_k^V - C'_{it} \gamma_{k,t} \right)^2,$$

for each $k = K, \ldots, K$. The weights satisfy $\sum_{k=K}^K \hat{w}_k^D = 1$ and each $\hat{w}_k^D$ is proportional to the residual sum of squares from an auxiliary FD regression

$$\min_{\delta_k^W, \delta_k^V, \{\gamma_{k,t}\}} \sum_{i=1}^N \sum_{t=1}^{T-k} \left( \Delta_k D_{it} - \Delta_k W'_{it} \delta_k^W - V'_{it} \delta_k^V - C'_{it} \gamma_{k,t} \right)^2.$$

Proof. Applying the FWL theorem to (23) yields

$$\hat{\beta}_{GFE}^D = \frac{\sum_{i=1}^N \sum_{k=K}^K \sum_{t=1}^{T-k} \Delta_k Y_{it} \left( \Delta_k D_{it} - \Delta_k W'_{it} \delta_k^W - V'_{it} \delta_k^V - C'_{it} \gamma_{k,t} \right)}{\sum_{i=1}^N \sum_{k=K}^K \sum_{t=1}^{T-k} \Delta_k D_{it} \left( \Delta_k D_{it} - \Delta_k W'_{it} \delta_k^W - V'_{it} \delta_k^V - C'_{it} \gamma_{k,t} \right)}.$$

where $\hat{\delta}_k^W, \hat{\delta}_k^V,$ and $\hat{\gamma}_{k,t}$ are the coefficients from an auxiliary regression (C.3) for each $k = K, \ldots, K$. On the other hand, applying the FWL theorem to (C.2) yields

$$\hat{\beta}_{GFD,k}^D = \frac{\sum_{i=1}^N \sum_{t=1}^{T-k} \Delta_k Y_{it} \left( \Delta_k X_{it} - \Delta_k W'_{it} \hat{\delta}_k^W - V'_{it} \hat{\delta}_k^V - C'_{it} \hat{\gamma}_{k,t} \right)}{\sum_{i=1}^N \sum_{t=1}^{T-k} \Delta_k X_{it} \left( \Delta_k X_{it} - \Delta_k W'_{it} \hat{\delta}_k^W - V'_{it} \hat{\delta}_k^V - C'_{it} \hat{\gamma}_{k,t} \right)}.$$

Therefore, the weighted-average expression (C.1) holds, where the weights are given by

$$\hat{w}_k^D = \frac{\sum_{i=1}^N \sum_{t=1}^{T-k} \left( \Delta_k X_{it} - \Delta_k W'_{it} \hat{\delta}_k^W - V'_{it} \hat{\delta}_k^V - C'_{it} \hat{\gamma}_{k,t} \right)^2}{\sum_{k'=K}^K \sum_{i=1}^N \sum_{t=1}^{T-k'} \left( \Delta_{k'} X_{it} - \Delta_{k'} W'_{it} \hat{\delta}_{k'}^W - V'_{it} \hat{\delta}_{k'}^V - C'_{it} \hat{\gamma}_{k,t} \right)^2}.$$

Theorem (C.1) suggests that a weighted-average relationship between a TWFE coefficient and FD coefficients, which fails in the presence of time-varying covariates as seen in Theorem 3 can be restored by a generalized TWFE regression.

### C.2 Motivating a Generalized TWFE Regression by a Linear Model

Section 4 provides a causal interpretation of a generalized TWFE regression without using a linear model. However, it is possible to motivate a generalized TWFE regression by a linear
model.

Following the way in which Angrist and Pischke (2017) motivates a linear regression equation in cross-sectional data, I produce a linear regression equation using the assumption of constant causal effects, and conditional common trends, and linearity. Suppose $Y_{it}(d) - Y_{it}(d_0) = \beta^D(d - d_0)$, i.e., per-unit treatment effects are constant. Then, the $k$-period difference in observed outcomes $Y_{it} \equiv Y_{it}(D_{it})$ satisfies

$$\Delta_k Y_{it} = \beta^D \Delta_k D_{it} + \Delta_k Y_{it}(d_0). \quad (C.4)$$

Make a conditional common trends assumption

$$E[\Delta_k Y_{it}(d_0) | \Delta_k D_{it}, \Delta_k W_{it}, V_{it}, C_i] = E[\Delta_k Y_{it}(d_0) | \Delta_k W_{it}, V_{it}, C_i]$$

as in Assumption $\text{CT}'$ and assume linearity of the conditional mean

$$E[\Delta_k Y_{it}(d_0) | \Delta_k W_{it}, V_{it}, C_i] = \Delta_k W_{it}' \beta^W_k + V_{it}' \beta^V_k + C_i' \gamma_{k,t}^*.$$ 

Then, the equation $(C.4)$ yields a linear regression equation

$$\Delta_k Y_{it} = \beta^D \Delta_k D_{it} + \Delta_k W_{it}' \beta^W_k + V_{it}' \beta^V_k + C_i' \gamma_{k,t}^* + \epsilon_{ikt},$$

where $\epsilon_{ikt} \equiv \Delta_k Y_{it}(d_0) - E[\Delta Y_{it}(d_0) | \Delta_k W_{it}, V_{it}, C_i]$. Since $E[\epsilon_{ikt} | \Delta_k D_{it}, \Delta_k W_{it}, V_{it}, C_i] = 0$, a regression of $\Delta_k Y_{it}$ on $(\Delta_k D_{it}, \Delta_k W_{it}, V_{it})$ gives a consistent estimator of $\beta^D$ as the coefficient on $\Delta_k D_{it}$. 

46