Difference-in-Differences when Parallel Trends Holds Conditional on Covariates*

Carolina Caetano[†] Brantly Callaway[‡]

September 9, 2024

Abstract

In this paper, we study difference-in-differences identification and estimation strategies when the parallel trends assumption holds after conditioning on covariates. We consider empirically relevant settings where the covariates can be time-varying, time-invariant, or both. We uncover a number of weaknesses of commonly used two-way fixed effects (TWFE) regressions in this context, even in applications with only two time periods. In addition to some weaknesses due to estimating linear regression models that are similar to cases with cross-sectional data, we also point out a collection of additional issues that we refer to as *hidden linearity bias* that arise because the transformations used to eliminate the unit fixed effect also transform the covariates (e.g., taking first differences can result in the estimating equation only including the change in covariates over time, not their level, and also drop time-invariant covariates altogether). We provide simple diagnostics for assessing how susceptible a TWFE regression is to hidden linearity bias based on reformulating the TWFE regression as a weighting estimator. Finally, we propose simple alternative estimation strategies that can circumvent these issues.

JEL Codes: C14, C21, C23

Keywords: Difference-in-Differences, Time-Varying Covariates, Time-Invariant Covariates, Hidden Linearity Bias, Two-way Fixed Effects Regression, Doubly Robust Estimation, Conditional Parallel Trends, Treatment Effect Heterogeneity

*Some of the results in this paper were originally in "Difference-in-differences with time-varying covariates" (Caetano, Callaway, Payne, and Sant'Anna (2022)). This paper and our companion paper, "Difference-in-differences with bad controls" (Caetano, Callaway, Payne, and Rodrigues (2023)), replace that paper. The code for the new estimation approaches proposed in the paper is provided in the `R` `pte` package, available at `https://github.com/bcallaway11/pte`. Code for the TWFE and AIPW diagnostics discussed in the paper is available in the `R` `twfeweights` package, which is available at `https://github.com/bcallaway11/twfeweights.` We thank Kyle Butts, Andrew Goodman-Bacon, Pedro Sant'Anna, Tymon Sloczynski, seminar participants at Auburn University, Central European University, Columbia University, the University of Connecticut, the University of Duisburg-Essen, Emory University Quantitative Theory and Methods Department, Florida State University, George Washington University, the University of Illinois Urbana-Champaign, McMaster University, Northwestern University, the Urban Institute, the University of Virginia, Washington University in St. Louis, the University of Wisconsin Ag Econ Department, Zhongnan University, the University of Zurich, and participants at the 2022 International Association of Applied Econometrics, the 17th IZA & 4th IZA/CREST Conference: Labor Market Policy Evaluation, the 2023 Kansas Econometrics Workshop, the University of Michigan Workshop on TWFE Regressions, the 2022 Midwest Econometrics Group Conference, the 2022 North American Summer Meetings of the Econometric Society, and the 2022 Southern Economics Association Conference for helpful comments.

[†]John Munro Godfrey, Sr. Department of Economics, University of Georgia. carol.caetano@uga.edu

[‡]John Munro Godfrey, Sr. Department of Economics, University of Georgia. brantly.callaway@uga.edu

1 Introduction

Difference-in-differences is one of the most popular identification strategies in empirical work in economics. Textbook explanations of difference-in-differences often consider a setting with only two time periods and two groups (a treated group and an untreated group) under the assumption that the two groups have the same trend in untreated potential outcomes over time. In this setting, two-way fixed effects (TWFE) regressions have good properties (e.g., being robust to treatment effect heterogeneity) while being convenient to use in empirical work. The two most common extensions to this textbook setting in empirical work are (i) to settings with multiple periods and variation in treatment timing across units and (ii) to include covariates in the parallel trends assumption. Concerning multiple periods and variation in treatment timing (the first common extension), several recent papers discuss important weaknesses of using TWFE regressions to implement difference-in-differences identification strategies and propose alternative estimation strategies that address these weaknesses (de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), and Callaway and Sant'Anna (2021), among others).

In this paper, we focus on the inclusion of covariates in the parallel trends assumption (the second common extension of the textbook setting mentioned above). The aim of including covariates in the analysis is that the parallel trends assumption can be weakened to hold only locally among units with the same observed characteristics rather than holding at the aggregate level (Heckman, Ichimura, Smith, and Todd (1998) and Abadie (2005))—this can be especially important in settings where the covariates are not well-balanced between the treated and untreated group. To give an example, below we consider an application with state-level panel data and a state-level treatment. In this setting, empirical researchers often include covariates such as a state's population, median income, or region, with the intention of estimating the causal effects of the treatment by comparing paths of outcomes for treated and untreated states located near each other with similar populations and income levels.

We pay careful attention to the different types of covariates that can show up in the parallel trends assumption: time-varying covariates and/or time-invariant covariates. Interestingly, the econometrics literature on DiD and empirical applications of DiD have often thought about and used covariates in substantially different ways. In the econometrics literature, it is common to assume that the covariates are all time-invariant or, if there are time-varying covariates, to use a pre-treatment value of the time-varying covariates as a time-invariant covariate; see, for example, Abadie (2005), Bonhomme and Sauder (2011), Sant'Anna and Zhao (2020), and Callaway and Sant'Anna (2021). On the other hand, in empirical work, the most common way to include covariates is in the following TWFE regression

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + X_{it}'\beta + e_{it} \tag{1}$$

where θ_t is a time fixed effect, η_i is individual-level unobserved heterogeneity (i.e., an individual fixed effect), D_{it} is a binary treatment indicator, and X_{it} are time-varying covariates. In the TWFE regression in Equation (1), α is the coefficient of interest. It is sometimes interpreted as "the causal effect of the treatment" or, in the presence of treatment effect heterogeneity, α is often loosely interpreted as some kind of average treatment effect parameter. Being able to include covariates is one of the original main attractions of using a TWFE regression to implement a DiD identification strategy. For example, Angrist and Pischke (2008) write: "A second advantage of regression-DD is that it facilitates empirical work with regressors." Relative to the econometrics literature discussed above, one immediately noticeable difference with the TWFE regression is that it does not include time-invariant covariates; if time-invariant covariates enter the TWFE model in an analogous way to the time-varying covariates (i.e., with a time-invariant coefficient), then they will be absorbed into the unit fixed effect. This is a common explanation for not including time-invariant covariates in difference-in-differences applications.

In the current paper, we uncover several limitations of this TWFE regression. Although TWFE regressions with only two time periods are known to be robust to treatment effect heterogeneity under unconditional parallel trends, we show that TWFE regressions that rely on conditional parallel trends assumptions are susceptible to a number of problems *even in the case with only two time periods*. In particular, we show that TWFE regressions can include non-neglible misspecification bias terms for any of three reasons: (1) violations of certain linearity conditions on the model for untreated potential outcomes over time, (2) paths of untreated potential outcomes that depend on the level of time-varying covariates in addition to (or instead of) the change in the covariates over time, and (3) paths of untreated potential outcomes that depend on time-invariant covariates. Arguably, the first issue mentioned above is expected—similar conditions show up in the literature on interpreting cross-sectional regressions under unconfoundedness (Goldsmith-Pinkham, Hull, and Kolesár (2022), Blandhol, Bonney, Mogstad, and Torgovitsky (2022), and Hahn (2023)) and assuming a linear model for untreated potential outcomes is often a key step for motivating linear models for the outcome itself (see Angrist and Pischke (2008) for a number of examples).

Issues (2) and (3) are more subtle, and we refer to the bias that arises from these issues as *hidden linearity bias*. Unlike Issue (1), there is no analog of hidden linearity bias in cross-sectional settings. Hidden linearity bias arises because, to estimate the parameters in Equation (1), the researcher removes the unit fixed effect by transforming the model (e.g., via a within transformation or taking first differences). For example, consider the simple setting with only two time periods. In that case, α is estimated by taking first differences to eliminate the unit fixed effect. A consequence of taking first differences is that the covariates are also differenced. This means that the estimated model ultimately only controls for *changes* in the time-varying covariates. To see the possible negative implications here, consider again our example about state-level policies: controlling for the change in a state's population and/or median income may not end up controlling well for the level of these covariates (e.g., states with similar changes in population could have quite different levels of population) or for other time-invariant covariates such as region.[1]

An important question for empirical researchers is how much the bias discussed above matters in

[1]It is worth also mentioning an alternative framing of hidden linearity bias. Modern empirical work often views linear regressions as linear projections rather than interpreting linearity as being literally true in the sense of the model correctly specifying a linear conditional mean. In cross-sectional settings, the linear projection view can be attractive as linearity itself may be a strong assumption, but using the linear model in estimation is still convenient and has other good properties, such as being the best linear approximation to a possibly nonlinear conditional expectation function. In the DiD setting, however, the distinction between linear approximation and a correctly specified linear conditional mean is more important. A linear conditional mean rationalizes the transformations that eliminate the unit fixed effects and their effects on the functional form of the covariates. On the other hand, if Equation (1) is interpreted as a linear approximation, then the transformations used to eliminate the unit fixed effect effectively change the identification strategy from one where the researcher conditions on covariates in the parallel trends assumption in a general way to one where the researcher instead conditions on changes in time-varying covariates over time in the parallel trends assumption. See Section 2.2 for more details.

practice. It is hard to measure this bias directly because the terms involve differences between conditional expectations that are likely to be challenging to estimate nonparametrically in most applications. Instead, to address this question, we propose simple diagnostic tools to assess the sensitivity of TWFE regressions to hidden linearity bias. Our idea is to recast α from the TWFE regression as a re-weighting estimator (see Aronow and Samii (2016) and Chattopadhyay and Zubizarreta (2023) for related ideas in the context of unconfoundedness and cross-sectional data). We recover the "implicit regression weights" (these weights involve linear projections rather than conditional expectations, so they are straightforward to calculate), but instead of applying the weights to the outcomes (which would recover α), we apply the weights to the levels of time-varying covariates and to time-invariant covariates. If the implicit regression weights balance the levels of time-varying covariates and time-invariant covariates, this suggests that hidden linearity bias is likely to be small; on the other hand, if there is a high degree of imbalance after applying the implicit regression weights, this implies that α may be quite sensitive to violations of the linearity conditions that rationalize only conditioning on transformations in the time-varying covariates over time in the estimated model.

In applications where none of the three issues mentioned above occur, TWFE regressions deliver a weighted average of conditional ATTs. However, even in this scenario, TWFE regressions still suffer from additional drawbacks. First, the weights are non-transparent to the researcher—it is difficult to determine whether the TWFE weights are reasonable without making several auxiliary calculations. Second, it is possible that the weights on the conditional ATTs can be negative, even in the setting with two time periods. Negative weights is an issue that has been discussed extensively in the literature (see, for example, de Chaisemartin and D'Haultfœuille (2020) and Blandhol, Bonney, Mogstad, and Torgovitsky (2022)) and often is an indication of an unreasonable weighting scheme. Third, the weights have a "weight-reversal" property related to the one pointed out in Słoczyński (2022) under unconfoundedness with cross-sectional data, which amounts to systematically weighting common values of the covariates too little and weighting uncommon values of the covariates too much (see Section 3 for more details). How much the undesirable properties of these weights matter in practice depends crucially on how heterogeneous the conditional ATTs are—in settings where treatment effects vary substantially across covariates, the differences between the weighting schemes will matter more. In summary, the requirements for the TWFE regression in Equation (1) to estimate the ATT are stringent. They include (1) linearity of the path of untreated potential outcomes, (2) no dependence of parallel trends on time-invariant covariates, (3) a parallel trends condition that depends on time-varying covariates only through their change over time, and (4) homogeneous treatment effects across different values of the covariates.

We propose several new estimation strategies that do not suffer from *any* of the limitations of the TWFE regression discussed above. Our estimators build on recently developed estimation strategies in the DiD literature, particularly the AIPW estimators proposed in Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021).[2] For example, in applications with time-varying covariates, one very

[2] AIPW estimators involve estimating both an outcome regression model for untreated potential outcomes and a model for the propensity score. These estimators are doubly robust in the sense that they deliver consistent estimates of the ATT if either the outcome regression model or the propensity score model is correctly specified. They are also straightforward to adapt to settings where the researcher estimates the outcome regression and propensity score using modern machine learners.

simple strategy that works well for balancing covariates is to include both the change and levels of the time-varying covariates across periods in the outcome regression model and the propensity score model. We also describe how to reformulate AIPW estimators of the ATT under conditional parallel trends as re-weighting estimators. Combining these results with our previously mentioned TWFE diagnostics allows an empirical researcher to compare the covariate balancing properties of the leading estimators for the ATT during the "design phase" of a study, providing a means by which a researcher can assess alternative estimation strategies before using the outcome at all (Ho, Imai, King, and Stuart (2007), Rubin (2008), and Imbens and Rubin (2015)).

We conclude the paper by revisiting an application from Cheng and Hoekstra (2013) on the effects of stand-your-ground laws on homicides. We find that including time-varying covariates, such as a state's population and/or median income, in a TWFE regression balances the average of the within-transformed covariates but often does little to improve (and in some cases makes worse) covariate balance in terms of the levels of the same covariates or in terms of time-invariant covariates. Using our approach, covariate balance is substantially improved. Our estimates of the effects of stand-your-ground laws on homicides are mostly qualitatively similar to those reported in Cheng and Hoekstra (2013), though we find somewhat less evidence for stand-your-ground laws increasing homicides.

The outline of the paper is as follows. Section 2 introduces the notation and main assumptions that we use in the paper, provides some preliminary identification results, and then discusses models that can rationalize the conditional parallel trends assumption that we use throughout the paper. Section 3 discusses the limitations of TWFE regressions in the context of conditional parallel trends. Section 4 proposes simple diagnostics for assessing the extent of hidden linearity bias in TWFE regressions. Sections 2 to 4 focus on the baseline DiD setup with two time periods and two groups because many of our main insights can be seen in this setting. Section 5 extends these arguments to settings with multiple periods and variation in treatment timing across units. In Section 6, we propose alternative estimation strategies that circumvent TWFE regressions' limitations. Finally, in Section 7, we revisit an application from Cheng and Hoekstra (2013) on the effects of stand-your-ground laws on homicides.

2 Setup

In this section, we introduce the notation and main assumptions we use in the paper, provide some preliminary identification results, and then discuss models that can rationalize the conditional parallel trends assumption we use throughout the paper.

Notation: For much of the paper, we consider a setting with two time periods. We denote the time periods by t^*-1 and t^* and refer to these as the first time period and the second time period, respectively. We consider the canonical setting where no units are treated in the first period. Let D_i be a binary treatment indicator; because no units are treated in the first time period, we omit a time subscript on D_i. Let X_{it} denote a $k \times 1$ vector of time-varying covariates for unit i in time period t, and let Z_i denote an $l \times 1$ vector of time-invariant covariates. Let Y_{it} denote the observed outcome for unit i in time period t, and let $Y_{it}(1)$ and $Y_{it}(0)$ denote treated and untreated potential outcomes for unit i in time period t. We can relate observed outcomes to potential outcomes by $Y_{it^*} = D_i Y_{it^*}(1) + (1 - D_i) Y_{it^*}(0)$ and

$Y_{it^*-1} = Y_{it^*-1}(0)$. In other words, in the second time period, we observe treated potential outcomes for treated units and untreated potential outcomes for untreated units. In the first time period, because no units are treated yet, we observe untreated potential outcomes for all units.[3]

2.1 Identification

Following the vast majority of the difference-in-differences literature, we target identifying the average treatment effect on the treated (ATT), which is given by

$$ATT := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0)|D = 1]$$

We make the following assumptions:

Assumption 1 (Random Sampling)**.** *The observed data consists of* $\{Y_{it^*}, Y_{it^*-1}, X_{it^*}, X_{it^*-1}, Z_i, D_i\}_{i=1}^n$ *which are independent and identically distributed.*

Assumption 2 (Overlap)**.** *There exists some* $\epsilon > 0$ *such that* $\mathrm{P}(D = 1) > \epsilon$ *and* $\mathrm{P}(D = 1|X_{t^*}, X_{t^*-1}, Z) < 1 - \epsilon$.

Assumption 3 (Conditional Parallel Trends)**.**

$$\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 1] = \mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0].$$

Assumption 1 says that we have access to an iid sample of two periods of panel data. Assumption 2 is a standard version of an overlap condition that is often invoked in the DiD literature (e.g., Abadie (2005)) and in the treatment effects literature more broadly. In practice, it says that, for all treated units, there exist untreated units with the same characteristics. Assumption 3 says that the path of untreated potential outcomes is the same, on average, for the treated group as the untreated group after conditioning on time-varying covariates X_{t^*} and X_{t^*-1} and time-invariant covariates Z. This is the main conditional parallel trends assumption we use throughout the paper. Assumption 3 also implicitly rules out "bad controls" (i.e., that the time-varying covariates that show up in the parallel trends assumption are affected by the treatment); see Caetano, Callaway, Payne, and Sant'Anna (2022) for a detailed discussion of this case.

Under Assumptions 1 to 3, the ATT is identified, and, in particular, it is given by

$$ATT = \mathbb{E}[\Delta Y_{t^*}|D = 1] - \mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D = 0]\Big|D = 1\Big] \tag{2}$$

We provide this result formally in Proposition S1 in Appendix SA in the Supplementary Appendix but note here that it follows using the same arguments as in existing work on difference-in-differences such as Heckman, Ichimura, Smith, and Todd (1998) and Abadie (2005), up to separately keeping track of the time-varying and time-invariant covariates. The expression in Equation (2) says that the ATT can be recovered in our setting by comparing the mean path of outcomes experienced by the treated group

[3] The discussion above implicitly imposes a SUTVA assumption (i.e., that potential outcomes only depend on the treatment status of a unit itself) and a no-anticipation assumption (that pre-treatment outcomes are not affected by eventually participating in the treatment). These are standard conditions in the DiD literature. We discuss no-anticipation in more detail in Section 5. We also suppose throughout the paper that all expectations exist and take all statements conditional on covariates to hold almost surely.

relative to the path of outcomes that the treated group would have experienced if it had not participated in the treatment. Under the conditional parallel trends assumption, the latter counterfactual path of untreated potential outcomes can be recovered by taking the path of outcomes conditional on time-varying and time-invariant covariates for the untreated group and then averaging it over the distribution of covariates for the treated group (this step allows for the distribution of time-varying and time-invariant covariates to be systematically different for the treated group relative to the untreated group). We note that the conditional expectation in the second term, $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D = 0]$ may, in many applications, be challenging to estimate though we defer these estimation issues to Section 6.

2.2 Models that Rationalize Parallel Trends Assumptions

Before proceeding to our results on interpreting TWFE regressions, this section discusses models that can rationalize parallel trends assumptions conditional on covariates. This exercise provides some intuition about the types of conditions that allow us to causally interpret α from the TWFE regression (see Assumption 4 below) and the types of problems that can arise when these conditions are not satisfied. Moreover, the more general models have a similar flavor as the alternative estimation strategies we propose later in the paper.

A natural starting point is to connect the unconditional parallel trends assumption to a two-way fixed effects model for untreated potential outcomes (as in, for example, Blundell and Costa Dias (2009), Gardner, Thakral, Tô, and Yap (2023), and Borusyak, Jaravel, and Spiess (2024)); that is,

$$Y_{it}(0) = \theta_t + \eta_i + e_{it},$$

where θ_t is a time fixed effect, η_i is time-invariant unobserved heterogeneity (i.e., an individual fixed effect), and e_{it} are idiosyncratic, time-varying unobservables.[4] The unconditional parallel trends assumption holds in this model for untreated potential outcomes under the condition that $\mathbb{E}[\Delta e_t|D = 1] = \mathbb{E}[\Delta e_t|D = 0]$ for all time periods. However, this setting allows η_i to be distributed differently between the treated and untreated group and does not impose any modeling assumptions on treated potential outcomes. Notice that this argument relies crucially on the additive separability of η_i.

As discussed above, the econometrics literature on difference-in-differences often considers the case where the parallel trends assumption is invoked conditional on time-invariant covariates. In that case, the analogous model for untreated potential outcomes is given by

$$Y_{it}(0) = g_t(Z_i) + \eta_i + e_{it},$$

where the distribution of η can vary across treated and untreated groups (as well as vary with Z) and the conditional parallel trends assumption holds under the condition that $\mathbb{E}[\Delta e_t|Z, D = 1] = \mathbb{E}[\Delta e_t|Z, D = 0]$ (see, for example, Heckman, Ichimura, and Todd (1997) for a discussion of this kind of model).[5] As in the unconditional model above, the key condition on this model is the additive separability of η, while the effect of the time-invariant covariate Z on untreated potential outcomes can be quite general

[4] To be clear on notation, throughout the paper, we use θ_t and η_i (and similar notation) as generic notation for time and unit fixed effects, and these are not the same as the corresponding terms in Equation (1).

[5] To see this, notice that $\mathbb{E}[\Delta Y_t(0)|Z, D = 1] = g_t(Z) - g_{t-1}(Z) = \mathbb{E}[\Delta Y_t(0)|Z, D = 0]$ which implies that conditional parallel trends holds.

and vary over time (in fact, if its effect does not vary over time, there is no need to include Z in the parallel trends assumption as its influence on the outcome is absorbed into the unit fixed effect η).

In this setup, the main challenge for implementing the identification strategy is estimating $g_t(z)$. A natural way to proceed is to parameterize the model as

$$Y_{it}(0) = \theta_t + Z_i'\delta_t + \eta_i + e_{it}.$$

which further implies that $ATT = \mathbb{E}[\Delta Y_t|D = 1] - \left(\tilde{\theta}_t - \mathbb{E}[Z|D = 1]'\tilde{\delta}_t\right)$ where $\tilde{\theta}_t := \theta_t - \theta_{t-1}$ and $\tilde{\delta}_t := (\delta_t - \delta_{t-1})$ which can both be consistently estimated from the regression of ΔY_t on Z using only observations from the untreated group.[6]

Given the discussion above with time-invariant covariates in the parallel trends assumption, when, instead, the parallel trends assumption is invoked conditional on both time-varying covariates and time-invariant covariates, the natural motivating model is

$$Y_{it}(0) = g_t(Z_i, X_{it}) + \eta_i + e_{it},$$

which implies that

$$\Delta Y_{it}(0) = g_t(Z_i, X_{it}) - g_{t-1}(Z_i, X_{it-1}) + \Delta e_{it}.$$

The same sorts of arguments as above imply that the conditional parallel trends assumption in Assumption 3 holds given this model for untreated potential outcomes. Similar to the previous case, the main practical challenge is that $g_t(z, x_t)$ is likely to be difficult to estimate nonparametrically, and a natural way to parameterize this model is

$$Y_{it}(0) = \theta_t + Z_i'\delta_t + X_{it}'\beta_t + \eta_i + e_{it}. \tag{3}$$

Taking first differences implies that

$$\Delta Y_{it}(0) = \tilde{\theta}_t + Z_i'\tilde{\delta}_t + \Delta X_{it}'\beta_t + X_{it-1}'\tilde{\beta}_t + \Delta e_{it}, \tag{4}$$

where $\tilde{\beta}_t := (\beta_t - \beta_{t-1})$. In this model, the path of untreated potential outcomes can depend on time-invariant covariates, the level of time-varying covariates, and how time-varying covariates change over time. Because untreated potential outcomes are observed for the untreated group, the parameters in the model above can be recovered from a regression of the change in outcomes over time on time-invariant covariates, the change in time-varying covariates, and the level of the time-varying covariates in the pre-treatment period using data from the untreated group.

Equation (4) provides a natural baseline model for the path of untreated potential outcomes. It provides a direct way to implement the conditional parallel trends assumption in Assumption 3 if one is willing to assume linearity. However, we show below that interpreting α in the TWFE regression, even as a weighted average of causal effect parameters, requires additional restrictions on this model—

[6] This is closely related to regression adjustment estimators (see, for example, Heckman, Ichimura, Smith, and Todd (1998), Imbens and Wooldridge (2009), and Sant'Anna and Zhao (2020) for related discussion).

particularly, that β_t and δ_t do not vary over time. In this case, Equation (4) reduces to

$$\Delta Y_{it}(0) = \tilde{\theta}_t + \Delta X_{it}\beta + \Delta e_{it} \tag{5}$$

so that the path of untreated potential outcomes no longer depends on time-invariant covariates or the level of time-varying covariates, embedding substantive restrictions on how covariates can affect paths of untreated potential outcomes over time and, effectively, changing the identification strategy. That the TWFE regression relies on auxiliary conditions such as these is an important drawback. Heuristically, the hidden linearity bias that we emphasize below comes from relying on a restricted model for untreated potential outcomes like the one in Equation (5) when the correct model is a more flexible like the one in Equation (4). In contrast to the TWFE regression, the alternative estimators that we propose in Section 6 explicitly include levels and changes in time-varying covariates as well as time-invariant covariates and, hence, do not suffer from hidden linearity bias.

3 Interpreting TWFE Regressions

This section considers how to interpret α in the TWFE regression in Equation (1). We continue to focus on the setting with two time periods where no one is treated in the first time period and where some, but not all, units become treated in the second time period. This is a favorable setting for TWFE regressions as it does not introduce well-known problems related to using already-treated units in the comparison group (de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021)).

For interpreting the TWFE regression, many of our results involve linear projections. Let $\mathrm{L}(D|\Delta X_{t^*})$ denote the (population) linear projection of D on ΔX_{t^*}.[7] That is,

$$\mathrm{L}(D|\Delta X_{t^*}) := \Delta X_{t^*}' \mathbb{E}[\Delta X_{t^*} \Delta X_{t^*}']^{-1} \mathbb{E}[\Delta X_{t^*} D] = \Delta X_{t^*}' \gamma$$

Similarly, for $d \in \{0, 1\}$, define

$$\mathrm{L}_d(\Delta Y_{t^*}|\Delta X_{t^*}) := \Delta X_{t^*}' \mathbb{E}[\Delta X_{t^*} \Delta X_{t^*}' | D = d]^{-1} \mathbb{E}[\Delta X_{t^*} \Delta Y_{t^*} | D = d] = \Delta X_{t^*}' \beta_d$$

which is the linear projection of ΔY_{t^*} on ΔX_{t^*} for the treated group (when $d = 1$) and for the untreated group (when $d = 0$), respectively.

Notice that, with exactly two periods, it is helpful to equivalently re-write Equation (1) as

$$\Delta Y_{it^*} = \alpha D_i + \Delta X_{it^*}' \beta + \Delta e_{it^*} \tag{6}$$

We view Equation (6) as a linear projection model rather than a linear conditional expectation/structural model, allowing for heterogeneous treatment effects. Our interest in this section is in determining what kind of conditions are required to interpret α as the ATT or at least as a weighted average of some un-

[7] All of the linear projections in this section include an intercept—this involves a slight abuse of notation where, for example, we augment ΔX_{t^*} so that it includes an intercept in addition to the change in time-varying covariates over time. Similarly, we also slightly abuse notation in Equation (6) by taking β to include an extra parameter in its first position corresponding to the intercept. For all the results below that involve linear projections, we assume that they are well-defined. This typically involves an extra assumption that a second-moment matrix, such as $\mathbb{E}[\Delta X_{t^*} \Delta X_{t^*}']$, is positive definite. We provide the vast majority of our results in the paper in terms of population (rather than sample) quantities. For expressions that only involve means and linear projections (which applies to many of our results below), analogous results hold for the corresponding sample quantities.

derlying treatment effect parameters. To start with, using well-known Frisch-Waugh-Lovell arguments, notice that we can write

$$\alpha = \mathbb{E}\left[\frac{(D - \mathrm{L}(D|\Delta X_{t^*}))\Delta Y_{t^*}}{\mathbb{E}[(D - \mathrm{L}(D|\Delta X_{t^*}))^2]}\right] \tag{7}$$

Next, we provide our first main result on interpreting α in terms of underlying causal effect parameters along with some additional bias terms.

Theorem 1. *Under Assumptions 1 to 3, α from Equation (6) can be expressed as*

$$\begin{aligned}
\alpha = {} & \mathbb{E}\Big[w(\Delta X_{t^*})ATT(X_{t^*}, X_{t^*-1}, Z)\Big|D=1\Big] \\
& + \mathbb{E}\Big[w(\Delta X_{t^*})\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=0]\Big) && \text{(A)} \\
& \qquad + \Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=0] - \mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=0]\Big) && \text{(B)} \\
& \qquad + \Big(\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=0] - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big\}\Big|D=1\Big] && \text{(C)}
\end{aligned}$$

where

$$w(\Delta X_{t^*}) := \frac{1 - \mathrm{L}(D|\Delta X_{t^*})}{\mathbb{E}[(1 - \mathrm{L}(D|\Delta X_{t^*}))|D=1]}$$

which are weights that have the following properties: (i) $\mathbb{E}[w(\Delta X_{t^*})|D=1] = 1$ and (ii) $w(\Delta X_{t^*})$ can be negative if there exist values of ΔX_{t^*} among the treated group such that $\mathrm{L}(D|\Delta X_{t^*}) > 1$.

The result in Theorem 1 indicates that α is equal to a weighted average of underlying conditional ATTs (we discuss the nature of the weights in more detail below) plus several undesirable bias terms.[8] That $w(\Delta X_{t^*})$ has mean one suggests that these bias terms, in general, should be a first-order concern for empirical researchers. This contrasts several recent papers on interpreting regressions in different contexts where the regression coefficient ends up including bias terms but where the weights have mean zero (e.g., Sun and Abraham (2021), de Chaisemartin and D'Haultfœuille (2023), and Goldsmith-Pinkham, Hull, and Kolesár (2022)).

The bias in Term (A) arises because the regression in Equation (6) does not include time-invariant covariates. This term suggests that failing to include time-invariant covariates in the TWFE regression when the path of untreated potential outcomes actually depends on time-invariant covariates undesirably contributes to how α is calculated. In our application to stand-your-ground laws, this type of bias term could arise by failing to include state-level time-invariant covariates that affect the path of untreated potential outcomes. A leading example would be failing to include a region indicator if trends in homicides (absent the policy) are different across different regions of the country.[9]

[8] The proof of Theorem 1 (especially the parts concerning the weights) is mechanically related to work on interpreting cross-sectional regressions under the assumption of unconfoundedness or other related settings (Angrist (1998), Aronow and Samii (2016), Słoczyński (2022), Ishimaru (2024), Goldsmith-Pinkham, Hull, and Kolesár (2022), Blandhol, Bonney, Mogstad, and Torgovitsky (2022), and Hahn (2023)). The hidden linearity bias terms (discussed in detail below) that show up in the expression for α are specific to the DiD setting that we consider here.

[9] As a point of clarification, it is common in empirical difference-in-differences applications that use state-level data to include region-time fixed effects (i.e., to include a region indicator with a time-varying coefficient). This partially, though not entirely, addresses the issues discussed in this section; see Wooldridge (2021, in particular, Equation 5.15 and the discussion in Section 5.2). Perhaps a better example comes from DiD applications that use individual-level data where it is less common to include time-invariant covariates with time-varying coefficients in the TWFE regression. For example, it

Term (B) arises when the path of untreated potential outcomes depends on the levels of time-varying covariates instead of only on the change in covariates over time. Together, we refer to Terms (A) and (B) as *hidden linearity bias*. To motivate this, when a researcher estimates the model in Equation (6), they likely realize that they are making linearity assumptions; on the other hand, it is also well-known that regressions can still have good properties in this setting, such as being the best linear approximation of a possibly nonlinear conditional expectation. This term indicates that an additional implication of linearity is that the estimated model, as a by-product of differencing out the unit fixed effect, only ends up including the change in the covariate over time. In the case where linearity is actually correct, then there is no issue here. Still, if the model is viewed as an approximation, then an undesirable implication of the TWFE model is that the researcher ends up only controlling for the change in the covariates over time and does not explicitly control for the levels of the covariates. In the stand-your-ground application, one of the covariates that Cheng and Hoekstra (2013) consider is a state's population. Presumably, the idea is to compare paths of outcomes for treated states with a high population to paths of outcomes for untreated states that also have a high population (or, likewise, a middle or low population). However, an implication of the TWFE regression is that, effectively, the researcher instead compares states that have similar population changes over time. And, of course, states with similar changes in population can be quite different in terms of the level of their populations.

Term (C) is non-zero when the conditional expectation of the change in untreated potential outcomes conditional on covariates is nonlinear in the change in covariates over time. This type of linearity condition is the one that researchers would likely suspect to be implicit in the TWFE regression. A similar term shows up in cross-sectional settings with different papers discussing various conditions under which it is equal to zero (Angrist (1998), Blandhol, Bonney, Mogstad, and Torgovitsky (2022), and Hahn (2023)). In general, this term is non-zero, though it may be reasonable to hope that the conditional expectation is close to being linear in many cases.

Next, we provide an additional assumption that serves to eliminate the bias terms discussed above.

Assumption 4 (Additional Assumptions to Rule Out Bias Terms)**.**

(A) The path of untreated potential outcomes does not depend on time-invariant covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D=0] = \mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, D=0]$

(B) The path of untreated potential outcomes only depends on the change in time-varying covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, D=0] = \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}, D=0]$

(C) The path of untreated potential outcomes is linear in the change in time-varying covariates. That is, $\mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}, D=0] = \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$

Theorem 2. *Under Assumptions 1 to 3, and if, in addition, Assumption 4 also holds, then*

$$\alpha = \mathbb{E}\Big[w(\Delta X_{t^*}) ATT(X_{t^*}, X_{t^*-1}, Z)\Big| D=1\Big]$$

is uncommon in labor economics to include a person's race as a covariate in a TWFE regression because it does not vary over time despite the fact that it seems likely that the path of many labor market outcomes depends on race.

where the weights $w(\Delta X_{t^*})$ *are the same ones defined in Theorem 1. If, in addition,* $ATT(X_{t^*}, X_{t^*-1}, Z)$ *is constant across all values of* (X_{t^*}, X_{t^*-1}, Z)*, then*

$$\alpha = ATT$$

Theorem 2 provides sufficient conditions for α from Equation (6) to be equal to a weighted average of conditional ATTs under the conditional parallel trends assumption in Assumption 3. The proof of Theorem 2 is provided in Appendix A. The intuition for the result is that the conditions in Assumption 4 together imply that

$$\mathbb{E}[\Delta Y_{t^*} | X_{t^*}, X_{t^*-1}, Z, D = 0] = \mathrm{L}_0(\Delta Y_{t^*} | \Delta X_{t^*})$$

This is sufficient for the bias terms in Theorem 1 to be equal to 0, and, thus, α is equal to a weighted average of $ATT(X_{t^*}, X_{t^*-1}, Z)$.

The result in Theorem 2 suggests several potential issues with the TWFE regressions as in Equation (6). First, the additional conditions in Assumption 4 are likely to be strong in many applications, and, perhaps more importantly, empirical researchers do not typically acknowledge that these assumptions are embedded in the TWFE estimation strategies that are commonly used in empirical work.

Second, even if one is willing to maintain the additional assumptions in Assumption 4, α from the TWFE regression is still hard to interpret for several reasons. To explain these reasons, first notice that maintaining these additional assumptions in Assumption 4 implies that all of the weights, conditional ATTs, and linear projections in the first part of Theorem 2 are identified and directly estimable. The first possible issue with interpreting α is that, although the weights have mean one, it is possible to have negative weights for some values of $ATT(X_{t^*}, X_{t^*-1}, Z)$. This can happen for values of the covariates among the treated group where $\mathrm{L}(D|\Delta X_{t^*}) > 1$. This is possible because $\mathrm{L}(D|\Delta X_{t^*})$ is a linear projection of a binary treatment on ΔX_{t^*}, which is not restricted to be between 0 and 1. In the literature, negative weights have often been emphasized as being particularly problematic (see, for example, de Chaisemartin and D'Haultfœuille (2020) and Blandhol, Bonney, Mogstad, and Torgovitsky (2022)). Blandhol, Bonney, Mogstad, and Torgovitsky (2022) refer to the class of parameters that can be written as a weighted average of conditional treatment effects where the weights are non-negative to be *weakly causal*; and, for example, in the presence of negative weights, it is possible to come up with examples where $ATT(X_{t^*}, X_{t^*-1}, Z)$ is positive for all values of the covariates, but α could be negative due to the weighting scheme. In empirical work, estimating $\mathrm{L}(D|\Delta X_{t^*})$ and checking if there are negative weights is straightforward. Another issue is that, even if there are no negative weights, the weights have a "weight-reversal" property (we adapt this terminology from Słoczyński (2022) who points out a related weighting issue in the context of unconfoundedness and cross-sectional data). Notice that the ideal weighting scheme would be for $w(\Delta X)$ to be uniformly equal to one—in which case, $\alpha = ATT$. Relative to this natural baseline, the weights in Theorem 2 indicate that α puts too much weight on conditional ATTs for values of the covariates that are relatively uncommon among the treated group relative to the untreated group and puts too little weight on conditional ATTs for values of the covariates that are relatively common among the treated group relative to the untreated group.

Finally, if, in addition to all the previous conditions, conditional ATTs are constant across different values of the covariates, then α will be equal to the ATT. This is a treatment effect homogeneity condition with respect to the covariates. It *is* somewhat weaker than individual-level treatment effect homogeneity, and it allows for treatment effects to be systematically different for treated units relative to untreated units. Instead, what it says is that, for the treated group, treatment effects cannot be systematically different across different values of the covariates. However, this assumption is likely to be extremely strong in most economic applications, and it is not commonly discussed/considered in empirical work.

These results differ greatly from our earlier result on identifying the ATT in Equation (2). That result did not require any of the additional assumptions in Assumption 4.

Remark 1 (Alternative conditions on the propensity score for interpreting α)**.** One can also show that α is equal to a weighted average of conditional ATTs under restrictions on the propensity score (rather than restrictions on $\mathbb{E}[\Delta Y_t(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]$ as above); namely, $\mathrm{P}(D = 1|X_{t^*}, X_{t^*-1}, Z) = \mathrm{L}(D|\Delta X_{t^*})$. See Angrist (1998), Aronow and Samii (2016), and Słoczyński (2022) for results along these lines with cross-sectional data under unconfoundedness. In Appendix SA.3 in the Supplementary Appendix, we argue that, in the panel data context that we consider, linearity conditions are less plausible on the propensity score than on the outcome models discussed above. Moreover, some leading cases where the propensity score would be linear by construction in cross-sectional settings do not apply in our setting. See the Supplementary Appendix for more details.

Remark 2 (Comparison to conditions for other estimation strategies)**.** Interestingly, very similar restrictions as the ones discussed in Assumption 4 arise in some recently proposed "heterogeneity robust" versions of difference-in-differences. For example, the imputation approaches proposed in Gardner, Thakral, Tô, and Yap (2023) and Borusyak, Jaravel, and Spiess (2024) involve estimating the model $Y_{it}(0) = \theta_t + \eta_i + X_{it}'\beta + e_{it}$ (see Gardner, Thakral, Tô, and Yap (2023, Eq. (7)) and Borusyak, Jaravel, and Spiess (2024, Eq. (5))) which, in the two-period context considered here, implicitly uses the assumption that $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0] = \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$—the same condition as implied by Assumption 4. Alternatively, the regression adjustment version of Callaway and Sant'Anna (2021) implicitly uses the assumption that $\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0] = \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*-1}, Z)$—besides linearity, this condition effectively says that the path of untreated potential outcomes does not depend on X_{t^*} once one controls for X_{t^*-1} and Z. The estimators we propose below do not include either of these types of auxiliary assumptions. See Appendix SC.1 in the Supplementary Appendix for a more detailed discussion along these lines.

4 Covariate Balance Diagnostics

This section develops diagnostic tools for assessing covariate balance inherited from two different estimation strategies. The first part of this section considers diagnostics for the TWFE regression in Equation (6). The second part of this section considers diagnostics of augmented inverse propensity score weighting (AIPW) estimators of the ATT under conditional parallel trends along the lines of the alternative estimators we propose later in the paper.

4.1 TWFE Diagnostics

Theorem 1 highlights several potential sources of bias from using the TWFE regression in Equation (6). In this section, we consider the problem of quantitatively assessing *how much* these bias terms matter in practice. This is not an easy task as the conditional expectations in Terms (A)-(C) of Theorem 1 are challenging to estimate without imposing additional functional form assumptions. The misspecification bias terms in Terms (A)-(C) amount to violations of linearity that come from differences between $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D = 0]$ and $\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$. Below, we propose a simple approach to assess the sensitivity of α from the TWFE regression to possible violations of this linearity condition.

To motivate the results in this section, notice that if we could find "balancing weights" $\vartheta_0(X_{t^*}, X_{t^*-1}, Z)$ that re-weight the untreated group such that it has the same distribution of (X_{t^*}, X_{t^*-1}, Z) as the treated group, then it would be the case that

$$\begin{aligned}\mathbb{E}[\Delta Y_{t^*}(0)|D = 1] &= \mathbb{E}\Big[\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]\Big|D = 1\Big] \\ &= \mathbb{E}\Big[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\mathbb{E}[\Delta Y_{t^*}(0)|X_{t^*}, X_{t^*-1}, Z, D = 0]\Big|D = 0\Big] \\ &= \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\Delta Y_{t^*}|D = 0]\end{aligned}$$

and, therefore, that we could recover $ATT = \mathbb{E}[\Delta Y_{t^*}|D = 1] - \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)\Delta Y_{t^*}|D = 0]$. In other words, if we could balance the distribution of covariates for the untreated group relative to the treated group, then we could recover the path of untreated potential outcomes for the treated group by looking at the mean path of outcomes for the untreated group after it has been re-weighted to have the same distribution of covariates as the treated group. These sorts of balancing weights are related to a large number of weighting estimators. For example, the weights from propensity score re-weighting satisfy this property (Rosenbaum and Rubin (1983) and Hirano, Imbens, and Ridder (2003)). Other examples include entropy balancing (Hainmueller (2012)) and covariate balancing propensity score (Imai and Ratkovic (2014)).

One useful property of balancing weights that we exploit heavily below is that they balance functions of the covariates across groups. That is, for some function of the covariates g,

$$\mathbb{E}[g(X_{t^*}, X_{t^*-1}, Z)|D = 1] = \mathbb{E}[\vartheta_0(X_{t^*}, X_{t^*-1}, Z)g(X_{t^*}, X_{t^*-1}, Z)|D = 0] \tag{8}$$

Equation (8) can be used to validate particular balancing estimators. For example, if one estimates the propensity score in the propensity score re-weighting balancing weights using a logit or probit model, then it is fairly common for researchers to check that the analog of Equation (8) actually does balance observed covariates across groups. For approaches like entropy balancing and the covariate balancing propensity score that explicitly balance certain functions of the covariates, it is relatively common to check if the weights balance higher-order terms or interactions that were not explicitly balanced.

Returning to α from the TWFE regression, one useful insight is that it can be written as a re-weighting estimator. Starting from the expression in Equation (7), it follows from the law of iterated expectations that

$$\alpha = \mathbb{E}[w_1(\Delta X_{t^*})\Delta Y_{t^*}|D = 1] - \mathbb{E}[w_0(\Delta X_{t^*})\Delta Y_{t^*}|D = 0] \tag{9}$$

where

$$w_1(\Delta X_{t^*}) = \frac{\pi(1 - \mathrm{L}(D|\Delta X_{t^*}))}{\mathbb{E}[(D - \mathrm{L}(D|\Delta X_{t^*}))^2]} \quad \text{and} \quad w_0(\Delta X_{t^*}) = \frac{(1-\pi)\mathrm{L}(D|\Delta X_{t^*})}{\mathbb{E}[(D - \mathrm{L}(D|\Delta X_{t^*}))^2]}$$

where we define $\pi := \mathrm{P}(D = 1)$. We refer to the weights $w_d(\Delta X_{t^*})$ as *implicit regression weights* below. Notice that these weights are simple to calculate, as the most complicated terms are linear projections. And, to be clear, applying these weights to the path of outcomes separately for the treated group and untreated group recovers α from the TWFE regression. Building on the intuition for weighting estimators discussed earlier in this section, the diagnostics we propose in this section come from applying these weights to functions of the covariates to check how well the weights balance the covariates across groups. In the context of cross-sectional data under the assumption of unconfoundedness, Aronow and Samii (2016) and Chattopadhyay and Zubizarreta (2023) derive related weights and discuss a number of properties of these types of weights. For our purposes, the most notable property is that these weights will balance (in mean) the covariates that show up in the regression; thus, in our case, they will balance ΔX_{t^*} across groups. See Proposition S3 in the Supplementary Appendix for a more detailed explanation of why this is the case.[10] Although the weights balance the mean of ΔX_{t^*}, they do not necessarily balance the distribution/means of the levels of time-varying covariates (that is, X_{t^*} or X_{t^*-1}) or of time-invariant covariates Z. Thus, our strategy below is to assess the sensitivity of the TWFE regression to violations of linearity by comparing terms such as

$$\begin{aligned} & \mathbb{E}[w_1(\Delta X_{t^*})X_{t^*}|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})X_{t^*}|D=0] \\ \text{or} \quad & \mathbb{E}[w_1(\Delta X_{t^*})X_{t^*-1}|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})X_{t^*-1}|D=0] \\ \text{or} \quad & \mathbb{E}[w_1(\Delta X_{t^*})Z|D=1] \quad \text{to} \quad \mathbb{E}[w_0(\Delta X_{t^*})Z|D=0] \end{aligned}$$

If these terms are all close to each other, it suggests that the implicit regression weights effectively balance time-invariant covariates and the levels of time-varying covariates between the treated group and untreated group, and, hence, that α from the TWFE regression is not much affected by hidden linearity bias. On the other hand, if these terms are not close to each other, it suggests that α from the TWFE regression could be sensitive to violations of linearity. And, to be precise, if linearity is exactly correct, then failing to balance time-invariant covariates and the levels of time-varying covariates is a non-issue; however, if the researcher mainly invokes a linear model for simplicity (and/or because of its good approximation properties for conditional expectations), then large differences in the terms above would suggest that the TWFE could perform poorly with respect to "controlling for" time-invariant covariates and levels of time-varying covariates.

[10] It is also worth pointing out that, although the weights balance the mean of ΔX_{t^*}, they balance with respect to a different covariate profile than the one for the ATT; for example, the correct weighting scheme for the ATT would have $w_1(\Delta X_{t^*}) = 1$ and $w_0(\Delta X_{t^*})$ to be balancing weights such that applying them to the untreated group would re-weight it to have the same distribution of ΔX_{t^*} as for the treated group. Neither of these holds, and this suggests that, even if the implicit TWFE weights did balance the distribution of covariates, it would still not be sufficient for α to be equal to the ATT; see Chattopadhyay and Zubizarreta (2023) for an extensive discussion of this property of the weights and an explicit expression for the covariate profile for which the weights balance.

Relationship to Strategies in Empirical Work

The ideas presented in this section are broadly similar to the idea of using covariates as outcomes to assess balance, which is relatively common in empirical work in economics (see Pei, Pischke, and Schwandt (2019) for a detailed discussion of this strategy). This approach is not feasible for assessing balance with respect to covariates that do not vary over time or for time-varying covariates that are included in the TWFE regression. Alternatively, some papers check for balance in terms of pre-treatment characteristics (see, for example, Goodman-Bacon and Cunningham (2019, Table 3)). The working paper version of Goodman-Bacon (2021) discusses comparing the averages of time-varying covariates (including levels) for early, late, and never-treated groups (see, in particular, pp. 20-21 at http://goodman-bacon.com/pdfs/ddtiming.pdf as well as Almond, Hoynes, and Schanzenbach (2011) and Bailey and Goodman-Bacon (2015) as examples of empirical work using this sort of strategy). Relative to these strategies, a main advantage of the weighting strategy discussed in this section is that one can directly use the implicit regression weights from a main TWFE specification used in a particular application and assess balance for functions of covariates that are included in the model.

4.2 AIPW Diagnostics

The main class of estimators that we suggest as alternatives to the TWFE regression are augmented inverse propensity score weighting (AIPW) estimators. These estimators involve estimating both an outcome regression model and a model for the propensity score. In this section, we introduce the particular AIPW estimands that we consider. Following a similar motivation as in the previous section for TWFE regressions, we recast our AIPW approach as a weighting estimator. Then, we can apply these implicit AIPW weights to the covariates or functions of the covariates, thereby allowing us to assess how well this estimation strategy balances covariate distributions for the treated and untreated groups.[11] As a step towards developing an AIPW estimator, it is a straightforward extension of the identification results in Equation (2) to show that[12]

$$ATT = \mathbb{E}\left[\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\Big|D=1\right] - \mathbb{E}\left[w_0^{aipw}\left(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\right)\Big|D=0\right] \tag{10}$$

[11] The results in this section build on several recent papers that have shown that, at least in some important cases, ostensible outcome models can often be reinterpreted as weighting estimators; these include Robins, Sued, Lei-Gomez, and Rotnitzky (2007), Kline (2011), and Chattopadhyay and Zubizarreta (2023). The results in this section are most similar to the ones in Chattopadhyay and Zubizarreta (2023). That paper considers the case with cross-sectional data under unconfoundedness and provides a general result (Theorem 3) on re-formulating AIPW estimators as weighting estimators in that context. Besides differences related to DiD relative to cross-sectional settings, there are still some differences between our results in this section and their results. They mainly focus on ATE rather than ATT, though they briefly mention how their results apply to the ATT in their Supplementary Appendix (see, in particular, the discussion on p.4). That said, their expression for the weights differs conceptually from ours as our results depend to a large extent on linear projections of odds ratios rather than adjusted differences in means of the covariates across groups. In addition, our weights are slightly numerically different from the weights that we get when we use the `lmw` R package (Chattopadhyay, Greifer, and Zubizarreta (2023)) in settings where the results are comparable. Finally, we provide a direct proof delivering the implicit AIPW weights for the ATT rather than deriving it as a byproduct of a general result.

[12] Given the expression for ATT in Equation (2), deriving Equation (10) follows from the same line of argument as other AIPW estimators (Robins, Rotnitzky, and Zhao (1994), Słoczyński and Wooldridge (2018), and Sant'Anna and Zhao (2020)): the first term is equal to the ATT by Equation (2), and the second term is equal to zero by the law of iterated expectations.

where[13]

$$w_0^{aipw} := \frac{\varpi_0^{aipw}}{\mathbb{E}[\varpi_0^{aipw}|D=0]} \qquad \text{with} \qquad \varpi_0^{aipw} := \frac{(1-\pi)p(X_{t^*}, X_{t^*-1}, Z)}{\pi\big(1-p(X_{t^*}, X_{t^*-1}, Z)\big)}$$

What is interesting and useful about this expression for the ATT arises in estimation. To estimate the ATT based on this expression requires first-step estimation of $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]$ and $p(X_{t^*}, X_{t^*-1}, Z)$. In this section, we specify a linear working model, $\mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)$, for $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]$. Similarly, let $\tilde{p}(X_{t^*}, X_{t^*-1}, Z)$ denote a working model for $p(X_{t^*}, X_{t^*-1}, Z)$; leading choices include a logit or probit model, but there are other possibilities.[14] We allow for the possibility that either or both of these models are misspecified. Given these working models for the outcome regression and the propensity score, we define

$$\widetilde{ATT} = \mathbb{E}\left[\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)\Big|D=1\right] - \mathbb{E}\left[\tilde{w}_0^{aipw}\big(\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)\big)\Big|D=0\right] \tag{11}$$

where

$$\tilde{w}_0^{aipw} := \frac{\tilde{\varpi}_0^{aipw}}{\mathbb{E}[\tilde{\varpi}_0^{aipw}|D=0]} \qquad \text{with} \qquad \tilde{\varpi}_0^{aipw} := \frac{(1-\pi)\tilde{p}(X_{t^*}, X_{t^*-1}, Z)}{\pi\big(1-\tilde{p}(X_{t^*}, X_{t^*-1}, Z)\big)}$$

$\widetilde{ATT}$ is an AIPW working model estimand corresponding to the expression for ATT in Equation (10) but with working models replacing the outcome regression and propensity score. The sample analog of $\widetilde{ATT}$ is doubly robust, in the sense that $\widetilde{ATT} = ATT$ if either $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] = \mathrm{L}_0(\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z)$ or $p(X_{t^*}, X_{t^*-1}, Z) = \tilde{p}(X_{t^*}, X_{t^*-1}, Z)$, (i.e., if either the outcome regression model or the propensity score model is correctly specified). The following proposition shows that $\widetilde{ATT}$ can be written as a re-weighting estimator.

Proposition 1. *To conserve on notation, let $X = (X_{t^*}, X_{t^*-1}, Z)$. Define γ_0 as the linear projection coefficient from projecting $p(X)/\big(1-p(X)\big)$ on X; similarly define $\tilde{\gamma}_0$ as the linear projection coefficient from projecting $\tilde{p}(X)/\big(1-\tilde{p}(X)\big)$ on X. Then, under Assumptions 1 to 3,*

$$\widetilde{ATT} = \mathbb{E}\left[\vartheta_1^{aipw}\Delta Y_{t^*}\Big|D=1\right] - \mathbb{E}\left[\vartheta_0^{aipw}\Delta Y_{t^*}\Big|D=0\right]$$

where ϑ_1^{aipw} and ϑ_0^{aipw} are weights which are defined as

$$\vartheta_1^{aipw} := 1 \qquad \textit{and} \qquad \vartheta_0^{aipw} := \tilde{w}_0^{aipw} + \frac{\gamma_0' X}{\mathbb{E}[\gamma_0' X|D=0]} - \frac{\tilde{\gamma}_0' X}{\mathbb{E}[\tilde{\gamma}_0' X|D=0]}$$

where $\mathbb{E}[\vartheta_1^{aipw}|D=1] = \mathbb{E}[\vartheta_0^{aipw}|D=0] = 1$ (i.e., the weights have mean one). It is possible for ϑ_0^{aipw} to be negative for some values of X. In addition, the weights satisfy the following covariate balancing

[13] All of the weights in this section are functions of (X_{t^*}, X_{t^*-1}, Z), but we omit this dependence to conserve on notation.

[14] To be clear, the proof of Proposition 1 does not require any substantive restrictions on the model for the propensity score, but it does use linearity of the outcome regression model. That said, the outcome regression model could include interactions, higher order terms, etc.

properties

$$\mathbb{E}[\vartheta_0^{aipw} X_{t^*}|D=0] = \mathbb{E}[X_{t^*}|D=1]$$
$$\mathbb{E}[\vartheta_0^{aipw} X_{t^*-1}|D=0] = \mathbb{E}[X_{t^*-1}|D=1]$$
$$\mathbb{E}[\vartheta_0^{aipw} Z|D=0] = \mathbb{E}[Z|D=1]$$

Proposition 1 shows that the AIPW working model estimand in Equation (11) can be re-formulated as a weighting estimator. It is possible for the weights to be negative; in applications, it is straightforward to calculate the sample analog of the weights.[15] The main takeaway from Proposition 1 is that, unlike the implicit TWFE weights discussed above, the implicit AIPW weights balance the levels of time-varying covariates and time-invariant covariates across groups.

Remark 3 (Regression adjustment and IPW as special cases of AIPW)**.** Two special cases of the AIPW working model estimand discussed above are worth mentioning. If we set $\tilde{p}(X_{t^*}, X_{t^*-1}, Z) = \pi$ (this amounts to "not including any covariates in the propensity score working model"), then the second term in Equation (11) is equal to zero, and the expression for $\widetilde{ATT}$ reduces to a regression adjustment estimand. Similarly, if we do not include any covariates in the outcome regression model, $\widetilde{ATT}$ becomes the inverse propensity score weighting (IPW) estimand. This means that our results in this section also cover those two cases.[16]

5 Multiple Periods and Variation in Treatment Timing

The above discussion has focused on the setting with exactly two periods. In this section, we expand those arguments to the case where there are more time periods and where there can be variation in treatment timing across different units. This setting is common in empirical work in economics and has been studied in several recent papers (de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), among others). The proofs of all of the results in this section are provided in Appendix SB in the Supplementary Appendix.

We need to introduce some more notation and assumptions for the arguments in this section. First, let T denote the number of time periods. In this section, we allow for T to be larger than two, but we focus on "short" panels where T is considered to be fixed.

Assumption MP-1 (Staggered Treatment Adoption)**.** *For all units and time periods $t = 2, \dots, T$, $D_{it-1} = 1 \implies D_{it} = 1$.*

[15] To see that all of the components ϑ_0^{aipw} can be calculated easily, notice that the first and third term only depend on $\tilde{p}(X_{t^*}, X_{t^*-1}, Z)$, which comes from the proposed model for the propensity score. The second term is less obvious as it depends on $p(X_{t^*}, X_{t^*-1}, Z)$, the actual (unknown) propensity score. However, the proof of Proposition 1 shows that $\gamma_0 = \dfrac{\pi}{1-\pi}\mathbb{E}[XX'|D=0]^{-1}\mathbb{E}[X|D=1]$, see in particular Equation (S1) in the proof of Lemma 3 in Appendix SA in the Supplementary Appendix, which can be directly estimated.

[16] More generally, suppose that one includes the covariates $X^{or} := C^{or}(X_{t^*}, X_{t^*-1}, Z)$ in the working outcome regression model, and the covariates $X^{ps} := C^{ps}(X_{t^*}, X_{t^*-1}, Z)$ in the working model for the propensity score (where C^{or} and C^{ps} are functions that could include subsets of the covariates, higher order terms, interactions, etc.), then the results in Proposition 1 continue to apply with $\tilde{p}(X^{ps})$ replacing $\tilde{p}(X_{t^*}, X_{t^*-1}, Z)$, $p(X^{or})$ replacing $p(X_{t^*}, X_{t^*-1}, Z)$, and all linear projections being on X^{or} rather than on X.

Assumption MP-1 says that once a unit becomes treated in one period, it remains treated in subsequent periods. Under Assumption MP-1, a unit's entire sequence of treatments is fully characterized by its "group" where group refers to the time period when the unit became treated. Let G_i denote a unit's group and denote the full set of groups by $\mathcal{G} \subseteq \{2, \ldots, T+1\}$. This notation implicitly drops units that are already treated in the first period. We also use the convention of setting $G_i = T+1$ among units that do not participate in the treatment in any period from $2, \ldots, T$,[17] and we define $\bar{\mathcal{G}} := \mathcal{G} \setminus \{T+1\}$ as the set of groups that participate in the treatment in any period. It is also convenient to define a binary indicator of being in the never-treated group: let $U_i = 1$ for units that never participate in the treatment and $U_i = 0$ otherwise.

Let Y_{it} denote the observed outcome for unit i in time period t. Given Assumption MP-1, we define potential outcomes based on a unit's group; that is, let $Y_{it}(g)$ denote the potential outcome for unit i in time period t if it were in group g. In terms of potential outcomes, the observed outcome is $Y_{it} = Y_{it}(G_i)$. In other words, the observed outcome is the potential outcome according to unit i's actual group. To make the notation more transparent, we also define $Y_{it}(0)$ to be unit i's potential outcome in time period t if it never participated in the treatment. We make the following assumption.

Assumption MP-2 (No-Anticipation)**.** *For $t < G_i$ (i.e., pre-treatment periods for unit i), $Y_{it} = Y_{it}(0)$.*

Assumption MP-2 says that, in periods before a unit is treated, its observed outcomes are untreated potential outcomes. This rules out that the treatment affects outcomes in periods before the treatment actually occurs. Next, define X_{it} to be a $k \times 1$ vector of time-varying covariates, and let $\mathbf{X}_i := (X'_{i1}, X'_{i2}, \ldots, X'_{iT})'$ denote the $Tk \times 1$ vector that stacks the time-varying covariates across periods. Finally, we continue to use Z_i to denote an $l \times 1$ vector of time-invariant covariates.

Assumption MP-3 (Multi-Period Sampling)**.** *The observed data consists of $\{Y_{i1}, \ldots, Y_{iT}, X_{i1}, \ldots, X_{iT}, Z_i, G_i\}_{i=1}^n$ which are independent and identically distributed.*

Assumption MP-4 (Multi-Period Overlap)**.** *There exists some $\epsilon > 0$ such that, for all $g \in \mathcal{G}$, $\mathrm{P}(G = g) > \epsilon$ and $\mathrm{P}(U = 1|\mathbf{X}, Z) > \epsilon$.*

Assumption MP-5 (Multi-Period Parallel Trends)**.** *For $t = 2, \ldots, T$ and for all $g \in \mathcal{G}$,*

$$\mathbb{E}[\Delta Y_t(0)|\mathbf{X}, Z, G = g] = \mathbb{E}[\Delta Y_t(0)|\mathbf{X}, Z]$$

Assumptions MP-3 to MP-5 extend Assumptions 1 to 3 to a setting with more than two time periods and variation in treatment timing. The setup and assumptions considered here are standard in the DiD literature. We discuss common, empirically relevant extensions in Appendix SC.2 in the Supplementary Appendix.

[17] In the literature, it is somewhat more common to set $G_i = \infty$ for never-treated units. Either convention is essentially arbitrary, but setting $G_i = T + 1$ unifies some of the notation for the TWFE decomposition results presented below.

5.1 Identification

For identification, following Callaway and Sant'Anna (2021), Wooldridge (2021) and several other papers, we target identifying group-time average treatment effects, which are defined as

$$ATT(g,t) := \mathbb{E}[Y_t(g) - Y_t(0)|G = g]$$

$ATT(g,t)$ is the average treatment effect for group g in period t. We also define the conditional-on-covariates version of group-time average treatment effects

$$ATT_{g,t}(\mathbf{x}, z) := \mathbb{E}[Y_t(g) - Y_t(0)|\mathbf{X} = \mathbf{x}, Z = z, G = g]$$

Next, we provide an identification result.

Proposition 2. *Under Assumptions MP-1 to MP-5, for $t \geq g$,*

$$ATT_{g,t}(\mathbf{X}, Z) = \mathbb{E}[Y_t - Y_{g-1}|\mathbf{X}, Z, G = g] - \mathbb{E}[Y_t - Y_{g-1}|\mathbf{X}, Z, U = 1]$$

and

$$\begin{aligned} ATT(g,t) &= \mathbb{E}[ATT_{g,t}(\mathbf{X}, Z)|G = g] \\ &= \mathbb{E}[Y_t - Y_{g-1}|G = g] - \mathbb{E}\Big[\mathbb{E}[Y_t - Y_{g-1}|\mathbf{X}, Z, U = 1]\Big|G = g\Big] \end{aligned}$$

The proof of Proposition 2 is provided in Appendix SB.1 in the Supplementary Appendix—it closely mimics the identification result for $ATT(g,t)$ in Callaway and Sant'Anna (2021) except for that some of the covariates can be time-varying. It generalizes the result in Equation (2) from a setting with two time periods to one with staggered treatment adoption. Proposition 2 says that conditional group-time average treatment effects are identified in the setting considered in this section and are equal to the mean path of outcomes between $(g-1)$ (which is the most recent pre-treatment period for group g) and period t conditional on both time-varying and time-invariant covariates for group g relative to the same path of outcomes for never-treated units.[18] The second part of the result says that $ATT(g,t)$ can be recovered by averaging $ATT_{g,t}(\mathbf{X}, Z)$ over the distribution of time-varying and time-invariant covariates for group g.

Group-time average treatment effects are important building blocks for our results below on interpreting TWFE regressions. However, unlike α from the TWFE regression in Equation (1), they are functional parameters in the sense that they can vary arbitrarily across g and t. Therefore, it is more natural to compare α from the TWFE regression to an aggregated causal effect parameter; in particular, we consider the following overall average treatment effect on the treated parameter

$$ATT^o := \mathbb{E}\Big[\bar{Y}^{post} - \bar{Y}(0)^{post}\Big|U = 0\Big]$$

[18] The same set of assumptions also rationalize using different comparison groups such as the not-yet-treated group (i.e., where the comparison group is defined by $D_t = 0$ rather than $U = 1$). See Callaway (2023) for a related discussion.

where, for units that ever participate in the treatment, we define

$$\bar{Y}_i^{post} := \frac{1}{T - G_i + 1} \sum_{t=G_i}^{T} Y_{it} \qquad \text{and} \qquad \bar{Y}_i(0)^{post} := \frac{1}{T - G_i + 1} \sum_{t=G_i}^{T} Y_{it}(0)$$

which are the average observed outcome and average untreated potential outcome, respectively, across unit i's post-treatment time periods. Thus, ATT^o is the average treatment effect across the population that participates in the treatment in any time period. It is straightforward to show that

$$ATT^o = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} w^o(g,t) ATT(g,t) \tag{12}$$

where $w^o(g,t) := \bar{p}_g/(T - g + 1)$ and $\bar{p}_g := \mathrm{P}(G = g | G \in \bar{\mathcal{G}})$, which is the probability of being in group g conditional on being among the set of groups that ever participates in the treatment.[19]

5.2 TWFE Decomposition

Next, we provide a decomposition of α from Equation (1) but in the case considered in this section with more than two time periods and staggered treatment adoption. The discussion below often uses double-demeaned random variables (i.e., transformed random variables that have had unit and time fixed effects removed); for example, we define $\ddot{Y}_{it} := Y_{it} - \bar{Y}_i - \mathbb{E}[Y_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[Y_s]$. We focus on estimating α from Equation (1) by fixed effects estimation. Thus, after applying the double-demeaning transformation, we ultimately use the following estimating equation for α:

$$\ddot{Y}_{it} = \alpha \ddot{D}_{it} + \ddot{X}_{it}'\beta + \ddot{e}_{it} \tag{13}$$

Before providing our main results, we need to introduce some more notation. First, notice that a unit's group fully determines $\ddot{D}_{it}$; i.e., $\ddot{D}_{it} = h(G_i, t)$ where

$$h(g,t) := \mathbb{1}\{t \geq g\} - \frac{T - g + 1}{T} - \mathbb{E}[D_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s]$$

Next, building on the notation in the main text, define the population linear projection of $\ddot{D}_{it}$ on $\ddot{X}_{it}$ as

$$\mathrm{L}(\ddot{D}_t | \ddot{X}_t) = \ddot{X}_t' \mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ddot{X}_s \ddot{X}_s'\right]^{-1} \mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ddot{X}_s \ddot{D}_s\right] =: \ddot{X}_t'\Gamma$$

Next, define the population linear projection of $(Y_{it} - Y_{ig-1})$ on $(X_{it} - X_{ig-1})$ using the never-treated group

$$\mathrm{L}_0\Big(Y_t - Y_{g-1} \big| X_t - X_{g-1}\Big) =: \lambda_{0,t,g-1} + \big(X_t - X_{g-1}\big)'\Lambda_{0,t,g-1}$$

where $\lambda_{0,t,g-1}$ is the intercept and $\Lambda_{0,t,g-1}$ is the slope coefficient, both of which can vary by the period t and the base period $(g-1)$. Furthermore, define Λ_0 as the vector of coefficients from a TWFE

[19] There are a number of other interesting aggregated parameters that show up in the literature. Besides ATT^o, the leading example is the event study, though several others are discussed in Callaway and Sant'Anna (2021). We discuss event studies in particular in more detail in Remark S9 in the Supplementary Appendix, and our provided code can immediately be used to produce event studies.

regression of Y_{it} on X_{it} using only the never-treated group (see Equation (S5) in Appendix SB.1 in the Supplementary Appendix for the complete expression), and define $\lambda_t := \mathbb{E}[Y_t - X_t'\Lambda_0 | U = 1]$.

An important first step for many of our results below is to re-write α as

$$\alpha = \frac{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[(\ddot{D}_t - \ddot{X}_t'\Gamma)\ddot{Y}_t\Big]}{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[(\ddot{D}_t - \ddot{X}_t'\Gamma)^2\Big]} \tag{14}$$

which holds from using Frisch-Waugh-Lovell arguments. We start by providing a decomposition of α.

Proposition 3. *Under Assumptions MP-1, MP-3 and MP-4,*

$$\alpha = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,G{=}g] - \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U{=}1]\Big)\Big|G=g\Big] \tag{A}$$

$$+\sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big\{\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U{=}1] - \Big((\lambda_t-\lambda_{g-1}) + (X_t\text{-}X_{g-1})'\Lambda_0\Big)\Big\}\Big|G=g\Big] \tag{B}$$

$$+\sum_{g\in\bar{\mathcal{G}}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,G{=}g] - \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U{=}1]\Big)\Big|G=g\Big] \tag{C}$$

$$+\sum_{g\in\bar{\mathcal{G}}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big\{\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X},Z,U{=}1] - \Big((\lambda_t-\lambda_{g-1}) + (X_t\text{-}X_{g-1})'\Lambda_0\Big)\Big\}\Big|G=g\Big] \tag{D}$$

where

$$w_{g,t}^{twfe}(\ddot{X}_t) := \frac{\Big(h(g,t) - \ddot{X}_t'\Gamma\Big)\pi_g}{\sum_{l\in\bar{\mathcal{G}}}\sum_{s=l}^{T}\mathbb{E}\Big[\Big(h(l,s) - \ddot{X}_{is}'\Gamma\Big)\Big|G=l\Big]\pi_l}$$

which have the following properties

(i) $\sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big|G=g\Big] = 1$ *and* $\sum_{g\in\mathcal{G}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big|G=g\Big] = -1.$

(ii) It is possible for $w_{g,t}^{twfe}(\ddot{X}_t)$ *to be negative for some values of* $\ddot{X}_t$ *with* $g\in\bar{\mathcal{G}}$ *and* $t\in\{1,\ldots,T\}$.

Proposition 3 is a main result in this part of the paper. It provides a decomposition of α from Equation (1) in a setting with staggered treatment adoption. It is a decomposition in the sense that it only relies on regularity assumptions and does not invoke identification assumptions such as parallel trends or no-anticipation. Terms (A)-(D) differ along two dimensions. First, Terms (A) and (B) involve post-treatment periods only, while Terms (C) and (D) involve only pre-treatment periods. Second, Terms (A) and (C) involve differences between conditional expectations of paths of outcomes between group g and the never-treated group. Once we invoke parallel trends and no-anticipation, these expressions in Term (A) will become group-time average treatment effects, while the expressions in Term (C) will be equal to zero (more details below). Term (C) involves violations of parallel trends in pre-treatment periods, which, as can be seen from the proposition, will contribute to our eventual estimate of α. Terms (B) and

(D) are post-treatment and pre-treatment misspecification bias terms, respectively; these terms include hidden linearity bias terms similar to the ones we emphasized in the two-period case. We consider these in substantially more detail below. That the weights on Term (B) sum to one (as opposed to, say, zero) indicates that the importance/magnitude of misspecification bias is on par with the magnitude of the treatment effects themselves. That the weights on Terms (C) and (D) are negative arises because these are pre-treatment periods and, hence, "comparison periods". That these weights sum to negative one indicates that pre-treatment violations of parallel trends and pre-treatment misspecification bias are as important for the resulting estimate of α as the treatment effects themselves.

Next, we add the assumptions of parallel trends and no-anticipation. To conserve on notation, define

$$\xi_{t,g-1}(\mathbf{X}, Z) := \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X}, Z, U{=}1] - \Big((\lambda_t - \lambda_{g-1}) + (X_t\text{-}X_{g-1})'\Lambda_0\Big)$$

which corresponds to the underlying components of the misspecification bias terms in Terms (B) and (D) in Proposition 3.

Theorem 3. *Under Assumptions MP-1 to MP-5,*

$$\alpha = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\Big(ATT_{g,t}(\mathbf{X}, Z) + \xi_{t,g-1}(\mathbf{X}, Z)\Big)\Big|G = g\Big] + \sum_{g\in\bar{\mathcal{G}}}\sum_{t=1}^{g-1}\mathbb{E}\Big[w_{g,t}^{twfe}(\ddot{X}_t)\xi_{t,g-1}(\mathbf{X}, Z)\Big|G = g\Big]$$

The weights are the same as in Proposition 3 and satisfy the same properties.

Theorem 3 shows, when we additionally invoke the parallel trends assumption and no-anticipation assumption, that α from the TWFE regression in Equation (1) is equal to a weighted average of conditional-on-covariates group-time average treatment effects plus two misspecification bias terms that are unaffected by parallel trends and no-anticipation. This result is analogous to (and extends) the result in Theorem 1 in the case with exactly two periods. Like the earlier case, the weights on conditional group-time average treatment effects are (i) driven by the estimation method and (ii) can be negative. Next, we provide a result that decomposes the misspecification bias terms in Theorem 3.

Proposition 4. *Under Assumptions MP-1 to MP-5, the misspecification bias terms in Proposition 3 and Theorem 3 can be decomposed as*

$$\begin{aligned}
\xi_{t,g-1}(\mathbf{X}, Z) = {} & \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X}, Z, U{=}1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X}, U{=}1]\Big) && \text{(MB-1)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|\mathbf{X}, U{=}1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|X_t, X_{g-1}, U{=}1]\Big) && \text{(MB-2)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|X_t, X_{g-1}, U{=}1] - \mathbb{E}[Y_t\text{-}Y_{g-1}|(X_t\text{-}X_{g-1}), U{=}1]\Big) && \text{(MB-3)}\\
& + \Big(\mathbb{E}[Y_t\text{-}Y_{g-1}|(X_t\text{-}X_{g-1}), U{=}1] - \big(\lambda_{0,t,g-1} + (X_t\text{-}X_{g-1})'\Lambda_{0,t,g-1}\big)\Big) && \text{(MB-4)}\\
& + \Big(\big(\lambda_{0,t,g-1} - (\lambda_t - \lambda_{g-1})\big) + (X_t\text{-}X_{g-1})'(\Lambda_{0,t,g-1} - \Lambda_0)\Big) && \text{(MB-5)}
\end{aligned}$$

Next, we provide a discussion of the components of the misspecification bias terms in Proposition 4 along with a set of sufficient conditions to eliminate them from the expression for α in Theorem 3. These conditions rationalize interpreting α from Equation (1) as a weighted average of $ATT_{g,t}(\mathbf{X}, Z)$. We discuss these conditions in words below and state them formally in Assumption MP-6 in Appendix SB.1 in the Supplementary Appendix.

Conditions to Eliminate Misspecification Bias

(1) The path of untreated potential outcomes does not depend on time-invariant covariates.

(2) The path of untreated potential outcomes does not depend on time-varying covariates in other periods besides (g–1) and t.

(3) The path of untreated potential outcomes only depends on the change in time-varying covariates between periods (g–1) and t.

(4) The path of untreated potential outcomes is linear in the change in time-varying covariates.

(5) The effect of the change in time-varying covariates over time on the path of untreated potential outcomes is constant across time periods.

Each condition serves to set the corresponding term in Proposition 4 equal to 0. Conditions (1)-(3) are all required to deal with the multiple-period version of hidden linearity bias: that transforming the model to eliminate the unit fixed effect also changes the functional form of the time-varying covariates and eliminates the time-invariant covariates, and, hence, effectively results in changing the parallel trends assumption. Condition (4) is a linearity condition, similar to Condition (C) in Assumption 4 in the two-period case. As discussed earlier, this is an expected/natural condition, given that we are considering the properties of a linear model. Condition (5) does not have an immediate analog from the two-period case. It says that, while the path of untreated potential outcomes can depend on the magnitude of changes of time-varying covariates over time, the effect of a specific change in the covariates should not vary across time periods. A good alternative way to view these conditions is as restrictions on the linear model for untreated potential outcomes in Equation (3) considered in Section 2.2:

$$Y_{it}(0) = \theta_t + \eta_i + Z_i'\delta_t + X_{it}'\beta_t + e_{it}. \tag{15}$$

Condition (1) is satisfied if $\delta_t = \delta$. Conditions (2)-(5) are satisfied if, additionally, $\beta_t = \beta$. Next, we provide a result interpreting α from the TWFE regression under the additional assumptions that rule out misspecification bias and additionally provide extra conditions for α to be equal to the ATT.

Theorem 4. *Under Assumptions MP-1 to MP-5 and MP-6,*

$$\alpha = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T} \mathbb{E}\left[w_{g,t}^{twfe}(\ddot{X}_t) ATT_{g,t}(\mathbf{X}, Z)\Big|G=g\right]$$

where the weights $w_{g,t}^{twfe}(\ddot{X}_t)$ are the same ones as in Theorem 3. If, in addition, $ATT_{g,t}(\mathbf{X}, Z) = ATT(g,t)$, then

$$\alpha = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T} \left\{\mathbb{E}\left[w_{g,t}^{twfe}(\ddot{X}_t)\Big|G=g\right] ATT(g,t)\right\}$$

If, in addition, $ATT_{g,t}(\mathbf{X}, Z) = ATT$, then

$$\alpha = ATT$$

This result is the multiple period version of Theorem 2 from the earlier case with only two time periods. The first part shows that, when one additionally includes the five conditions discussed above,

α from the TWFE regression can be interpreted as a weighted average of conditional average treatment effects. Even if these conditions hold, the weights on conditional $ATT(g, t)$s are (i) driven by the estimation method, (ii) difficult to rationalize except under extra assumptions restricting treatment effect heterogeneity, and (iii) as pointed out by de Chaisemartin and D'Haultfœuille (2020), the weights can be negative. The second and third parts layer on additional restrictions on treatment effect heterogeneity. For the second part, even if $ATT_{g,t}(\mathbf{X}, Z)$ does not vary across covariates (a very strong additional condition), one will still recover weighted averages of $ATT(g, t)$, the weights will still be difficult to interpret, and the weights can still be negative. Finally, if we fully rule out any forms of systematic treatment effect heterogeneity, then α will be equal to the ATT. The main takeaway from this result is that, even if one is willing to make strong auxiliary assumptions about the path of untreated potential outcomes as in Assumption MP-6, the weights on the conditional average treatment effects will still be difficult to interpret and can be negative unless one is willing to impose additional assumptions that sharply limit treatment effect heterogeneity.

Relationship to Other Papers The results in Theorems 3 and 4 are related to several other results in the literature. Although they mainly consider interpreting TWFE regressions with multiple periods and variation in treatment timing in a setting without covariates in the parallel trends assumption, both de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (2021) include some results for TWFE regressions that include time-varying covariates. Some of our results, particularly the first part of Theorem 4, are closely related to Theorem S4 in Online Appendix 3.3 of de Chaisemartin and D'Haultfœuille (2020). In that theorem, de Chaisemartin and D'Haultfœuille (2020) essentially take as a starting point the combination of our Assumptions MP-5 and MP-6 and show that, under a conditional parallel trends assumption that involves only changes in observed covariates and linearity assumptions that their main results related to multiple periods and variation in treatment timing essentially continue to apply. Our weights in Theorem 4 are the same as in that paper, though we expand it in important ways by allowing for time-invariant covariates in the parallel trends assumption and by providing conditions under which the expression for α can be simplified. Our results in Theorem 3 that provide possible sources of misspecification bias are also new to the literature.

Goodman-Bacon (2021, Section 5.2) provides a decomposition of α into a "within" component and "between" component. The between component arises due to variation in treatment timing and can be expressed as an adjusted-by-covariates 2x2 difference-in-differences comparison.[20] The within component comes from variation in the covariates within a particular group and is, therefore, related to our expression for α in the setting with only two time periods. Relative to Goodman-Bacon (2021), we further decompose this type of term into several more primitive objects that highlight that researchers should be careful in interpreting "within" components as averages of causal effects unless they are willing to invoke extra assumptions. In work first made publicly available after the first version of our paper, Lin and Zhang (2022) build on some of our results and the event study decomposition in Sun and Abraham (2021) and show that an additional bias term can arise in event study regressions

[20]The main results in Goodman-Bacon (2021) (for the case without covariates) show that α from the TWFE regression can be written as a weighted average of all possible 2x2 difference-in-differences type comparisons among groups whose treatment status changes between two periods relative to groups whose treatment status does not change across periods. The term discussed here is similar in spirit to these terms in the unconditional setting.

that include time-varying covariates. Finally, Ishimaru (2022, Section 2.2), like de Chaisemartin and D'Haultfœuille (2020), provides conditions under which TWFE regressions that include covariates can be interpreted as weighted averages of underlying treatment effect parameters. These include a version of conditional parallel trends that holds when one conditions on the change in covariates over time[21] and an assumption on the linearity of the propensity score conditional on changes in observed covariates over time.[22]

5.3 Covariate Balance Diagnostics with Multiple Periods

Next, we discuss how to extend our TWFE and AIPW diagnostics in Section 4 to settings with multiple periods and variation in treatment timing. Similar to the case with two periods, the goal in this section is to show that (i) α from the TWFE regression and $\widetilde{ATT}^{aipw,o}$ from our AIPW estimator can be recast as weighting estimators and then (ii) to apply the implicit weights to levels of the time-varying covariates and the time-invariant covariates in order to understand how well each of these balances covariates for treated groups relative to the never-treated group. As above, this provides a way to assess the sensitivity of the TWFE regression to hidden linearity bias.

Toward this end (and like for the case with two periods), notice that if we could find balancing weights that, for a particular group g in time period t, balance the distribution of the time-varying and time-invariant covariates for the never-treated group relative to group g, then we could recover $ATT(g,t)$ by applying these weights to the path of outcomes for the never-treated group. In particular, for some group $g \in \bar{\mathcal{G}}$ and post-treatment time period $t \geq g$, let $\vartheta_{g,t}(\mathbf{X}, Z)$ denote balancing weights that re-weight the never-treated group so that, after applying the weights, it has the same distribution of $(\mathbf{X}, Z)$ as group g. Given these balancing weights and using very similar arguments as in Section 4, one can show that

$$ATT(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G = g] - \mathbb{E}[\vartheta_{g,t}(\mathbf{X}, Z)(Y_t - Y_{g-1})|U = 1] \tag{16}$$

and that

$$ATT^o = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} \left\{ \mathbb{E}\left[Y_t - Y_{g-1} \middle| G = g \right] - \mathbb{E}\left[\vartheta_{g,t}(\mathbf{X}, Z)(Y_t - Y_{g-1}) \middle| U = 1 \right] \right\} w^o(g,t) \tag{17}$$

Equations (16) and (17) show that group-time average treatment effects and ATT^o can, under conditional parallel trends, be recovered by re-weighting the never-treated group to have the same distribution of $\mathbf{X}$ and Z as each group.

[21] Ishimaru (2022) does point out that "conditioning on [changes in time-varying covariates] may not be sufficient to make parallel trends plausible."

[22] One way that the decomposition in Ishimaru (2022) is more general than the one in the current paper is that it does not require the treatment to be binary. Ishimaru (2022) also considers an interesting extension on decomposing a modified TWFE regression that additionally includes time-varying coefficients on time-varying coefficients. Based on his result, it seems likely that this sort of regression would not suffer from issues related to parallel trends depending on the levels of time-varying covariates rather than only changes in time-varying covariates over time. However, it appears that this regression would still suffer from the other issues mentioned in this section; that said, this is a distinct (and much less commonly used in empirical work) regression from the TWFE regression in Equation (1).

Covariate Balance Diagnostics for TWFE Regressions

Next, we turn to reformulating the TWFE regression in Equation (1) as a weighting estimator. In Proposition S7 in the Supplementary Appendix, we show that α can be rewritten in terms of implicit regression weights. We show that

$$\alpha = \sum_{g\in\bar{\mathcal{G}}}\sum_{t=g}^{T} \bar{w}^{twfe}(g,t)\left\{\mathbb{E}\left[w_{g,t}^{1,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\Big|G=g\right]-\mathbb{E}\left[w_{g,t}^{0,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\Big|U=1\right]\right\} \tag{18}$$
$$+\sum_{g\in\bar{\mathcal{G}}}\sum_{t=1}^{g-1} \bar{w}^{twfe}(g,t)\left\{\mathbb{E}\left[w_{g,t}^{1,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\Big|G=g\right]-\mathbb{E}\left[w_{g,t}^{0,twfe}(\mathbf{X},Z)(Y_t-Y_{g-1})\Big|U=1\right]\right\}+r$$

where $\bar{w}^{twfe}(g,t) := \mathbb{E}[w_{g,t}^{twfe}(\ddot{X}_t)|G=g]$ and

$$w_{g,t}^{1,twfe}(\mathbf{X},Z) := \frac{(\ddot{D}_t-\ddot{X}_t'\Gamma)}{\mathbb{E}[(\ddot{D}_t-\ddot{X}_t'\Gamma)|G=g]} \quad \text{and} \quad w_{g,t}^{0,twfe}(\mathbf{X},Z) := \frac{(\ddot{D}_t-\ddot{X}_t'\Gamma)}{\mathbb{E}[(\ddot{D}_t-\ddot{X}_t'\Gamma)|U=1]}$$

and where r is a remainder term.[23]

AIPW Estimands with Multiple Periods

Next, we consider AIPW working model estimands for group-time average treatment effects and the overall average treatment effect with multiple periods and variation in treatment timing. This is the population version of our main alternative estimator to the TWFE regression; see the next section for further details. Define the AIPW working model estimand for $ATT(g,t)$ as

$$\widetilde{ATT}^{aipw}(g,t) = \mathbb{E}\Big[(Y_t\text{–}Y_{g-1})-\mathrm{L}_{g,t}^0(Y_t\text{–}Y_{g-1}|\mathbf{X},Z)\Big|G=g\Big]-\mathbb{E}\Big[\tilde{w}_{g,t}^{0,aipw}(\mathbf{X},Z)\big((Y_t\text{–}Y_{g-1})-\mathrm{L}_{g,t}^0(Y_t\text{–}Y_{g-1}|\mathbf{X},Z)\big)\Big|U=1\Big] \tag{19}$$

where $\mathrm{L}_{g,t}^0(Y_t\text{–}Y_{g-1}|\mathbf{X},Z)$ is the projection of $Y_t\text{–}Y_{g-1}$ onto $\mathbf{X}$ and Z in the untreated group,[24] and where

$$\tilde{w}_{g,t}^{0,aipw} := \frac{\tilde{\varpi}_{g,t}^{0,aipw}(\mathbf{X},Z)}{\mathbb{E}[\tilde{\varpi}_{g,t}^{0,aipw}(\mathbf{X},Z)|U=1]} \qquad \text{and} \qquad \tilde{\varpi}_{g,t}^{0,aipw}(\mathbf{X},Z) = \frac{\pi_0\tilde{p}_{g,t}(\mathbf{X},Z)}{\pi_g\big(1-\tilde{p}_{g,t}(\mathbf{X},Z)\big)}$$

where $\pi_g := \mathrm{P}(G=g)$, $\pi_0 := \mathrm{P}(U=1)$, and $\tilde{p}_{g,t}(\mathbf{X},Z)$ denotes the probability limit of a working model for the generalized propensity score

$$p_{g,t}(\mathbf{X},Z) := \mathrm{P}(G=g|\mathbf{X},Z,\mathbb{1}\{G=g\}+U=1)$$

[23] The remainder term is a byproduct of using $(g-1)$ as a base period in the decomposition of α presented here. In the discussion after Proposition S7 in the Supplementary Appendix, we argue that this term is likely to be small in most applications, and, indeed, in all of the diagnostics that we report in our application with multiple periods, this term is negligible. We also provide a decomposition that uses period one as the base period that involves exactly the same weights but does not include a remainder term in Proposition S6. Our reasons for preferring the decomposition using $(g-1)$ as the base period are that (i) it allows for a direct comparison with the AIPW working model estimand discussed below, where their differences are fully accounted for by differences in implicit weighting schemes and (ii) it allows us to quantify how much pre-treatment violations of parallel trends contribute to α.

[24] We index $\mathrm{L}_{g,t}^0(Y_t\text{–}Y_{g-1}|\mathbf{X},Z)$ by (g,t) to allow for this linear projection to include only a subset of the time-varying covariates (see the next section for details), and this subset could change across groups and/or time periods. As for the case with two periods, the main requirement is that the outcome regression working model be linear, though it can include interactions and higher order terms.

which is the conditional probability of being in group g conditional on being in group g or the never-treated group. We index $p_{g,t}(\mathbf{X}, Z)$ and $\tilde{p}_{g,t}(\mathbf{X}, Z)$ by g and t to allow the generalized propensity score and our working model for it to change across time periods, particularly with respect to which time-varying covariates are included in the model. It is straightforward to see, along the lines of the case with two time periods, that the sample analog of $\widetilde{ATT}^{aipw}(g,t)$ is doubly robust for $ATT(g,t)$ (see the next section for more details). In addition, define

$$\widetilde{ATT}^{aipw,o} = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} \widetilde{ATT}^{aipw}(g,t) w^o(g,t) \tag{20}$$

which is AIPW working model estimand for ATT^o.

Covariate Balance Diagnostics for AIPW

Next, we show that $\widetilde{ATT}^{aipw,o}$ can be rewritten in terms of weighted averages of paths of outcomes over time for each group relative to the never-treated group. In particular, building on the arguments from the two-period case discussed above, in Proposition S8 in the Supplementary Appendix, we show that

$$\widetilde{ATT}^{aipw,o} = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} w^o(g,t) \left\{ \mathbb{E}\left[\vartheta_{g,t}^{1,aipw}(\mathbf{X}, Z)(Y_t - Y_{g-1}) \Big| G = g \right] - \mathbb{E}\left[\vartheta_{g,t}^{0,aipw}(\mathbf{X}, Z)(Y_t - Y_{g-1}) \Big| U = 1 \right] \right\} \tag{21}$$

where $w^o(g,t)$ are the same weights as in ATT^o above, and where

$$\vartheta_{g,t}^{1,aipw}(\mathbf{X}, Z) := 1 \quad \text{and} \quad \vartheta_{g,t}^{0,aipw}(\mathbf{X}, Z) := \tilde{w}_{g,t}^{aipw}(\mathbf{X}, Z) + \frac{\gamma_{g,t}' \mathbf{W}_{g,t}}{\mathbb{E}[\gamma_{g,t}' \mathbf{W}_{g,t} | U = 1]} - \frac{\tilde{\gamma}_{g,t}' \mathbf{W}_{g,t}}{\mathbb{E}[\tilde{\gamma}_{g,t}' \mathbf{W}_{g,t} | U = 1]}$$

where $\mathbf{W}_{g,t}$ are the covariates used in $\mathrm{L}_{g,t}^0(Y_t - Y_{g-1} | \mathbf{X}, Z)$, $\gamma_{g,t}$ is the linear projection coefficient from projecting $p_{g,t}(\mathbf{X}, Z)/(1 - p_{g,t}(\mathbf{X}, Z))$ on $\mathbf{W}_{g,t}$, and $\tilde{\gamma}_{g,t}$ is the linear projection coefficient from projecting $\tilde{p}_{g,t}(\mathbf{X}, Z)/(1 - \tilde{p}_{g,t}(\mathbf{X}, Z))$ on $\mathbf{W}_{g,t}$.

Comparison of TWFE and AIPW Covariate Balance Properties

It is worthwhile to compare the covariate balancing properties from the TWFE regression to AIPW. On the inside, both are weighted averages of the path of outcomes for group g relative to the untreated group, where the weights depend on the covariates. All of these inner weights, both for TWFE and AIPW, have mean one by construction. However, in practice, these covariate-specific weights can be much different from each other. For one thing, the TWFE weights are regression-type weights that only depend on transformed values of the time-varying covariates and not directly on the levels of time-varying covariates or on time-invariant covariates at all. As formulated above, the implicit AIPW weights will balance whatever covariates are included in $\mathbf{W}_{g,t}$. As we discuss in Section 6, it may be necessary in many applications to do some dimension reduction (so that $\mathbf{W}_{g,t}$ is of lower dimension than $(\mathbf{X}, Z)$), especially with respect to the time-varying covariates. Thus, for example, if one ultimately estimates $ATT(g,t)$ by AIPW including $(X_t - X_{g-1})$, X_{g-1}, and Z as covariates, then AIPW balances the covariates that are included in the model, but would not necessarily balance other transformations of the time-varying covariates (e.g., this specification would not necessarily balance $\bar{X}$). For another difference, the inner weights balance towards different implied target populations. The AIPW weights balance towards group g, the correct target population for $ATT(g,t)$. On the other hand, the TWFE effectively re-weights both group g and the never-treated group. In general, the inner implicit TWFE weights or the inner implicit AIPW weights can be negative.

The outer weighting scheme is different, as the TWFE regression weights come from the group-specific means of $w_{g,t}^{twfe}(\ddot{X}_t)$ and the AIPW weights are the same as for ATT^o. For TWFE, the outer post-treatment weights sum to one, while the outer pre-treatment weights sum to negative one, which holds by the same argument as in Proposition 3 above. Similarly, the TWFE regression can be affected by violations of parallel trends in pre-treatment periods, while AIPW is not.

Like the case with two periods, one of our main interests in the application is to apply the implicit TWFE and AIPW weights to the covariates themselves to assess their sensitivity to hidden linearity bias. For example, in the application, we assess covariate balance by applying both sets of weights to $\bar{X}$ and Z to check how well TWFE and AIPW balance the covariates.

6 Alternative Estimation Strategies

This section discusses alternative estimation strategies that do not suffer from the limitations of TWFE regressions discussed above. We mainly focus on the setting considered in Section 5 with multiple periods and variation in treatment timing across units, noting that this case generalizes the two-period case. First, we discuss AIPW estimators in this context. AIPW estimators are attractive in the context that we consider, and many existing results are straightforward to adapt to our setting. Second, from an empirical perspective, the main complication is that, with panel data and time-varying covariates, the dimension of the covariates can be very large. We discuss several different dimension reduction techniques in the second part of this section.

To start with, define $m_{g,t}(\mathbf{X}, Z) := \mathbb{E}[Y_t(0) - Y_{g-1}(0)|\mathbf{X}, Z, U = 1]$. We refer to $m_{g,t}(\mathbf{X}, Z)$ as an outcome regression model. We continue to use $p_{g,t}(\mathbf{X}, Z)$ to denote the generalized generalized propensity score. Let $\hat{m}_{g,t}(\mathbf{X}, Z)$ and $\hat{p}_{g,t}(\mathbf{X}, Z)$ denote estimators of $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$, respectively. Then, we consider AIPW estimators of $ATT(g,t)$ of the form

$$\widehat{ATT}^{aipw}(g,t) := \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{w}_{g,t}^{1,aipw}(\mathbf{X}_i, Z_i) - \hat{w}_{g,t}^{0,aipw}(\mathbf{X}_i, Z_i)\Big)\Big((Y_t - Y_{g-1}) - \hat{m}_{g,t}(\mathbf{X}_i, Z_i)\Big)$$

which, after slightly re-arranging terms, is the sample analog of Equation (19) (where here we also allow for the possibility of a nonlinear model for the outcome regression), and where

$$\hat{w}_{g,t}^{1,aipw}(\mathbf{X}_i, Z_i) := \frac{\mathbb{1}\{G_i = g\}}{\hat{\pi}_g} \quad \text{and} \quad \hat{w}_{g,t}^{0,aipw}(\mathbf{X}_i, Z_i) := \frac{\mathbb{1}\{U_i = 1\}\frac{\hat{p}_{g,t}(\mathbf{X}_i, Z_i)}{1-\hat{p}_{g,t}(\mathbf{X}_i, Z_i)}}{\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\{U_j = 1\}\frac{\hat{p}_{g,t}(\mathbf{X}_j, Z_j)}{1-\hat{p}_{g,t}(\mathbf{X}_j, Z_j)}}$$

AIPW estimators have been well-studied and have several known properties, which we discuss next.

Remarks on AIPW Estimation

1. If we specify parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$—leading choices are a linear model for the outcome regression and logit or probit for the generalized propensity score, then $\widehat{ATT}^{aipw}(g,t)$ is doubly robust for $ATT(g,t)$. Double robustness means that if either the outcome regression or the propensity score model is correctly specified, then the estimator is consistent for $ATT(g,t)$. See Robins, Rotnitzky, and Zhao (1994), Scharfstein, Rotnitzky, and Robins (1999), and Słoczyński and Wooldridge (2018) for general results on the double robustness property of AIPW estimators and Sant'Anna and Zhao (2020) for the specific case of DiD.

2. Given parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$, asymptotic normality of $\widehat{ATT}^{aipw}(g,t)$ holds under Assumptions MP-1 to MP-5 and weak/standard regularity conditions based on standard results for AIPW estimators. In particular, if both the outcome regression and propensity score models are correctly specified, then $\sqrt{n}\big(\widehat{ATT}^{aipw}(g,t) - ATT(g,t)\big)$ has mean 0, is asymptotically normal, and achieves the semiparametric efficiency bound. If only one of the models is correctly specified, then the estimator is still asymptotically normal, although the estimator's variance will be larger than in the case where both models are correctly specified. See Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021) for more details.

3. The estimator $\widehat{ATT}^{aipw}(g,t)$ can be used to construct an estimator of the overall average treatment effect, ATT^o, by averaging over all groups and time periods. In particular,

$$\widehat{ATT}^{o} = \sum_{g \in \bar{\mathcal{G}}} \sum_{t=g}^{T} \hat{w}^o(g,t) \widehat{ATT}^{aipw}(g,t)$$

where $\hat{w}^o(g,t) = \frac{\hat{\mathrm{P}}(G=g|U=0)}{T-g+1}$. This estimator is consistent for ATT^o and is asymptotically normal under the same conditions discussed above. Similar results hold for event studies or other aggregated parameters that can be expressed as weighted averages of $ATT(g,t)$. These results follow directly from ones provided in Callaway and Sant'Anna (2021); see that paper for more details.

4. Regression adjustment is a special case of AIPW estimation when (i) we use a linear model for $m_{g,t}(\mathbf{X}, Z)$ and (ii) we use the estimator of the generalized propensity score $\hat{p}_{g,t}(\mathbf{X}, Z) = \hat{\mathrm{P}}(G = g | \mathbb{1}\{G = g\} + U = 1)$ (i.e., no covariates are included in the generalized propensity score model). In this case $\widehat{ATT}^{aipw}(g,t)$ simplifies to

$$\widehat{ATT}^{ra}(g,t) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{G_i = g\}}{\hat{\pi}_g} \Big(Y_{it} - Y_{ig-1} - \hat{m}_{g,t}(\mathbf{X}_i, Z_i) \Big)$$

where we use superscript on $\widehat{ATT}^{ra}(g,t)$ to indicate that this is a regression adjustment estimator of $ATT(g,t)$. In this case, consistent and asymptotically normal estimation of $ATT(g,t)$ hinges on correct specification of the outcome regression $m_{g,t}(\mathbf{X}, Z)$. One reason that regression adjustment is an important special case is that it is fairly common in empirical work to have groups with a small number of observations. In this case, estimating the generalized propensity score may be highly unstable and unreliable; hence, regression adjustment may be a more attractive alternative in these types of applications.

5. Several extensions to the AIPW estimator discussed above apply immediately to our case. These include allowing for anticipation effects and using alternative comparison groups (such as the not-yet-treated group rather than the never-treated group)—see Callaway and Sant'Anna (2021) and Callaway (2023) for more details. These issues are the same as in other work, so we do not elaborate on them here. However, they are often important in empirical work, and these extensions are all available in our code.

6. In cases where the researcher does not wish to specify parametric models for $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$, essentially the same estimator can be used but with machine learners or nonparametric estimators replacing the parametric models. In terms of algorithm, the only modification is to use cross-fitting where the outcome regression and the propensity score are estimated on a different subset of the data than the one used to estimate $ATT(g,t)$. Then, asymptotically normal estimation of $ATT(g,t)$ only requires fast-enough estimation of $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$. Many machine learners satisfy these conditions, given that these functions are smooth enough. See Chang (2020) and Callaway, Drukker, Liu, and Sant'Anna (2023) for results specific to DiD and Chernozhukov et al. (2018) and Rothe and Firpo (2019) for general results.

Dimension Reduction

The main practical challenge is that, by construction, the dimension of the covariates in $m_{g,t}(\mathbf{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z)$ is likely to be high as $\mathbf{X}$ is of dimension Tk where k is the number of time-varying covariates. For example, if there are five time-varying covariates, zero time-invariant covariates, and ten time periods, then the dimension of $\mathbf{X}$ is fifty. This suggests that, in most applications, reducing the dimension of the covariates will be desirable. One leading dimension-reducing assumption is that

$$m_{g,t}(\mathbf{X}, Z) = m_{g,t}(X_t\text{–}X_{g-1}, X_{g-1}, Z) \quad \text{and} \quad p_{g,t}(\mathbf{X}, Z) = p_{g,t}(X_t\text{–}X_{g-1}, X_{g-1}, Z)$$

which says that, in terms of time-varying covariates, the outcome regressions and generalized propensity scores only depend on (i) the change in the time-varying covariates from the base period to the current period and (ii) the level of the time-varying covariates in the base-period, rather than the covariates across all time periods. This type of specification includes both types of covariates that show up in Callaway and Sant'Anna (2021) and in imputation approaches such as Gardner, Thakral, Tô, and Yap (2023) and Borusyak, Jaravel, and Spiess (2024).

This is not the only possible choice, however. Another option is to assume that $m_{g,t}(\mathbf{X}, Z) = m_{g,t}(\bar{X}, Z)$ and $p_{g,t}(\mathbf{X}, Z) = p_{g,t}(\bar{X}, Z)$ where $\bar{X}$ is the average of the time-varying covariates across all time periods. Another alternative is to choose the covariates in the outcome regression and generalized propensity score in a data-driven way. In this vein, for one set of results in the application, we use the LASSO to choose the time-varying covariates to include in the outcome regression model and generalized propensity score, where we construct a dictionary of time-varying components based on their principal components.

In practice, a researcher can choose any of a number of approaches to deal with the high-dimension of the covariates. Although we think there are a few leading possibilities, rather than necessarily advocating a particular approach to dimension reduction, we instead emphasize that any approach to dimension reduction should be a carefully and transparently considered step of the analysis rather than being inherited from the estimation strategy as is the case with TWFE regressions.

7 Application

In this section, we consider a small-scale version of an application on stand-your-ground laws that we derive from Cheng and Hoekstra (2013). Although there is some variation across states regarding the specifics of particular stand-your-ground laws, in essence, a stand-your-ground law removes the duty to retreat in a potentially violent altercation. Cheng and Hoekstra (2013) study a period from 2000-2010 where 20 states implemented stand-your-ground laws. Before 2000, no states had implemented these policies, and they were implemented in a staggered fashion. Cheng and Hoekstra (2013) consider a number of outcomes in their paper, but here, we only consider one of their main outcomes: the number of homicides in a state. This is an interesting outcome as stand-your-ground laws have contrasting theoretical implications on homicides. On the one hand, there are some possible deterrence effects where stand-your-ground laws reduce the number of violent altercations, leading to fewer homicides. On the other hand, stand-your-ground laws could increase the deadliness of a given violent altercation, leading to more homicides. Cheng and Hoekstra (2013) find that stand-your-ground laws increase homicides.[25]

Below, we provide two sets of results using two different subsets of the data. First, we use a subset of the data that only includes the years 2000 and 2010. This first dataset is in line with our arguments above for the case of exactly two periods. While we report estimates of the effects of stand-your-grand policies on homicides, much of our main interest is in how different estimation strategies (based on *the same identification strategies*) balance the distribution of covariates which we are able to assess using the covariate balance diagnostics that we proposed for TWFE and AIPW earlier in the paper. Table 1, in the Appendix, provides summary statistics using the two-period subset of the data and for the full set of covariates used in Cheng and Hoekstra (2013). There are notable and large differences in several covariates across states that implemented a stand-your-ground law at some point relative to those that did not. The most notable differences are in terms of geography (treated states were substantially more likely to be in the South or Midwest), median income (treated states lower), poverty rate (treated states higher), number of prisoners (treated states higher), per capita welfare expenditures (treated states lower), and unemployment rate (treated states higher). There are also differences in how some covariates changed over time. Most notably, the poverty rate and the number of prisoners increased more in treated states, while median income decreased more in treated states. Finally, the summary statistics indicate moderate but nontrivial differences in population between treated and untreated states (treated states tending to be larger). For our second set of results, we mimic Cheng and Hoekstra (2013)'s setting much more closely—we use the full data of all 50 states across all available years, we

[25] Our application is not meant to replicate their paper as we make several simplifications and freely modify their approach to emphasize the contributions of the current paper. We discuss several additional differences and simplifications that we made in Appendix SD in the Supplementary Appendix.

use the same set of covariates as in one of their main specifications, and we use sampling weights in the analysis as in their paper.

Many of our results in this section are reported in figures that summarize covariate balance after applying the implicit weighting schemes for different estimation strategies that we developed earlier in the paper. The figures report the standardized difference between the treated and comparison group for each covariate being considered, where the standardized difference is the difference between the average value of the covariate for the treated group relative to the untreated group scaled by the pooled standard deviation of the covariate. We report both raw covariate balance and covariate balance after applying the implicit TWFE or AIPW weights. To give a sense of magnitude, standardized differences of around 0.1 or smaller are typically considered small differences, standardized differences around 0.3 are medium/important differences, and standardized differences of 0.5 or larger are considered large differences—see, for example, Imbens and Rubin (2015) for a textbook discussion.

7.1 Results with Two Periods and only Population and Region as Covariates

For the first set of results, we consider a highly simplified setting. The outcome is the log of the number of homicides in a state. We consider one time-varying covariate: the log of a state's population, and one time-invariant covariate: the Census region (Midwest, Northeast, South, or West) that a state is located in.[26] The intuition for the identification strategy in this section is that a researcher would like to estimate the impact of the policy by comparing the change in log homicides among treated and untreated states that have similar populations and are located in the same region of the country.[27]

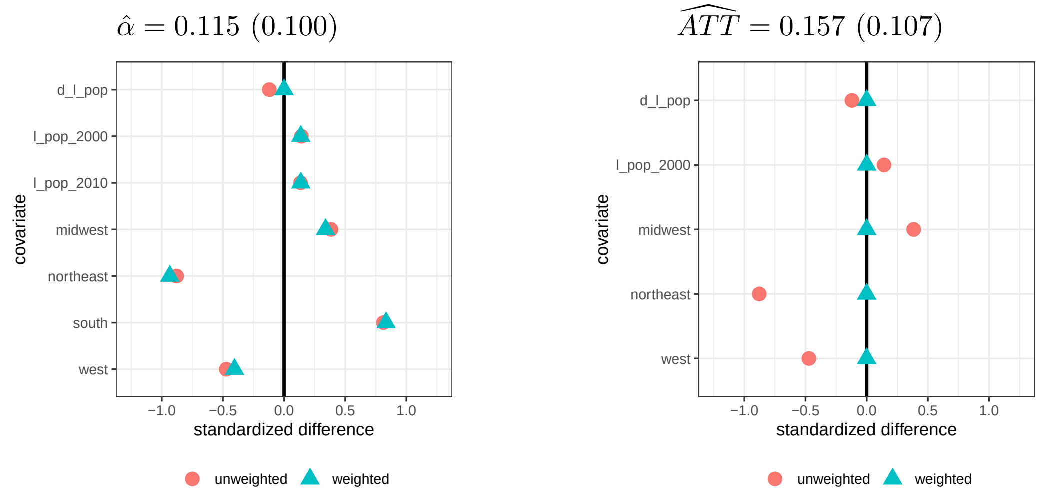
Figure 1 provides two sets of estimates meant to (i) summarize the causal effect of stand-your-ground laws on homicides and (ii) assess covariate balance inherited from each estimation strategy. Panel (a) contains results from the canonical TWFE regression in Equation (1) (recall that, in this simple setting with two periods, this just amounts to running a regression of the change in log homicides on a treatment indicator and the change in log population). Panel (b) contains results from the AIPW estimation strategy discussed in Section 6 where the outcome regression model is a linear model and the propensity score model is a logit model, and both include the change in log population, the level of log population in the first period, and region indicators as covariates. The results are qualitatively similar. The AIPW estimate is roughly 40% larger in magnitude, though neither estimate is statistically different from zero nor are the estimates statistically different from each other—none of these results are too surprising given that we are only using two time periods worth of data in this section.

More interesting, however, are the covariate balance results reported in the figure. Notice that, with just the raw data (the red circles in both panels), there is a relatively small amount of imbalance in the change in log population between the treated and untreated group, there are moderate differences in

[26] While it is common in empirical research to omit time-invariant covariates from a TWFE regression, covariates like Census region are an exception to this—they are often included as region-by-year fixed effects. In this section, we do not include region-by-year fixed effects in order to illustrate our results on specifications that do not include time-invariant covariates. In Appendix SD in the Supplementary Appendix, we provide analogous results for specifications that include region-by-year fixed effects.

[27] Because the TWFE regressions in this section do not explicitly include region, most empirical papers would not describe their identification strategy exactly as we have worded it in this sentence. However, many papers would argue that the TWFE regression effectively controls for *any* time-invariant covariate (i.e., region plus all other time-invariant characteristics of a state) because the TWFE regression includes a unit fixed effect.

Figure 1: Two Period Covariate Balance using TWFE and AIPW



(a) TWFE

(b) AIPW

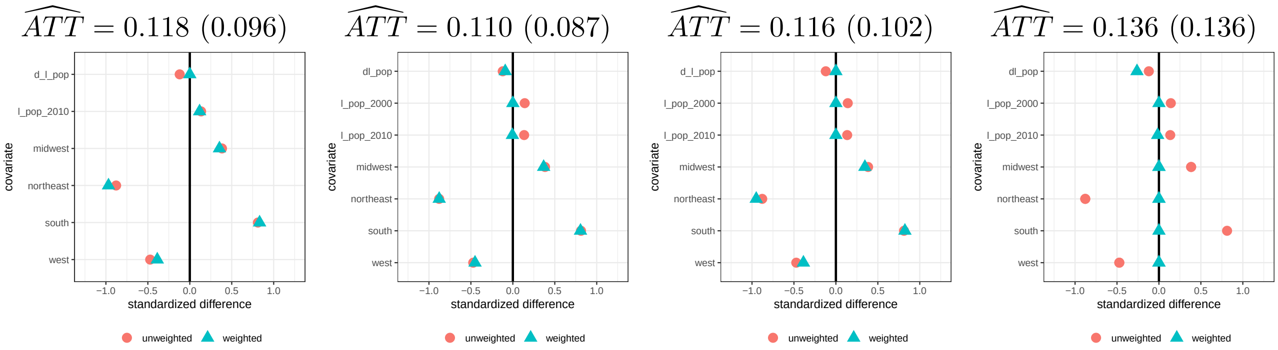
Notes: The figure reports estimates of the effects of stand-your-ground laws on homicides and covariate balance statistics using the two-period data discussed in the main text. The balance statistics are invariant to the outcome. Different covariates are displayed along the y-axis. `d_l_pop` is the change in the log of state-level population from 2000 to 2010; `l_pop_2000` and `l_pop_2010` are the level of the log of state-level population in 2000 and 2010, respectively; and **`midwest`**, **`northeast`**, **`south`**, **`west`** are indicators of Census region. The x-axis reports standardized differences for the mean of each covariate between the treated group and untreated group. The red circles provide the standardized difference for the raw difference, and the blue triangles show the standardized difference after applying the implicit weighting scheme from each estimation method. Panel (a) comes from regressing ΔY_{t^*} on D_{t^*} and ΔX_{t^*}. Panel (b) uses the AIPW estimation strategy discussed in the paper that includes $\Delta X_{t^*}, X_{t^*-1}$, and Z in both the outcome regression model and the propensity score.

terms of the levels of the log of population both in 2000 and 2010, and then there are large differences in the distribution of Census region. The results in Panel (a) from the TWFE regression balance the change in log population, which aligns with the theoretical properties of this kind of regression discussed earlier in the paper.[28] In terms of balancing the levels of the log of population, the regression essentially has no effect. The standardized difference is only 3% smaller using the implicit TWFE regression weights than in the raw data. In other words, *controlling for the change in the log of population in the TWFE regression does not result in the treated group and untreated being any more similar in terms of their levels of log population*—which, quite likely, was the main goal of including the state's population in the model to begin with. Similarly, the TWFE regression essentially does not affect the balance of the region indicators. Panel (b) provides results for the AIPW estimator proposed in the paper. This approach perfectly balances the means of all of the population and region covariates, and their averages are the same as the target population—this is by construction.

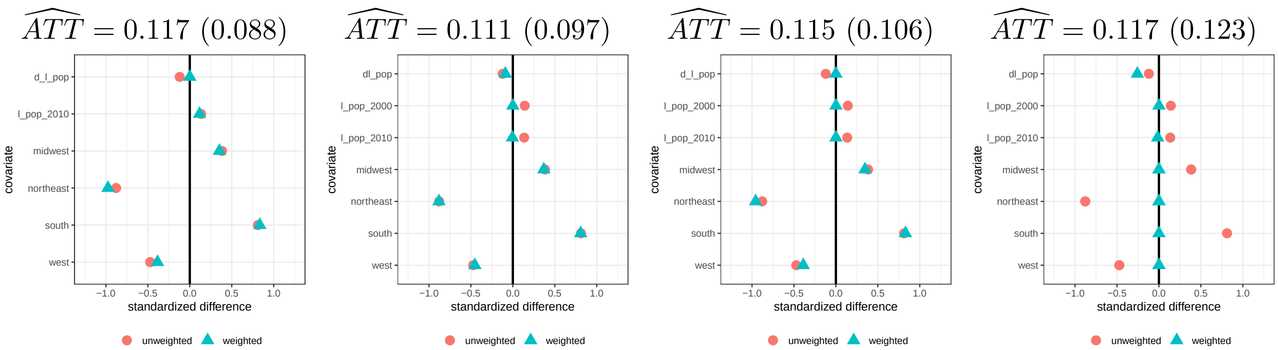
Given the results in the previous paragraph, an interesting follow-up question is: what is the main driver of the improved covariate balance across estimators? Figure 2 provides covariate balance statistics

[28] Recall that, ideally, the implicit weights will (i) balance the covariates between the treated and untreated group and (ii) make the covariates have the same distribution as for the treated group. All the figures in this section are geared toward checking (i). (ii) is satisfied by construction for all of the AIPW and regression adjustment estimators considered in this section, but it is not satisfied for the TWFE regression. This means that, here, while the implicit TWFE regression weights balance the change in log population, they do not result in the distribution of the change in log population being the same as it is for the treated group.

Figure 2: Additional Results for Two Period Covariate Balance using Regression Adjustment and AIPW



Reg. Adj. (a): ΔX_{t^*} Reg. Adj. (b): X_{t^*-1} Reg. Adj. (c): $\Delta X_{t^*},\ X_{t^*-1}$ Reg. Adj. (d): $X_{t^*-1},\ Z$



AIPW (a): ΔX_{t^*} AIPW (b): X_{t^*-1} AIPW (c): $\Delta X_{t^*}, X_{t^*-1}$ AIPW (d): $X_{t^*-1},\ Z$

Notes: The figure reports estimates of the effects of stand-your-ground laws on homicides and covariate balance statistics using the two-period data discussed in the main text. The balance statistics are invariant to the outcome. Different covariates are displayed along the y-axis. `d_l_pop` is the change in the log of state-level population from 2000 to 2010; `l_pop_2000` and `l_pop_2010` are the level of the log of state-level population in 2000 and 2010, respectively; and `midwest`, `northeast`, `south`, `west` are indicators of Census region. The x-axis reports standardized differences for the mean of each covariate between the treated group and untreated group. The red circles provide the standardized difference for the raw difference, and the blue triangles show the standardized difference after applying the implicit weighting scheme from each estimation method. The top row provides regression adjustment results with different covariate specifications indicated in each panel. The bottom row includes analogous results from AIPW estimation with different covariate specifications indicated in each panel; in each case, the same covariate specification is used for the propensity score as for the outcome regression model.

for eight different specifications which come from (i) either using regression adjustment or AIPW or (ii) varying the covariates included in the estimations among the following four sets of covariates: (a) the change in log population only, (b) the level of log population in 2000 only, (c) both the change in log population and the level of log population in 2000, and (d) the level of log population in 2000 and region. Two of these eight are particularly worth emphasizing. Specification (a), particularly in the regression adjustment case, corresponds to the specification used in imputation strategies that linearly include X_t. Specification (d), in either the regression adjustment or AIPW cases, corresponds to the default way to include covariates in Callaway and Sant'Anna (2021).

There are two main takeaways from this figure. First, in terms of covariate balance, regression adjustment and AIPW are very similar in all cases. Second, the regression adjustment specification (a), which only includes the change in log population, does not perform well in balancing the level of log population, especially compared to specifications that directly include the level of log population for at least one of the periods. On the other hand, Specification (d), which includes both the level of log population in 2000 and region, balances the covariates well, almost as well as our approach in Figure 1.

7.2 Results with More Periods and More Covariates

We use a specification much closer to the one used in Cheng and Hoekstra (2013) for a final set of results. First, we take the log of homicides per 100,000 people in the state as the outcome.[29] Second, we use sampling weights based on the state's average population across all time periods.[30] Third, we include a number of additional covariates: the log of the number of police per 100,000 population, the log of the number of incarcerated persons per 100,000, the log of government spending on assistance and subsidies per capita, the log of government spending on public welfare per capita, median household income, the poverty rate, the unemployment rate, and demographic indicators of the fraction of the state's population that are black males ages 15-24 and 25-44 or white males 15-24 and 25-44.[31] All three of these modifications come from Cheng and Hoekstra (2013). Another issue is that, because our unit of observation is the state, by construction, the size of many of our groups is very small. This results in AIPW estimation being infeasible (as it is impossible to estimate a generalized propensity score); therefore, we only report TWFE estimates and regression adjustment estimates. Finally, in line with Cheng and Hoekstra (2013) (but unlike the results above), all the estimates in this section include region-by-year fixed effects.

Figure 3 provides the results. First, relative to the previous results, there are larger differences in the estimates across different specifications. The TWFE estimate is positive and statistically different from zero, the regression adjustment specification that includes the changes in time-varying covariates (Panel (b)) is somewhat larger and marginally statistically significant, the regression adjustment specification that includes the pre-treatment levels of time-varying covariates (Panel (c)) is close to zero, and the regression adjustment specification that includes both levels and changes of time-varying covariates (Panel (d)) is roughly similar to the TWFE estimate though less precisely estimated. One source of differences between the TWFE results and the regression adjustment results is that the TWFE estimates are affected by violations of parallel trends in pre-treatment periods (i.e., the second term in Equation (18)). If we manually zero out the contribution of pre-treatment violations of parallel trends, then the TWFE estimate increases to 0.0879, a 31% increase.

The remaining differences are explained by different implicit weighting schemes. In the figure, covariate balance is assessed in terms of how well the implicit weights across all post-treatment periods balance the average of each covariate.[32] Relative to the raw data, the TWFE regression does improve

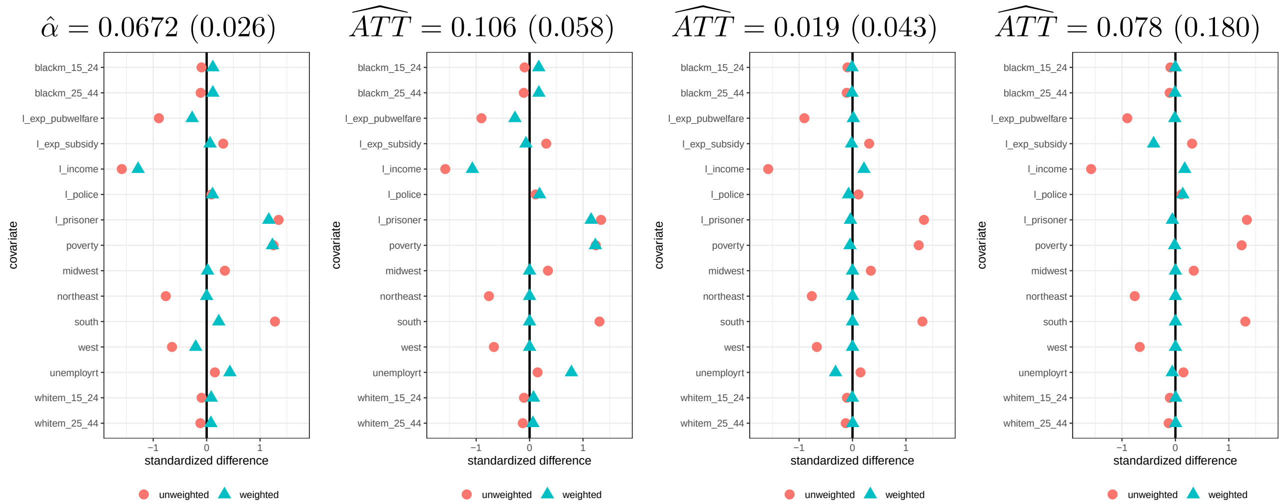
[29] The previous results did not normalize homicides to be per capita because population was included as a covariate.

[30] See Remark S8 in the Supplementary Appendix for additional discussion on how to extend our results presented above to include sampling weights.

[31] Regarding covariates, the only difference relative to Cheng and Hoekstra (2013) is that we do not include the lag of the log of the number of incarcerated persons per 100,000.

[32] In particular, for TWFE, we replace $(Y_t$–$Y_{g-1})$ in the first term in Equation (18) with $\bar{X}$. For AIPW, similarly, we replace $(Y_t$–$Y_{g-1})$ in Equation (21) with $\bar{X}$.

Figure 3: Multiple Period Covariate Balance with Additional Covariates



(a) TWFE

(b) Reg. Adj.: $\Delta X, Z$

(c) Reg. Adj.: X_{g-1}, Z

(d) Reg. Adj.: ΔX, X_{g-1}, Z

Notes: The figure reports estimates of the effects of stand-your-ground laws on homicides and covariate balance statistics using all available data from 2000-2010. The balance statistics are invariant to the outcome. Different covariates are displayed along the y-axis. See the main text as well as Table 1 for a detailed explanation of each covariate. The x-axis reports standardized differences for the mean of each covariate between the treated group and untreated group that come from our multi-period diagnostics for TWFE and regression adjustment/AIPW discussed in the main text. The red circles provide the standardized difference for the raw difference, and the blue triangles show the standardized difference after applying the implicit weighting scheme from each estimation method. We also use state-specific average population as sampling weights as in the main results in Cheng and Hoekstra (2013). The results in Panel (a) come from a TWFE regression that includes all the covariates listed in the figure. Panels (b)-(d) report regression adjustment results with different specifications for the covariates, as described in the main text.

covariate balance, though there are still some covariates that are severely unbalanced: median income, log of incarceration rate, and poverty rate, particularly. Regression adjustment that includes the change in covariates over time (Panel (b)) does not perform much better. The last two specifications perform better in terms of covariate balance. We also calculated the effective sample size for the untreated group for each of the regression adjustment specifications (see Remark S12 in the Supplementary Appendix for the specific calculation). The effective sample size across post-treatment periods is 63.1, 26.7, and 9.9 for the specifications in Panels (b), (c), and (d), respectively. That the effective sample size drops off substantially for the specification that includes both levels and changes of time-varying covariates is the likely explanation for the dramatic increase in the standard errors in Panel (d).[33] Taken together, the covariate balance and effective sample size results suggest that the specification in Panel (c) is the most suitable for this application.

Discussion

A main takeaway from our application is that, in terms of controlling for particular covariates in the parallel trends assumption, a first-order concern is the functional form under which the covariates enter

[33]The decrease in effective sample sizes between Panels (b)-(d) is not surprising. In our application, we have more imbalance in the levels of the covariates than in their changes, suggesting that it is "more difficult" to balance levels than changes in covariates; and it is still more difficult to balance both levels and changes.

the model. Approaches that inherit transformed covariates as a byproduct of the estimation strategy, whether it be TWFE, imputation/regression adjustment, or AIPW, perform poorly (at least in our example) in terms of balancing the levels of time-varying covariates or time-invariant covariates. On the other hand, regression adjustment and AIPW approaches that include *any* level of a time-varying covariate and time-invariant covariates (such as the default implementation of Callaway and Sant'Anna (2021)) performed substantially better. In some cases, including levels *and* changes in time-varying covariates and time-invariant covariates, performed better, although this was not uniformly true. We conjecture that a good heuristic for empirical work is to always include some version of the level of time-varying covariates (this could be a pre-treatment value of the covariates, its average across all time periods, or some other measure) and time-invariant covariates in the estimated model; then, in applications with enough data, one should then consider including changes in time-varying covariates as well.

8 Conclusion

We have considered difference-in-differences identification strategies when (i) the identification strategy hinges on comparing treated and untreated units with the same observed covariates and (ii) these covariates include time-varying and/or time-invariant variables. In this empirically common setting, researchers have most often implemented this identification strategy using a TWFE regression like the one in Equation (1). In the current paper, we have demonstrated a number of potential weaknesses of TWFE regressions in this context. Some of these weaknesses, such as lack of robustness to multiple periods and variation in treatment timing or being reliant on certain linearity conditions, are likely not surprising given existing work in the difference-in-differences literature. However, we also document several other weaknesses that we refer to as "hidden linearity bias." Hidden linearity bias arises because the transformations used to eliminate the unit fixed effect in the TWFE regression also change the functional form of the covariates. This transformation thus either effectively changes the identification strategy (to one where only the change in time-varying covariates is included in the parallel trends assumption) or relies heavily on a correctly specified linear model. It is not common in empirical work to engage with whether or not these conditions are reasonable—indeed, in most applications, these are likely to be strong and undesirable extra assumptions. We proposed several diagnostic tools for assessing the sensitivity of TWFE regressions that include covariates to hidden linearity bias. We also proposed an alternative estimation strategy, building on recent work in the DiD literature, that does not suffer from hidden linearity bias, does not require any auxiliary assumptions along the lines mentioned above, and is effectively no more complicated to implement in practice than the TWFE regression.

References

Abadie, Alberto (2005). "Semiparametric difference-in-differences estimators". *The Review of Economic Studies* 72.1, pp. 1–19.

Almond, Douglas, Hilary W Hoynes, and Diane Whitmore Schanzenbach (2011). "Inside the war on poverty: The impact of food stamps on birth outcomes". *The Review of Economics and Statistics* 93.2, pp. 387–403.

Angrist, Joshua D (1998). "Estimating the labor market impact of voluntary military service using Social Security data on military applicants". *Econometrica* 66.2, pp. 249–288.

Angrist, Joshua D and Jorn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Aronow, Peter M and Cyrus Samii (2016). "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60.1, pp. 250–267.

Bailey, Martha J and Andrew Goodman-Bacon (2015). "The War on Poverty's experiment in public medicine: Community health centers and the mortality of older Americans". *American Economic Review* 105.3, pp. 1067–1104.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2022). "When is TSLS actually late?" Working Paper.

Blundell, Richard and Monica Costa Dias (2009). "Alternative approaches to evaluation in empirical microeconomics". *Journal of Human Resources* 44.3, pp. 565–640.

Bonhomme, Stephane and Ulrich Sauder (2011). "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling". *Review of Economics and Statistics* 93.2, pp. 479–494.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2024). "Revisiting event-study designs: Robust and efficient estimation". *Review of Economic Studies*, rdae007.

Caetano, Carolina, Brantly Callaway, Robert Payne, and Hugo Sant'Anna (2022). "Difference in differences with time-varying covariates". Working Paper.

Callaway, Brantly (2023). "Difference-in-differences for policy evaluation". *Handbook of Labor, Human Resources and Population Economics.* Ed. by Zimmermann, Klaus F. Springer International Publishing, pp. 1–61.

Callaway, Brantly, David Drukker, Di Liu, and Pedro HC Sant'Anna (2023). "Double/debiased machine-learning estimator for difference-in-difference with multiple periods". Working Paper.

Callaway, Brantly and Pedro HC Sant'Anna (2021). "Difference-in-differences with multiple time periods". *Journal of Econometrics* 225.2, pp. 200–230.

Chang, Neng-Chieh (2020). "Double/debiased machine learning for difference-in-differences models". *The Econometrics Journal* 23.2, pp. 177–191.

Chattopadhyay, Ambarish, Noah Greifer, and Jose R Zubizarreta (2023). "lmw: Linear model weights for causal inference". Working Paper.

Chattopadhyay, Ambarish and José R Zubizarreta (2023). "On the implied weights of linear regression for causal inference". *Biometrika* 110.3, pp. 615–629.

Cheng, Cheng and Mark Hoekstra (2013). "Does strengthening self-defense law deter crime or escalate violence? Evidence from expansions to Castle Doctrine". *Journal of Human Resources* 48.3, pp. 821–854.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). "Double/debiased machine learning for treatment and structural parameters". *The Econometrics Journal* 21.1, pp. C1–C68.

de Chaisemartin, Clement and Xavier D'Haultfœuille (2020). "Two-way fixed effects estimators with heterogeneous treatment effects". *American Economic Review* 110.9, pp. 2964–2996.

— (2023). "Two-way fixed effects regressions with several treatments". forthcoming at *Journal of Econometrics.*

Gardner, John, Neil Thakral, Linh T Tô, and Luther Yap (2023). "Two-stage differences in differences". Working Paper.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2022). "Contamination bias in linear regressions". Working Paper.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225.2, pp. 254–277.

Goodman-Bacon, Andrew and Jamein P Cunningham (2019). "Changes in family structure and welfare participation since the 1960s: The role of legal services". Working Paper.

Hahn, Jinyong (2023). "Properties of least squares estimator in estimation of average treatment effects". *SERIEs* 14.3, pp. 301–313.

Hainmueller, Jens (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". *Political analysis*, pp. 25–46.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing selection bias using experimental data". *Econometrica* 66.5, pp. 1017–1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". *The Review of Economic Studies* 64.4, pp. 605–654.

Hirano, Keisuke, Guido Imbens, and Geert Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score". *Econometrica* 71.4, pp. 1161–1189.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007). "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". *Political analysis* 15.3, pp. 199–236.
Imai, Kosuke and Marc Ratkovic (2014). "Covariate balancing propensity score". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 243–263.
Imbens, Guido and Jeffrey Wooldridge (2009). "Recent developments in the econometrics of program evaluation". *Journal of Economic Literature* 47.1, pp. 5–86.
Imbens, Guido W and Donald B Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
Ishimaru, Shoya (2022). "What do we get from a two-way fixed effects estimator? Implications from a general numerical equivalence". Working Paper.
— (2024). "Empirical decomposition of the IV-OLS gap with heterogeneous and nonlinear effects". *Review of Economics and Statistics*, pp. 1–16.
Kline, Patrick (2011). "Oaxaca-Blinder as a reweighting estimator". *American Economic Review* 101.3, pp. 532–37.
Lin, Lihua and Zhengyu Zhang (2022). "Interpreting the coefficients in dynamic two-way fixed effects regressions with time-varying covariates". *Economics Letters* 216, p. 110604.
Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt (2019). "Poorly measured confounders are more useful on the left than on the right". *Journal of Business & Economic Statistics* 37.2, pp. 205–216.
Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky (2007). "Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable". *Statistical Science* 22.4, pp. 544–559.
Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American Statistical Association* 89.427, pp. 846–866.
Rosenbaum, Paul and Donald Rubin (1983). "The central role of the propensity score in observational studies for causal effects". *Biometrika* 70.1, pp. 41–55.
Rothe, Christoph and Sergio Firpo (2019). "Properties of doubly robust estimators when nuisance functions are estimated nonparametrically". *Econometric Theory* 35.5, pp. 1048–1087.
Rubin, Donald B (2008). "For objective causal inference, design trumps analysis". *The Annals of Applied Statistics* 2.3, pp. 808–840.
Sant'Anna, Pedro H. C. and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". *Journal of Econometrics* 219.1, pp. 101–122.
Scharfstein, Daniel O, Andrea Rotnitzky, and James M Robins (1999). "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *Journal of the American Statistical Association* 94.448, pp. 1096–1120.
Słoczyński, Tymon (2022). "Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights". *The Review of Economics and Statistics* 104.3, pp. 501–509.
Słoczyński, Tymon and Jeffrey M Wooldridge (2018). "A general double robustness result for estimating average treatment effects". *Econometric Theory* 34.1, pp. 112–133.
Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225.2, pp. 175–199.
Wooldridge, Jeff (2021). "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators". Working Paper.

A Proofs of Results with Two Periods

A.1 TWFE Regressions with Two Periods

To start with, we provide a useful lemma for our arguments involving linear projections.

Lemma 1. *Under Assumptions 1 and 2, for* $d \in \{0, 1\}$,

$$\mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_d(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D = d\Big] = \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\Delta Y_{t^*}\Big|D = d\Big]$$

The proof of Lemma 1 is provided in the Supplementary Appendix. Next, we provide a useful lemma for dealing with the denominator in the expression for α in Equation (7).

Lemma 2. *Under Assumptions 1 and 2,*

$$\mathbb{E}\Big[\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)^2\Big] = \mathbb{E}\Big[1 - \mathrm{L}(D|\Delta X_{t^*})\Big|D = 1\Big]\pi$$

The proof of Lemma 2 is provided in the Supplementary Appendix.

Over the following two propositions, we provide two decompositions of α that serve as important steps for showing our result in Theorem 1.

Proposition A1. *Under Assumptions 1 and 2, α from the regression in Equation (6) can be decomposed as*

$$\alpha = \mathbb{E}\left[w(\Delta X_{t^*})\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D = 1\right]$$

where $w(\Delta X_t)$ are the same weights as in Theorem 1.

Proof. Starting with the numerator from Equation (7), we have that

$$\begin{aligned}
&\mathbb{E}\Big[\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big] \\
&\quad = \mathbb{E}\Big[\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big|D = 1\Big]\pi - \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\Delta Y_{t^*}\Big|D = 0\Big](1-\pi) \\
&\quad = \mathbb{E}\Big[\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D = 1\Big]\pi \\
&\qquad - \mathbb{E}\Big[\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D = 0\Big](1-\pi) \\
&\quad = \mathbb{E}\Big[D\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big] - \mathbb{E}\Big[(1-D)\mathrm{L}(D|\Delta X_{t^*})\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big] && (22) \\
&\quad = \mathbb{E}\left[D\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\right] \\
&\qquad + \mathbb{E}\left[\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right] \\
&\quad = \mathbb{E}\left[\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D = 1\right]\pi && (23)
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the second equality holds by Lemma 1, the third equality holds by applying the law of iterated expectations to both terms, the fourth equality holds by adding and subtracting $\mathbb{E}\Big[D\big(1 - \mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big]$, and the last equality holds by applying the law of iterated expectations for the first term and because

$$\mathbb{E}\left[\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\right] = \mathbb{E}\left[\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)\Delta X_{t^*}'\right]\beta_0 = 0$$

where the first equality holds by the definition of $\mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})$, and the second equality holds because ΔX_{t^*} is uncorrelated with the projection error $\big(D - \mathrm{L}(D|\Delta X_{t^*})\big)$.

Combining the expression in Equation (23) with the expression for the denominator from Lemma 2 completes the proof, given the definition of the weights $w(\Delta X_{t^*})$. □

Proposition A1 says that α, the coefficient on the treatment variable in the TWFE regression in Equation (6), is equal to a weighted average of the linear projection of the change in outcomes over time on the change in covariates over time for the treated group relative to the linear projection of the change in outcomes over time on the change in covariates over time for the untreated group. One notable feature of this decomposition is that it is straightforward to compute both the weights and the linear projection terms that show up in the decomposition (notice that the weights themselves only depend on linear projections). It also serves as an important intermediate step for establishing our results on interpreting α in terms of underlying causal effect parameters below.

Proposition A2. *Under Assumptions 1 and 2, α from the regression in Equation (6) can be decomposed as*

$$\alpha = \mathbb{E}\Big[w(\Delta X_{t^*})\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\Big)\Big|D=1\Big] \tag{24}$$

$$+ \mathbb{E}\Big[w(\Delta X_{t^*})\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D=1\Big] \tag{25}$$

where $w(\Delta X_{t^*})$ are the same weights as in Theorem 1.

Proof. Starting from the numerator of the expression in Proposition A1, we have that

$$\begin{aligned}
&\mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*}) - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big|D=1\right]\pi \\
&\quad= \mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\Big)\Big|D=1\right]\pi \\
&\quad- \mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1] - \mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\right. \qquad (26)\\
&\qquad\qquad\left. -\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0] - \mathrm{L}_0(\Delta Y_{t^*}|\Delta X_{t^*})\Big)\Big\}\Big|D=1\right]\pi
\end{aligned}$$

which holds by adding and subtracting

$$\mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\Big)\Big|D=1\right]\pi$$

Next, notice that

$$\mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1]\Big|D=1\right] = \mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big|D=1\right] \tag{27}$$

which holds by the law of iterated expectations. Further, notice that

$$\mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\mathrm{L}_1(\Delta Y_{t^*}|\Delta X_{t^*})\Big|D=1\right] = \mathbb{E}\left[\big(1-\mathrm{L}(D|\Delta X_{t^*})\big)\Delta Y_{t^*}\Big|D=1\right] \tag{28}$$

which holds by Lemma 1. Plugging Equations (27) and (28) back into Equation (26) implies that the expression in that line is equal to 0. Therefore, combining that expression with the one for the denominator in Equation (7) from Lemma 2 completes the proof. □

Like Proposition A1, Proposition A2 is a decomposition in that the parallel trends assumption is not used here, and, therefore, α is, in general, equal to the expression provided in Proposition A2. Unlike Proposition A1, Proposition A2 includes conditional expectation terms that may be challenging to estimate in applications (especially without invoking extra functional form assumption). It says that α from the TWFE regression in Equation (6) consists of two terms. The first is a weighted average of the average path of outcomes for the treated group relative to the path of outcomes for the untreated group (conditional on having the same time-varying and time-invariant covariates). The second is a weighted average (where the weights are the same as for the first term) of the difference between the average path of outcomes for the untreated group conditional on time-varying and time-invariant covariates and the linear projection of ΔY_{t^*} on ΔX_{t^*} for the untreated group.

Proof of Theorem 1. First, it immediately follows from Assumptions 1 to 3 that $ATT(X_{t^*}, X_{t^*-1}, Z) = \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=1] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]$. This implies that Equation (24) in Proposition A2 is equal to $\mathbb{E}[w(\Delta X_{t^*})ATT(X_{t^*}, X_{t^*-1}, Z)|D=1]$. The second term comes from adding and subtracting terms to the expression in Equation (25) in Proposition A2. The properties of the weights hold immediately by their definitions. □

Proof of Theorem 2. The result holds immediately from Theorem 1 by noticing that Assumption 4 directly implies that Equations (A) to (C) are all equal to 0. □

A.2 AIPW Diagnostics with Two Periods

We start by providing a lemma that we use to prove Proposition 1 that is related to reformulating regression adjustment as a weighting estimator. See Remark S1 in the Supplementary Appendix for more details.

Lemma 3. *To conserve on notation, let $X = (X_{t^*}, X_{t^*-1}, Z)$. Under Assumptions 1 and 2,*

$$\mathbb{E}\left[\mathrm{L}_0(\Delta Y_{t^*}|X)\Big|D=1\right] = \mathbb{E}\left[\vartheta_0^{\mathrm{L}_0}\Delta Y_{t^*}\Big|D=0\right]$$

where $\vartheta_0^{\mathrm{L}_0}$ are weights which are defined as

$$\vartheta_0^{\mathrm{L}_0} := \frac{(1-\pi)}{\pi}\gamma_0' X$$

where γ_0 denotes the linear projection coefficient from projecting $p(X)/(1-p(X))$ on X among the untreated group. In addition, $\mathbb{E}[\vartheta_0^{\mathrm{L}_0}|D=0] = 1$ (i.e., the weights have mean one), and it is possible that the weights can be negative.

The proof of Lemma 3 is provided in the Supplementary Appendix.

Proof of Proposition 1. To conserve on notation, let $X = (X_{t^*}, X_{t^*-1}, Z)$. Then, we can re-write $\widetilde{ATT}$ as

$$\widetilde{ATT} = \mathbb{E}\left[\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X)\Big|D=1\right] - \mathbb{E}\left[\tilde{w}_0^{aipw}\big(\Delta Y_{t^*} - \mathrm{L}_0(\Delta Y_{t^*}|X)\big)\Big|D=0\right]$$

Re-arranging terms, we have that

$$\begin{aligned}\widetilde{ATT} &= \mathbb{E}\left[\Delta Y_{t^*}\Big|D=1\right] - \mathbb{E}\left[\mathrm{L}_0(\Delta Y_{t^*}|X)\Big|D=1\right] - \mathbb{E}\left[\tilde{w}_0^{aipw}\Delta Y_{t^*}\Big|D=0\right] + \mathbb{E}\left[\tilde{w}_0^{aipw}\mathrm{L}_0(\Delta Y_{t^*}|X)\Big|D=0\right] \\ &=: A - B - C + D \end{aligned} \tag{29}$$

Terms A and C are straightforward to deal with as they are already weighted averages of ΔY_{t^*}. For Term B, from Lemma 3 we have that

$$B = \mathbb{E}\left[\frac{(1-\pi)}{\pi}\gamma_0' X \Delta Y_{t^*}\Big|D=0\right]$$

where γ_0 is the projection coefficient from projecting $p(X)/(1-p(X))$ on X among the untreated group. Furthermore, notice that

$$\mathbb{E}\left[\frac{(1-\pi)}{\pi}\gamma_0'X\Big|D=0\right] = \mathbb{E}\left[\frac{(1-\pi)}{\pi}\frac{p(X)}{(1-p(X))}\Big|D=0\right] = 1$$

where the first equality holds because $\gamma_0'X$ is the linear projection of $p(X)/(1-p(X))$ on X among the untreated group, and, hence, the corresponding projection errors have mean zero for the untreated group, and the second equality holds by repeated application of the law of iterated expectations. Thus, we have that

$$B = \mathbb{E}\left[\vartheta_{0,B}^{aipw}\Delta Y_{t^*}\Big|D=0\right] \quad \text{where} \quad \vartheta_{0,B}^{aipw} = \frac{\gamma_0'X}{\mathbb{E}[\gamma_0'X|D=0]} \tag{30}$$

Next, turning to the "numerator" of Term D, notice that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\varpi}_0^{aipw}\mathrm{L}_0(\Delta Y_{t^*}|X)\Big|D=0\right] &= \frac{(1-\pi)}{\pi}\mathbb{E}\left[\frac{\tilde{p}(X)}{(1-\tilde{p}(X))}X'\mathbb{E}[XX'|D=0]^{-1}\mathbb{E}[X\Delta Y_{t^*}|D=0]\Big|D=0\right] \\
&= \frac{(1-\pi)}{\pi}\underbrace{\mathbb{E}\left[\frac{\tilde{p}(X)}{(1-\tilde{p}(X))}X'\Big|D=0\right]\mathbb{E}[XX'|D=0]^{-1}}_{\tilde{\gamma}_0'}\mathbb{E}[X\Delta Y_{t^*}|D=0] \\
&= \mathbb{E}\left[\frac{(1-\pi)}{\pi}\tilde{\gamma}_0'X\Delta Y_{t^*}\Big|D=0\right]
\end{aligned} \tag{31}$$

where the first equality holds by the definitions of $\tilde{\varpi}_0^{aipw}$ and $\mathrm{L}_0(\Delta Y_{t^*}|X)$, the second equality removes the nonrandom terms from the expectation, and the last equality holds by the definition of $\tilde{\gamma}_0$. Furthermore, for the "denominator" of Term D, notice that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\varpi}_0^{aipw}\Big|D=0\right] &= \frac{(1-\pi)}{\pi}\mathbb{E}\left[\frac{\tilde{p}(X)}{(1-\tilde{p}(X))}\Big|D=0\right] \\
&= \mathbb{E}\left[\frac{(1-\pi)}{\pi}\tilde{\gamma}_0'X\Big|D=0\right]
\end{aligned} \tag{32}$$

where the first equality holds by the definition of $\tilde{\varpi}_0^{aipw}$, and the second equality holds because $\tilde{\gamma}_0'X$ is the linear projection of $\tilde{p}(X)/(1-\tilde{p}(X))$ on X among the untreated group which has mean zero projection errors conditional on $D=0$. Combining the expressions in Equations (31) and (32), we have that

$$D = \mathbb{E}\left[\vartheta_{0,D}^{aipw}\Delta Y_{t^*}\Big|D=0\right] \quad \text{where} \quad \vartheta_{0,D}^{aipw} = \frac{\tilde{\gamma}_0'X}{\mathbb{E}[\tilde{\gamma}_0'X|D=0]} \tag{33}$$

Combining Equations (30) and (33) with Equation (29) yields the first part of the result. That the weights have mean one follows because (i) $\vartheta_1^{aipw} = 1$ and (ii) each of the components of ϑ_0^{aipw} have mean one; there are three of these terms, two are added and one is subtracted as in Equation (29). Furthermore, given that two of the components of ϑ_0^{aipw} involve linear projections, the weights can be negative. Finally, we show that the weights balance the covariates in Lemma S1 in the Supplementary Appendix. $\square$

Table 1: Summary Statistics

	Treated	Untreated	Difference	Std. Diff.
Outcomes				
log homicides (2000)	5.188	4.696	0.492	0.337
Δ log homicides	0.081	-0.030	0.111	0.337
Time-Invariant Covariates				
Midwest	0.333	0.172	0.161	0.384
Northeast	0.000	0.310	-0.310	-0.878
South	0.524	0.172	0.351	0.812
West	0.143	0.345	-0.202	-0.472
Levels of Time-Varying Covariates (2000)				
log population	15.114	14.971	0.143	0.142
log police	5.746	5.711	0.034	0.168
log prisoners	6.135	5.817	0.318	0.782
log welfare exp. per capita	6.840	6.973	-0.133	-0.442
log subsidies per capita	4.583	4.693	-0.109	-0.210
log median income	10.778	10.926	-0.149	-1.084
poverty rate	12.586	9.986	2.599	1.055
unemployment rate	4.052	3.690	0.363	0.417
% black males 15-24	1.988	2.834	-0.847	-0.196
% black males 25-44	3.506	5.309	-1.803	-0.223
% white males 15-24	11.307	10.429	0.878	0.050
% white males 25-44	24.515	24.577	-0.062	-0.002
Changes in Time-Varying Covariates				
Δ log population	0.110	0.117	-0.007	-0.120
Δ log police	-0.005	0.001	-0.006	-0.065
Δ log prisoners	0.096	0.037	0.060	0.425
Δ log welfare exp. per capita	0.398	0.355	0.043	0.202
Δ log subsidies per capita	0.265	0.166	0.099	0.228
Δ log median income	-0.076	-0.040	-0.037	-0.426
Δ poverty rate	2.684	2.111	0.573	0.462
Δ unemployment rate	4.933	4.855	0.078	0.040
Δ % black males 15-24	0.149	0.180	-0.031	-0.065
Δ % black males 25-44	-0.272	-0.663	0.390	0.342
Δ % white males 15-24	-0.359	0.274	-0.633	-0.340
Δ % white males 25-44	-3.594	-3.903	0.309	0.039

Notes: The table provides summary statistics for the outcomes, time-invariant covariates, and levels and changes in time-varying covariates. States are classified as being treated or untreated based on their treatment status in 2010. The column 'Difference' reports the difference between the average of each variable for the treated group relative to the untreated group. The column 'Std. Diff.' reports the standardized difference of each variable for the treated group relative to the untreated group, which is the difference divided by the pooled standard deviation.