Difference-in-Differences with Time Varying Covariates and Empirical Applications

by

Stroud Payne

(Under the Direction of Brantly Callaway)

Abstract

This paper examines how to implement difference-in-differences techniques when there are time-varying covariates. Two-way fixed effects (TWFE) models are popular in the current literature but have been shown to have biased results when solving models with time-varying covariates. This paper presents conditions under which researchers can still recover the average treatment effect of the treated (ATT) of some treatment when there are time-varying covariates and provides doubly robust estimators that work with these assumptions. In addition, the paper offers an example for how to use imputation techniques to estimate difference-in-differences models, using a data set on stand-your-ground laws from Cheng and Hoekstra, 2013. Imputation involves using untreated data to make predictions on the untreated potential outcomes of treated units. The paper also provides a proof for the asymptotic normality of imputation techniques in both the two-period case and the multiple-period case.

Index words: Difference-in-Differences, Two-way Fixed Effects, Imputation, Covariates, Potential Outcomes, Parallel Trends

Difference-in-Differences with Time Varying Covariates and Empirical Applications

by

Stroud Payne

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree

Master of Arts

Athens, Georgia

2022

Difference-in-Differences with Time Varying Covariates and Empirical Applications

by

Stroud Payne

Major Professor:	Brantly Callaway
Committee:	Carolina Caetano
	Leonard Goff

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2022

Acknowledgments

I would like to thank Brantly Callaway, Carolina Caetano, and Hugo Rodrigues, who cowrote Chapters 1, 2, 3, 4, and the "Main Proofs" section of the Appendix. I would like to also thank Leonard Goff for his advice and contributions to this paper.

Contents

Chapter 1

Introduction

In this paper, we study difference in differences identification strategies where (i) the parallel trends assumption holds only after conditioning on covariates, (ii) some or all of these covariates vary over time, and (iii) some of the time varying covariates could themselves be affected by the treatment.

A number of papers (e.g., Heckman et al., 1998, Sant'Anna and Zhao, 2020 Abadie, 2005) show that certain causal effect parameters, typically the average treatment effect on the treated (ATT), are identified under conditional parallel trends assumptions. These types of conditional parallel trends assumptions are attractive in applications where the path of untreated potential outcomes may differ among units with different characteristics. However, work in the econometrics literature typically considers the case where covariates involved in the parallel trends assumption either do not vary over time or are "pre-treatment" (that is, the value of a time-varying covariate is set to its value in the pre-treatment period; see Bonhomme and Sauder, 2011, Lechner, 2011 for some discussions on using pre-treatment values of time-varying covariates). In contrast, empirical work in economics often only includes covariates that vary over time. In this case, identification must implicitly assume that the treatment does not have an effect on the covariates themselves, which is implausible in some applications.

Covariates that could have been affected by participating in the treatment are often referred to as "post-treatment" or as "bad controls." The received wisdom seems to be that this type of covariate should not be included in empirical research.[1] However, we provide several examples below where it seems important to condition on the value of the covariate that *would have occurred in the absence of the treatment*; in these cases, it would not generally be sufficient to just "not include" this sort of covariate. We propose several different strategies for dealing with time-varying covariates that show up in the parallel trends assumption while also potentially being affected by the treatment.

[1]For example, Angrist and Pischke, 2008 discuss "bad controls" in the context of deciding whether or not to control for occupation when studying causal effects of graduating from college on earnings. In that case, occupation is likely to be affected by attending college and, therefore, can make comparisons in earnings among those with the same occupation who graduated or did not graduate from college hard to interpret (even if college were randomly assigned). Angrist and Pischke, 2008 note that "...we would do better to control only for variables that are not themselves caused by education." We return to a related example later in this section on the effect of job displacement on earnings where occupation is potentially affected by job displacement.

Difference in differences identification strategies are most often implemented using two-way fixed effects (TWFE) regressions. The most common version of a TWFE regression that includes covariates is the following

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + X_{it}'\beta + v_{it} \tag{1.1}$$

where θ_t is a time fixed effect, η_i is individual-level unobserved heterogeneity (i.e., an individual fixed effect), D_{it} is the treatment indicator, and X_{it} are time varying covariates. In the TWFE regression in Equation (1.1), α is the parameter of interest and it is often interpreted as "the causal effect of the treatment" or at least would be hoped to be a weighted average of underlying heterogeneous treatment effects. Being able to include covariates is one of the main attractions of using a TWFE regression to implement a DID design. For example, Angrist and Pischke, 2008 write: "A second advantage of regression-DD is that it facilitates empirical work with regressors other than switched-on/switched off dummy variables."[2] TWFE regressions have come under much scrutiny in recent work in terms of how well they perform for implementing DID identification strategies. In particular, TWFE regressions can perform very poorly in the presence of more than two time periods, variation in treatment timing across units, and treatment effect heterogeneity (particularly, treatment effect dynamics); see Goodman-Bacon, 2021, de Chaisemartin and D'Haultfœuille, 2020. Although with only two time periods, TWFE regressions are known to be reliable under unconditional parallel trends, here we point out a number of problems with TWFE regressions for implementing DID identification strategies that rely on conditional parallel trends assumptions *even in the case with only two time periods*.

In particular, we show that TWFE regressions can deliver poor estimates of the average treatment effect on the treated (which is the natural target parameter for DID identification strategies) for any of four reasons: (1) time-varying covariates that are themselves affected by the treatment, (2) ATTs and/or parallel trends assumptions that depend on the pre-treatment level of time varying covariates in addition to (or instead of) only the change in the covariates over time, (3) ATTs and/or paths of untreated potential outcomes that depend on time-invariant covariates, and (4) violations of strong functional form assumptions both for outcomes over time and for the propensity score. All four of these issues are common in applications in economics.

In applications where none of the four issues mentioned above occur, TWFE regressions deliver a weighted average of conditional ATTs where all the weights are positive. However, even in this best-case scenario, TWFE regressions still suffer from a "weight-reversal" property similar to the one pointed out in Słoczyński, 2020 under unconfoundedness with cross-sectional data. In our case, conditional ATTs for relatively uncommon values of the covariates among the treated group (relative to the untreated group) are given large weights while conditional ATTs for common values of the covariates among the treated group are given small weights. In order to get around this weight reversal issue, one needs to additionally rule out heterogeneous treatment effects across different values of the covariates. Adding this condition to the previous four implies that TWFE regressions deliver the ATT; however, we stress that these are a

[2] Angrist and Pischke, 2008 also briefly mention "bad controls" in the context of difference in differences (Section 5.2.1).

very stringent set of requirements for TWFE regressions to perform well for estimating the ATT when the parallel trends assumption depends on time-varying covariates.

We propose several new strategies for dealing with time-varying covariates that are required for the parallel trends assumption to hold. When the researcher is confident that the covariates evolve exogenously with respect to the treatment, we provide a doubly robust estimand for the ATT (these arguments are similar to the ones in Sant'Anna and Zhao, 2020 for the case with time invariant covariates). Doubly robust estimators have the property that they deliver consistent estimates of the ATT if *either* an outcome regression model or a propensity score model is correctly specified, thus giving researchers an extra chance to correctly specify a model relative to regression adjustment or propensity score weighting strategies. Besides this, our doubly robust estimands can also be used in the context of the double/debiased machine learning literature where the propensity score and outcome regression model can be estimated using a wide variety of modern machine learning techniques (see Chernozhukov et al., 2018 for the general case and Chang, 2020 in the context of DID).[3] When the time-varying covariates can be affected by the treatment, we provide sufficient (and easy-to-interpret) conditions under which the strategy of conditioning on "pre-treatment" covariates, which is common in the econometrics literature, is justified. We also discuss other cases where this strategy is not reasonable. In these cases, we propose regression adjustment-type and doubly robust-type expressions for the ATT. Finally, when a researcher is willing to make an additional functional form assumption for untreated potential outcomes, we propose some even simpler approaches based on regression adjustment (these approaches are also broadly similar to recent "imputation estimators" proposed in Liu et al., 2021, Gardner, 2021, Borusyak et al., 2021). We also show that stronger functional form assumptions for the model for untreated potential outcomes can allow for parallel trends-type assumptions for the covariates to be sufficient for identification of the ATT.

Before moving into our main arguments, we provide three examples to illustrate the types of questions that we address in the current paper. We revisit these applications at relevant parts of the paper.

Example 1 (Stand-your-ground laws)**.** Cheng and Hoekstra, 2013 study the effects of stand-your-ground laws on homicides and other crimes. They use state-level data and exploit variation in the timing of stand-your-ground laws across states in order to identify policy effects. For some of their results, they condition on time-varying covariates that include state-level demographics, the number of police officers in the state, the number of people incarcerated, median income, poverty rate, and spending on assistance and public welfare. Although it is debatable whether or not some of the these covariates could be affected by the treatment (particularly the number of police officers and the number of people incarcerated), by running TWFE regressions that include these covariates, Cheng and Hoekstra, 2013 at least implicitly argue that these covariates evolve exogenously from the treatment. Whether this is true or not, for exposition purposes we will assume that none of the covariates used in this example are affected by the treatment.

[3] Using machine learning in this context may be particularly useful because the expressions for the ATT involve conditioning on time-varying covariates across different time periods. In many applications, time-varying covariates may be highly serially correlated, and it may be challenging to specify simple parametric models involving these covariates in this context. However, machine learning estimators may perform much better in this context.

Example 2 (Shelter-in-place orders)**.** A number of recent papers study the effect of shelter-in-place orders on various outcomes including mobility (see, for example, Weill et al., 2021 and references therein), labor market outcomes (e.g., Gupta et al., 2020), and consumer spending (e.g., Chetty et al., 2020). Paths of all of these outcomes (in the absence of shelter-in-place orders) likely depend on the current number of Covid-19 cases due to individuals making different choices about staying at home or continuing to work based on the local "state" of the pandemic. This suggests that parallel trends assumptions ought to condition on the number of Covid-19 cases that would have occurred if the policy had not been implemented. Moreover, since Covid-related policies are *designed* to affect the number of Covid-19 cases, this would be a case with a time-varying covariate that is likely to be affected by the treatment.

Example 3 (Job Displacement)**.** Research on job displacement typically invokes parallel trends assumptions to identify causal effects of job displacement on workers' earnings. If, in the absence of job displacement, paths of earnings depend on the occupation, industry, or union status of a worker, then it would be desirable to condition on these variables in the parallel trends assumption. However, most empirical work on job displacement does not condition on these variables, presumably due to each of these possibly being affected by job displacement.[4] Moreover, Barnette et al., 2021 argue that differences in the distribution of pre-displacement occupations are likely an important explanation for the magnitude of effects of job displacement; similarly, Brand, 2006 reports relatively large effects of job displacement on occupation.

The examples above are broadly representative of applications that invoke DID identification assumptions with time varying covariates. The first example involves time-varying covariates that can reasonably be thought of as evolving exogenously with respect to the treatment. The following two examples both involve covariates that are potentially affected by the treatment. Later in the paper, we point out some further conceptual differences between these latter two examples.

Related Literature

Our paper shares a similar motivation to Zeldow and Hatfield, 2021 which considers different possible sources of bias due to controlling for time-varying covariates that are possibly affected by the treatment. That paper mainly considers how sensitive existing strategies are (e.g., controlling for only pre-treatment covariates or additionally including lagged outcomes) to covariates that can be affected by the treatment. Relative to that paper, we make explicit assumptions on how the treatment can affect the covariates and, under these extra conditions, are able to propose estimation strategies that are guaranteed to perform well (up to regularity conditions) in those cases.

Our paper is also related to the literature on causal inference with panel data using structural nested mean models Robins, 1997 and marginal structural models Robins et al., 2000; see Blackwell and Glynn, 2018 for a recent review. These approaches, however, are based on "sequential ignorability" assumptions rather than allowing for time-invariant unobserved heterogeneity. Sequential ignorability implies

[4]Some papers do include occupation, industry, and/or union controls in "robustness checks" and others study how effects of job displacement vary by whether or not a worker remains in the same industry, occupation, or union status following job displacement which is broadly similar to controlling for each of these (see, for example, Topel, 1991, Jacobson et al., 1993, Stevens, 1997).

that treated and untreated potential outcomes are independent of treatment status conditional on pre-treatment values of covariates (and possibly pre-treatment outcomes).[5] Unlike the bulk of this literature, the current paper focuses on the case where a researcher would like to invoke a parallel trends assumption – rather than sequential ignorability – for identification. However, the current paper also invokes an additional assumption on how treated and untreated potential covariates are generated; this type of assumption is not made in this literature. The reason for this is that the timing that we consider differs from what is typically considered in the literature on sequential ignorability; in our case, units potentially become treated, and then their covariate realizes (and may itself be affected by treatment). This covariate needs to be controlled for identification. By contrast, the sequential ignorability literature typically has the covariate realized first, then the treatment, then the outcome, and controlling for, effectively, the covariate in the previous period is sufficient for identifying parameters of interest. That said, like the current paper, that literature does take seriously how covariates evolve over time and how participating in the treatment can affect covariates themselves. Of papers broadly in this literature, the most similar to the current paper is Imai et al., 2018 which focuses on a conditional parallel trends assumption that can hold after conditioning on past values of the covariates as well as past values of the outcome.

Our paper is also related to the literature on mediation analysis. Like a mediator, our covariates can be affected by treatment participation. However, the mediation literature is typically interested in decomposing treatment effects into direct effects of the treatment and indirect effects due to the effect of the treatment on the mediator (see Huber, 2020 for a recent review of this literature). Our paper is less ambitious in that we only seek to identify the overall effect of the treatment on outcomes; the tradeoff is that we are able to generally make weaker assumptions than would be required to separately recover direct and indirect effects of participating in the treatment. That said, it would be interesting to extend our arguments to additionally identifying direct and indirect effects of participating in the treatment, and it seems likely that existing arguments from the mediation analysis literature could be applied in this case. Our paper is relatively more similar to Rosenbaum, 1984, Lechner, 2008, Flores and Flores-Lagunes, 2009; these papers consider identification of treatment effect parameters under unconfoundedness (and with cross-sectional data) where the covariates that are required for the unconfoundedness assumption to hold could have been affected by the treatment. Besides this, our paper is related to a large literature in econometrics on strict exogeneity and pre-determinedness in panel data models (see, for example, Arellano and Honoré, 2001).

Finally, our results on interpreting TWFE regressions build on work on interpreting cross-sectional regressions under the assumption of unconfoundedness and in the presence of treatment effect heterogeneity; this literature includes Angrist, 1998, Aronow and Samii, 2016, Słoczyński, 2020, Goldsmith-Pinkham et al., 2021, Ishimaru, 2021. Goodman-Bacon, 2021, de Chaisemartin and D'Haultfœuille, 2020, and Ishimaru, 2022, all of which provide decompositions of the TWFE regression in Equation (1.1). In some ways, the decompositions in these papers are more general than our decomposition as they all consider the case

[5] Another difference between the current paper and much of the sequential ignorability literature is that these papers are typically primarily interested in recovering causal effects of different treatment paths (e.g., where each unit can move into or out of the treatment in each period). The arguments in our paper could likely be extended in this direction but our main results apply to the case where there are only two time periods and treatment can only take place in the second time period.

with more than two time periods and with variation in treatment timing. On the other hand, our results zoom in on the "textbook" case with exactly two periods and where no one is treated in the first period; our decomposition emphasizes a number of possible limitations of TWFE regressions even in the case with exactly two periods. Indeed, moving to more complicated cases with more periods and variation in treatment timing would make the case for using TWFE regressions even weaker, as it would introduce additional issues particularly related to using already treated units as comparison units (which can lead to negative weights on underlying treatment effect parameters), as all three papers mentioned above imply. See Remark 4 below for a more detailed comparison.

Chapter 2

Identification

Notation and Setup

For this section, we focus on a baseline case where the researcher has access to two time periods of panel data. We label the second time period t^* and the first time period t^*-1, and use t to indicate a generic time period. In each time period, we observe outcomes Y_t, time-varying covariates X_t, and time-invariant covariates Z. As is standard in the DID literature, we suppose that no one is treated in the first time period. We use the binary variable D to indicate whether or not a unit participates in the treatment. Importantly for our setup, we allow for the possibility that the time-varying covariate can itself be affected by the treatment; in order to do this, we define $X_t(1)$ to be the value that the covariates would take if a unit participated in the treatment and $X_t(0)$ to be the value that the covariates would take if a unit did not participate in the treatment; for simplicity, we often refer to these as "treated potential covariates" and "untreated potential covariates." Next, we define treated potential outcomes as $Y_t(1, X_t(1))$ (this is the outcome that a unit would experience in time period t if they participated in the treatment and their covariates took on its value under the treatment) and untreated potential outcomes as $Y_t(0, X_t(0))$ (this is the outcome that a unit would experience in time period t if they did not participate in the treatment and their covariates took their values in the absence of the treatment). For most of the arguments in the current paper, it is sufficient to use the shorter notation $Y_t(1) := Y_t(1, X_t(1))$ and $Y_t(0) := Y_t(0, X_t(0))$. In this setup, the observed covariates in each time period are: $X_{t^*} = DX_{t^*}(1) + (1-D)X_{t^*}(0)$ and $X_{t^*-1} = X_{t^*-1}(0)$. In other words, in the second time period, we observe treated potential covariates for units that participate in the treatment, and we observe untreated potential covariates for units that do not participate in the treatment. In the first time period, since no units are treated yet, we observe untreated potential covariates for all units. Likewise, observed outcomes are given by $Y_{t^*} = DY_{t^*}(1) + (1-D)Y_{t^*}(0)$, and $Y_{t^*-1} = Y_{t^*-1}(0)$.

Following the vast majority of the DID literature, we target identifying the average treatment effect on the treated (ATT). It is given by

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

which is the average difference between treated and untreated potential outcomes among the treated group.

Throughout the paper, we make the following assumptions

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

Assumption 2 (Conditional Parallel Trends)**.**

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

Assumption 3 (Overlap)**.**

(a) $P(D = 1|X_{t^*}, X_{t^*-1}, Z) < 1$ *a.s.*

(b) $P(D = 1|X_{t^*-1}, W_{t^*-1}, Z) < 1$ *a.s.*

Assumption 1 says that we observe iid panel data. Assumption 2 says that, on average, the path of untreated potential outcomes is the same for the treated group as for the untreated group after conditioning on untreated potential covariates in time period t^*, pre-treatment covariates X_{t^*-1}, and time-invariant covariates Z. Relative to standard conditional parallel trends assumptions (Heckman et al., 1997, Callaway and Sant'Anna, 2021, Abadie, 2005), the set of covariates being conditioned on includes untreated potential covariates which are unobserved for the treated group and therefore can complicate existing identification strategies. Assumption 3 is an overlap assumption, and this type of assumption is standard in the treatment effects literature. Part (a) implies that, for any values of X_{t^*}, X_{t^*-1}, and Z, there will be some untreated units with those values of the covariates in the population. Part (b) is similar but holds for any values of X_{t^*-1}, W_{t^*-1}, and Z.

Next, we provide two distinct assumptions for dealing with covariates that vary over time.

Assumption Cov-Exogeneity. $(X_{t^*}(0)|X_{t^*-1}, Z, D = 1) \sim (X_{t^*}(1)|X_{t^*-1}, Z, D = 1)$

Assumption Cov-Unconfoundedness. $X_{t^*}(0) \perp\!\!\!\perp D|X_{t^*-1}, W_{t^*-1}, Z$ *where* W_{t^*-1} *is a vector of pre-treatment variables.*

We call the first assumption covariate exogeneity because it implies that participating in the treatment does not change the distribution of covariates for the treated group. This assumption is technically weaker than assumptions like, for all units $X_{it^*}(1) = X_{it^*}(0) = X_{it^*}$, although this would certainly be a leading case where this sort of condition might hold. Assumption Cov-Exogeneity allows for covariates to change values over time, but it imposes that (in distribution) they are not affected by participating in the treatment. This sort of condition may be reasonable in some applications (e.g., Example 1 above). In other cases, this assumption may be less reasonable (e.g., Examples 2 and 3 above).

Assumption Cov-Unconfoundedness is an unconfoundedness assumption for untreated potential covariates. It allows for the treatment to affect the time varying covariates, but it says that the distribution of untreated potential covariates is the same for the treated group and the untreated group after conditioning on the vector of pre-treatment covariates $(X_{t^*-1}, W_{t^*-1}, Z)$. This assumption allows us to recover the conditional distribution of untreated potential covariates for the treated group. This distribution is a key ingredient for identifying the ATT below.

In Assumption Cov-Unconfoundedness, we allow for the possibility that W_{t^*-1} is empty; in fact, this is a leading case. In this case, unconfoundedness for untreated potential covariates holds after conditioning on the lag of the time-varying covariates X_{t^*-1} and other time invariant covariates Z. Below, we connect this specific condition to the common practice in the econometrics literature on DID of conditioning on pre-treatment values of time-varying covariates. With a slight abuse of notation, we also allow for the possibility that W_{t^*-1} includes the lagged outcome Y_{t^*-1}, so that covariate unconfoundedness holds after conditioning on pre-treatment covariates, time invariant covariates, and the pre-treatment outcome. Interestingly, we show below that, under this condition, both the path of outcomes over time and the lag of the outcome show up in the expression for ATT which is unusual in DID applications (see, Chabé-Ferret, 2017 for related discussion). In the results below, we provide separate results that invoke either Assumption Cov-Exogeneity or Assumption Cov-Unconfoundedness.

Next, we state our main identification result.

Theorem 1. *Under Assumptions 1 and 2,*

(1) if, in addition, Assumption Cov-Exogeneity and Assumption 3(a) hold, then

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

(2) if, in addition, Assumption Cov-Unconfoundedness and Assumption 3(b) hold, then

$$\begin{aligned} ATT &= \mathbb{E}[\Delta Y_{t^*}|D=1] \\ &\quad - \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\big|X_{t^*-1}, W_{t^*-1}, Z, D=0\big]\Big|D=1\Big]. \end{aligned}$$

The intuition for part (1) of Theorem 1 is relatively straightforward. Under the conditional parallel trends assumption and when covariates evolve exogenously, one can recover the ATT by (i) taking the path of outcomes experienced by the treated group and adjusting it by the path of outcomes experienced by the untreated group (conditional on X_{t^*}, X_{t^*-1}, and Z) and then (ii) accounting for differences in the distribution of X_{t^*}, X_{t^*-1}, and Z across groups. This result is very similar to existing results with time invariant covariates (e.g., Heckman et al., 1997 as well as Lechner, 2011).

The intuition for part (2) is somewhat more complicated. The term $\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]$ is the average change in outcomes over time conditional on X_{t^*}, X_{t^*-1}, and Z among the untreated group. Under Assumption 2, this is the path of outcomes that, conditional on $X_{t^*}(0)$, X_{t^*-1}, and Z, the treated group would have experienced if they had not participated in the treatment. The next expectation is over the distribution of $X_{t^*}(0)$ (conditional on X_{t^*-1}, W_{t^*-1}, and Z) for the untreated group. Under Assumption Cov-Unconfoundedness, this is the same conditional distribution that $X_{t^*}(0)$ follows for the treated group. Finally, the outside expectation is over the distribution of X_{t^*-1}, W_{t^*-1}, and Z for the treated group and, therefore, allows for these variables to be distributed differently in the treated group relative to the untreated group.

Corollary 1 (Important Special Cases)**.** *Under Assumptions 1, 2, 3(b), and Cov-Unconfoundedness,*

(1) if, in addition, $W_{t^*-1} = \varnothing$, *then*

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

(2) if, in addition, $W_{t^*-1} = Y_{t^*-1}$, *then*

$$\begin{aligned} ATT &= \mathbb{E}[\Delta Y_{t^*}|D=1] \\ &\quad - \mathbb{E}\Big[\mathbb{E}\big[\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0]\big|X_{t^*-1}, Y_{t^*-1}, Z, D=0\big]\Big|D=1\Big]. \end{aligned}$$

Corollary 1 provides two important special cases for the results in part (2) of Theorem 1. The first part provides a formal justification for the common practice in the econometrics literature on DID with time varying covariates of including only "pre-treatment" covariates. In particular, this result says that, when unconfoundedness holds for the time varying covariate conditional on time-invariant covariates and other pre-treatment covariates, then it is sufficient for the researcher to only "account for" pre-treatment and time-invariant covariates in order to recover the ATT.[1] The second part of Corollary 2 is also interesting in that it relates the ATT to an expression that includes the lagged outcome. There are a number of papers that explore the idea of including lagged outcomes in a DID framework (e.g., Chabé-Ferret, 2017, Imai et al., 2018, Zeldow and Hatfield, 2021) though it is challenging to provide a justification for including lagged outcomes in DID settings — our approach justifies the inclusion of lagged outcomes (in the manner specified in the corollary) in cases where unconfoundedness for the time-varying covariate holds after conditioning on the lag of the outcome variable.

Next, we provide alternative expressions for ATT that are useful for estimation.

Corollary 2 (Doubly Robust Expressions for ATT)**.** *Under Assumptions 1 and 2,*

(1) if, in addition, Assumption Cov-Exogeneity and Assumption 3(a) hold, then

$$ATT = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{p(X_{t^*}, X_{t^*-1}, Z)(1-D)}{\mathbb{E}[D](1-p(X_{t^*}, X_{t^*-1}, Z))}\right)(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=0])\right], \tag{2.1}$$

where $p(X_{t^*}, X_{t^*-1}, Z) := P(D=1|X_{t^*}, X_{t^*-1}, Z)$.

[1] We provide the proof of this result in the appendix. The proof is relatively straightforward, but it appears to be a new contribution in the literature.

(2) *if, in addition, Assumption Cov-Unconfoundedness and Assumption 3(b) hold with* $W_{t^*-1} = \varnothing$, *then*

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

where $p(X_{t^*-1}, Z) := P(D=1|X_{t^*-1}, Z)$.

Both of the expressions in Corollary 2 involve both an outcome regression (the conditional expectation terms in each expression) and a propensity score. They are both doubly robust in the sense that a researcher can consistently estimate the ATT if *either* the model for the propensity score or the outcome regression model is correctly specified (references on double robustness include Robins et al., 1994, Scharfstein et al., 1999, Słoczyński and Wooldridge, 2018, Sant'Anna and Zhao, 2020). Besides this, they also provide a connection to the DID literature on estimating the ATT under conditional parallel trends using double/debiased machine learning; see, in particular, Chang, 2020. This may be particularly useful in the first case, where the propensity score and outcome regression depend on time-varying covariates in both periods. These can be practically difficult to estimate because, in many cases, X_{t^*} and X_{t^*-1} may be highly collinear. Conventional methods typically invoke functional form assumptions that impose, for example, that these functionals only depend on ΔX_{t^*}. As noted below, these sorts of restrictions may be implausible in many applications.

Remark 1. In cases where time-varying covariates may be affected by the treatment, we mainly focus on the case where an unconfoundedness type assumption holds for the time varying covariates. A natural alternative would be to invoke parallel trends assumptions for the time-varying covariates themselves. Importantly, though, our above arguments require identifying the entire conditional distribution of $X_{t^*}(0)$ for the treated group (not just its mean).[2] That said, difference in differences approaches that recover the distribution of untreated potential outcomes, such as Callaway and Li, 2019, Callaway et al., 2018, could be applied here (though note that these approaches require additional assumptions). Likewise, the change-in-changes approach in Athey and Imbens, 2006, Melly and Santangelo, 2015, which can recover distributions of untreated potential outcomes, could be applied to the time-varying covariates in this context. Another potential limitation of these approaches in this context is that they typically only point identify distributions of continuous outcomes and, therefore, would not be very suitable for a number of relevant applications that involve discrete time-varying covariates.

Remark 2. Although neither of our assumptions on untreated potential covariates in Assumptions Cov-Exogeneity and Cov-Unconfoundedness are directly testable, the condition that is given in Assumption Cov-Unconfoundedness can be "pre-tested" — that is, one can check if it holds in pre-treatment time periods. One simple idea is to compute pseudo-ATTs in pre-treatment periods; if both Assumption 2 (the conditional parallel trends assumption) and Assumption Cov-Unconfoundedness hold in pre-treatment

[2] In Chapter 4, we propose some alternative approaches where standard parallel trends assumptions for time-varying covariates can be used, though these approaches require imposing a linear model for the path of untreated potential outcomes in Assumption 2 that are not used in this section.

periods, then these pseudo-ATTs should be equal to 0. Alternatively, one can directly pre-test Assumption Cov-Unconfoundedness: for some pre-treatment period t, Assumption Cov-Unconfoundedness implies that the distribution of $X_t|X_{t-1}, W_{t-1}, Z, D = d$ is the same for both the treated and untreated groups. This sort of test could be implemented using results from the goodness-of-fit testing literature (e.g., Bierens, 1982, Stute, 1997).

To conclude this section, we revisit the three examples from the introduction.

Example 1 (Stand-your-ground, cont'd) Our example on stand-your-ground laws involved conditioning on time-varying covariates that evolved exogenously with respect to the treatment. This suggests that this example is most related to our results in part (1) of Theorem 1 and part (1) of Corollary 2. In particular, machine learning estimators of the propensity score and outcome regression functions in Corollary 2 are particularly attractive as they do not require strong functional form assumptions on these nuisance functions.[3]

Example 2 (Shelter-in-place, cont'd) In our example of shelter-in-place orders on various economic outcomes, the parallel trends assumption holds after conditioning on the number of Covid-19 cases that would have occurred if the policy had not been implemented. That is, "untreated potential Covid-19 cases" plays the role of $X_{t^*}(0)$ in this case. Callaway and Li, 2021 show that, under a SIRD model — which is the leading pandemic model in the epidemiology literature — controlling for the pre-treatment "state" of the pandemic is sufficient for unconfoundedness to hold. That is, the conditions in part (1) of Corollary 1 and part (2) of Corollary 2 hold when one wants to control for the number of untreated potential Covid-19 cases.

Example 3 (Job displacement, cont'd) Finally, recall our example on the effect of job displacement on earnings where the parallel trends assumption holds only after conditioning on, for example, "untreated potential occupation" — that is, the occupation that a worker would have had if they had not been displaced from their job. In this case, an unconfoundedness assumption for occupation may be more likely to hold if it conditions on (i) pre-treatment time-varying covariates (including pre-treatment occupation), (ii) time invariant covariates (such as demographics and education), *and* (iii) pre-treatment earnings. In particular, conditioning on pre-treatment earnings could be important if there are occupation specific wage premiums and high-earning workers are more likely to (in the absence of job displacement) stay in the same occupation over time relative to low-earning workers. This application would then be covered by the results from part (2) of Corollary 1.

[3] This particular application uses state-level data, so, in practice, it may be difficult to use machine learning approaches with only 50 or so observations. See Chapter 4 for some more parametric approaches that may be more suitable for applications with limited data. That said, the more general point here though is that, in cases where covariates evolve exogenously, machine learning estimators, given enough data, are likely to be attractive in many applications.

Chapter 3

Interpreting TWFE Regressions

In this section, we consider how to interpret α in the TWFE regression in Equation (1.1). We continue to consider the "textbook" case 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 best-case for TWFE regressions as it does not introduce well-known problems related to using already-treated units as comparison units that show up when using TWFE regressions with multiple periods, variation in treatment timing, and treatment effect heterogeneity (Goodman-Bacon, 2021, de Chaisemartin and D'Haultfœuille, 2020). In the case with exactly two periods, it is helpful to equivalently re-write Equation (1.1) as

$$\Delta Y_{it^*} = \alpha D_i + \Delta X'_{it^*}\beta + \Delta v_{it^*} \tag{3.1}$$

where we define $\Delta X_{t^*} := (1, X_{t^*} - X_{t^*-1})'$ which is the change in the covariate over time and is augmented with an intercept term for the time fixed effect. We also slightly abuse notation by taking β to include an extra parameter in its first position corresponding to the intercept. 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 underlying treatment effect parameters.

Denote the linear projection of ΔY_{t^*} on ΔX_{t^*} by
$\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}) := \Delta X'_{t^*}\mathbb{E}[\Delta X_{t^*}\Delta X'_{t^*}]^{-1}\mathbb{E}[\Delta X_{t^*}\Delta Y_{t^*}]$,

and define the corresponding projection error $e := \Delta Y_{t^*} - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*})$. Similarly, define the linear projection of D on ΔX_{t^*} as $\mathrm{L}(D|\Delta X_{t^*}) := \Delta X'_{t^*}\mathbb{E}[\Delta X_{t^*}\Delta X'_{t^*}]^{-1}\mathbb{E}[\Delta X_{t^*}D]$ and the corresponding projection error $u := D - \mathrm{L}(D|\Delta X_{t^*})$. Standard Frisch-Waugh type arguments imply that

$$\alpha = \frac{\mathbb{E}[De]}{\mathbb{E}[u^2]} \tag{3.2}$$

Below, to keep the notation concise, it is useful to define $X^{all}(d) := (X_{t^*}(d), X_{t^*-1}, Z)$. We also define $ATT_{X^{all}(0)}(X^{all}(0)) := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0)|X^{all}(0), D = 1]$ which is the ATT conditional on $X_{t^*}(0)$, X_{t^*-1}, and Z. And we further define $p(X^{all}(0)) = \mathrm{P}(D = 1|X^{all}(0))$. Next, we state a main result decomposing α from the TWFE regression.

Proposition 1. *Under Assumptions 1, 2, and 3(a),*

$$\alpha = \mathbb{E}[\omega_{ATT}(X^{all}(0))ATT_{X^{all}(0)}(X^{all}(0))|D=1]$$

$$+ \mathbb{E}\Bigg[\sum_{d\in\{0,1\}} \omega_d(X^{all}(0))\Bigg\{\Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}(0), X_{t^*-1}, Z, D=d] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=d]\Big) \quad \text{(A)}$$

$$+ \Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, Z, D=d] - \mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=d]\Big) \quad \text{(B)}$$

$$+ \Big(\mathbb{E}[\Delta Y_{t^*}|X_{t^*}, X_{t^*-1}, D=d] - \mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=d]\Big) \quad \text{(C)}$$

$$+ \Big(\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D=d] - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=d)\Big)\Bigg\}\Bigg|D=1\Bigg] \quad \text{(D)}$$

$$+ \mathbb{E}\left[\omega_e(X^{all}(0))\left\{L(\Delta Y_{t^*}|\Delta X_{t^*}, D=1) - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\right\}\Big|D=1\right] \quad \text{(E)}$$

where

$$\omega_{ATT}(X^{all}(0)) := \frac{1-p(X^{all}(0))}{\mathbb{E}[(1-L(D|\Delta X_{t^*}))|D=1]}$$

$$\omega_d(X^{all}(0)) := \frac{d\,p(X^{all}(0)) + (1-d)(1-p(X^{all}(0)))}{\mathbb{E}[(1-L(D|\Delta X_{t^*}))|D=1]}$$

$$\omega_e(X^{all}(0)) := \frac{(p(X^{all}(0)) - L(D|\Delta X_{t^*}))}{\mathbb{E}[(1-L(D|\Delta X_{t^*}))|D=1]}$$

The result in Proposition 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 a number of undesirable "bias" terms. We provide formal conditions under which each of these extra terms are equal to zero below. But, informally, term (A) contains bias from the treatment potentially affecting the covariates in time period t^*. Term (B) comes from ignoring time invariant covariates. Term (C) comes up when paths of outcomes depend on the levels of time-varying covariates instead of only on the change in covariates over time. Term (D) arises when the conditional expectation is nonlinear in the change in covariates over time. Term (E) is conceptually different from terms (A)-(D) and is non-zero when the propensity score is not equal to a linear projection of the treatment on the change in covariates over time.[1]

Next, we introduce several additional assumptions that are useful for eliminating the bias terms in Proposition 1. We also use the additional notation:

$$ATT_{X_{t^*}(0),X_{t^*-1}}(x_{t^*}(0), x_{t^*-1}) := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0)|X_{t^*}(0) = x_{t^*}(0), X_{t^*-1} = x_{t^*-1}]$$

[1] Without further assumptions, some of the expressions that involve $X_{t^*}(0)$ are not necessarily identified (this includes $ATT_{X^{all}(0)}(X^{all}(0)), \mathbb{E}[\Delta Y_{t^*}|X_{t^*}(0), X_{t^*-1}, Z, D=1]$ and all of the weights as they depend on $p(X^{all}(0))$. However, if we additionally invoke Assumption Cov-Exogeneity, then all of these terms are identified and Term (A) is equal to zero (see the discussion below for more details along these lines). The reason we do not invoke this assumption in Proposition 1 is to point out that time-varying covariates being affected by the treatment can itself be an additional complication for TWFE regressions.

and

$ATT_{\Delta X_{t^*}(0)}(\Delta x_{t^*}(0)) := \mathbb{E}[Y_{t^*}(1) - Y_{t^*}(0)|\Delta X_{t^*}(0) = \Delta x_{t^*}(0)]$ — these define different types of conditional ATTs.

Assumption 4 (Conditional ATTs and parallel trends do not depend on time invariant covariates)**.**

(a) $ATT_{X^{all}(0)}(X^{all}(0)) = ATT_{X_{t^*}(0), X_{t^*-1}}(X_{t^*}(0), X_{t^*-1})$ *a.s.*

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

Assumption 5 (Conditional ATTs and parallel trends only depend on change in time-varying covariates)**.**

(a) $ATT_{X_{t^*}(0), X_{t^*-1}}(X_{t^*}(0), X_{t^*-1}) = ATT_{\Delta X_{t^*}(0)}(\Delta X_{t^*}(0))$ *a.s.*

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

Assumption 6 (Linearity of conditional ATTs and paths of untreated potential outcomes)**.**

(a) There exists a δ_1 *such that* $ATT_{\Delta X_{t^*}(0)}(\Delta X_{t^*}(0)) = \Delta X_{t^*}(0)'\delta_1$

(b) There exists a δ_0 *such that* $\mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}(0), D = 0] = \Delta X_{t^*}(0)'\delta_0$

Assumption 7 (Linearity of propensity score in terms of change in time-varying covariates)**.**
There exists a δ_p *such that* $p(X^{all}(0)) = \Delta X_{t^*}(0)'\delta_p$.

The first part of Assumption 4 says that, conditional on $X_{t^*}(0)$ and X_{t^*-1}, conditional ATTs do not depend on time invariant covariates Z. The second part says that, conditional on $X_{t^*}(0)$ and X_{t^*-1}, the path of untreated potential outcomes does not depend on time invariant covariates Z. This implies that the conditional parallel trends assumption holds without conditioning on time invariant covariates (and thus strengthens Assumption 2). Assumption 5 is similar; the first part says that conditional ATTs further only depend on changes in time-varying covariates over time, and the second part says that the conditional parallel trends assumption only depends on the change in time-varying covariates over time rather than their level.

Assumption 6 says that conditional ATTs and paths of untreated potential outcomes are linear in changes in untreated potential covariates over time. Jointly, Assumption Cov-Exogeneity and Assumptions 4 to 6 imply that (i) time varying covariates are not affected by the treatment, (ii) that conditional ATTs (conditional on $X_{t^*}(0)$, X_{t^*-1}, and Z) only depend on the change in time-varying covariates (and not on their levels or time invariant covariates) and are linear in time-varying covariates, and (iii) that the conditional parallel trends assumption in Assumption 2 only depends on the change in time-varying covariates over time (and not on their levels or time invariant covariates) and is linear in time-varying covariates over time.

Assumption 7 says that the propensity score (conditional on $X_{t^*}(0)$, X_{t^*-1}, and Z) is linear in $\Delta X_{t^*}(0)$. This type of assumption is very common in the literature on interpreting regressions under unconfoundedness with cross-sectional data (e.g., Aronow and Samii, 2016, Słoczyński, 2020, Angrist, 1998, Ishimaru, 2021). In those cases, it sometimes holds by construction (e.g., when the covariates are all discrete and a full set of interactions is included in the model). In our case, though, it seems particularly implausible as (i) it requires the propensity score to only depend on changes in covariates over time, and (ii) even with fully interacted discrete regressors, the propensity score is unlikely to be linear in changes in the regressors over time.[2]

Proposition 2. *Under Assumptions 1, 2, 3(a), Cov-Exogeneity, 4, 5, and 6,*

$$\begin{aligned}\alpha &= \mathbb{E}[\omega_{ATT}(X^{all}(0))ATT_{X^{all}(0)}(X^{all}(0))|D=1] \\ &\quad + \mathbb{E}\left[\omega_e(X^{all}(0))\left\{L(\Delta Y_{t^*}|\Delta X_{t^*}, D=1) - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\right\}\big|D=1\right] \qquad \text{(E)}\end{aligned}$$

where ω_{ATT} and ω_e are defined in Proposition 1.

(a) If, in addition, Assumption 7 holds, then

$$\alpha = \mathbb{E}[\omega_{ATT}(X^{all}(0))ATT_{X^{all}(0)}(X^{all}(0))|D=1]$$

and $\mathbb{E}[\omega_{ATT}(X^{all}(0))|D=1] = 1$.

(b) If, in addition, Assumption 7 holds and $ATT_{\Delta X_{t^*}(0)}(\Delta x_{t^*}(0))$ is the same across all values of $\Delta x_{t^*}(0)$, then

$$\alpha = ATT$$

The proof of Proposition 2 is provided in the appendix. In the proof, we provide more specific results on which conditions are required for each term in terms (A)-(D) in Proposition 1 to be equal to 0.

The result in Proposition 2 suggests a number of potential issues with the TWFE regressions as in Equation (1.1). First, even if one is willing to maintain the additional assumptions in Assumption Cov-Exogeneity and Assumptions 4 to 6 (which are likely to be very strong in most applications), α from the TWFE regression is still hard to interpret. Maintaining these additional assumptions (particularly, Assumption Cov-Exogeneity) implies that all of the weights, conditional ATTs, and linear projections in the first part of Proposition 2 are identified and directly estimable. That said, the weights on conditional ATTs, ω_{ATT}, do not have the property that they have mean one and the nuisance expression in term (E) may be non-negligible.

The second part of Proposition 2 says that, if we are willing to assume that the propensity score is equal to the linear projection of the treatment on the change in time-varying covariates over time, then

[2] For example, suppose that the only covariate is binary. In the cross-sectional case considered by other papers mentioned above, the propensity score would be linear by construction. However, the change in the covariate over time would be a single variable that can take the values -1, 0, or 1; moreover, the change in a binary covariate over time is equal to 0 in cases when the covariate is equal to 1 in both periods or when the covariate is equal to 0 in both periods. This suggests that the propensity score would not be linear in the change in covariates over time even in this very simple case.

the weights on conditional ATTs will have mean one and the nuisance expression in term (E) will be equal to zero. Even in this case, the weights have a "weight-reversal" property analogous to the one pointed out in Słoczyński, 2020 in the context of unconfoundedness and cross-sectional data. What this means is that conditional ATTs are given more weight for values of the covariates that are uncommon among the treated group relative to the untreated group; and that conditional ATTs are given less weight for values of the covariates that are 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 still be systematically different for treated units relative to untreated units; however, for the treated group, treatment effects cannot be systematically different across different values of the covariates.

These results are much different from our earlier results in Chapter 2. Those results did not require any of the additional assumptions in Proposition 2. In fact, when covariates evolve exogenously with respect to the treatment (as under Assumption Cov-Exogeneity), then the doubly robust expressions for the ATT in part (1) of Corollary 2 only require that either the propensity score or the outcome regression model be correctly specified; in cases where these are estimated using machine learning, even these parametric assumptions can be substantially relaxed. Moreover, in contrast with the TWFE regressions considered in this section, our earlier additional results can accommodate cases where the time-varying covariates are affected by the treatment.

Remark 3. It is worth pointing out that all of the extra conditions considered in Proposition 2 are sufficient conditions rather than necessary conditions. For example, it would be possible for some violations of these assumptions to "offset" each other so that α *happens* to be equal to ATT. That said, there is no reason to expect this to happen in applications.

Remark 4. The result in Proposition 1 is related to several other decompositions of TWFE regressions that include covariates. All of these papers consider the same TWFE regression that we do in Equation (1.1). They also consider the more general case where there are more than two time periods and allow for variation in treatment timing. de Chaisemartin and D'Haultfœuille, 2020 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. That said, this suggests that in the two period case that we consider, TWFE regressions would recover the ATT. The explanation for this difference is that we do not impose those extra conditions in Proposition 1. Goodman-Bacon, 2021 provides a decomposition of α into a "within" component and "between" component. The between component arises due to variation in treatment timing, and, therefore does not show up in our case. The within component is analogous to our expression for α in Equation (3.2). Relative to Goodman-Bacon, 2021, we further decompose this term into a number of 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. Finally, Ishimaru, 2022, 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 (though, in both papers, the weights can be negative). These include a version of conditional parallel trends that holds when one conditions on the change in covariates over time[3] and an assumption on linearity of the propensity score conditional on changes in observed covariates over time.[4] None of these papers explicitly address the issue of time-varying covariates potentially being affected by the treatment.[5]

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

[4]Another way that the decomposition in Ishimaru, 2022 is more general than the one in the current paper is that paper 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 regression from the TWFE regression in Equation (1.1) that is much more commonly used in empirical work in economics.

[5]Goodman-Bacon, 2021 does remark that "Note that for covariates to aid in identification, [time-varying covariates] must be unaffected by the treatment to avoid bias from 'conditioning on a post-treatment variable'."

Chapter 4

Alternative Regression Adjustment/Imputation Approaches

In this section, we provide several alternative strategies that involve stronger parametric assumptions on the path of untreated potential outcomes than we made in Chapter 2. The approaches discussed in this section are generally simpler to estimate than would be the case for the expressions coming from Chapter 2 and, in some cases, can allow for weaker (or at least alternative) assumptions on how the treatment affects time-varying covariates. The strategies that we propose in this section are also able to avoid the issues with TWFE regressions pointed out in Chapter 3, and (when desired) can allow for the possibility that the treatment has an effect on the covariates.

The ideas in this section are broadly similar to regression adjustment strategies in the treatment effects literature (see, for example, Imbens and Wooldridge, 2009) and the imputation estimators proposed in Liu et al., 2021, Gardner, 2021, Borusyak et al., 2021 though they allow for (i) time-varying effects of time varying covariates and (ii) the possibility that the treatment directly affects the time-varying covariates.

To start with, it is well known (e.g. Blundell and Costa Dias, 2009, Gardner, 2021, Borusyak et al., 2021) that there is a close connection between unconditional parallel trends assumptions and the following model for untreated potential outcomes

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

where θ_t is a time fixed effect, η_i is time invariant unobserved heterogeneity (i.e., an individual fixed effect), and v_{it} are idiosyncratic, time varying unobservables. An unconditional version of parallel trends holds in this model for untreated potential outcomes under the condition that $\mathbb{E}[\Delta v_t|D=1] = \mathbb{E}[\Delta v_t|D=0]$ for all time periods (this would hold by construction if v_t is independent of treatment status in all time periods), but allows for η to be distributed differently across groups and does not impose any modeling assumptions on treated potential outcomes.

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

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

where the distribution of η can vary across groups (as well as vary with Z) and the key condition for the conditional parallel trends assumption to hold is that $\mathbb{E}[\Delta v_t|Z, D = 1] = \mathbb{E}[\Delta v_t|Z, D = 0]$ (see, for example, Heckman et al., 1997 for a discussion of this kind of model).[1]

In this setup, the main challenge is estimating $g_t(z)$ (though note that this is a practical, estimation challenge rather than an identification challenge). The natural way to parameterize this model is

$$Y_{it}(0) = Z_i'\delta_t + \eta_i + v_{it} \tag{4.1}$$

where we now take Z to include an intercept (and, therefore, δ_t absorbs the time fixed effect). Given this framework, $ATT = \mathbb{E}[\Delta Y_{t^*}|D = 1] - \mathbb{E}[Z|D = 1]'\delta^*_{t^*}$ where $\delta^*_{t^*} := (\delta_{t^*} - \delta_{t^*-1})$ which can be consistently estimated from the regression of ΔY_{t^*} on Z using only observations from the untreated group.[2]

The same sort of arguments imply that, when there are some covariates that vary over time (as above, we consider the case of a single time-varying covariate but note that it is straightforward to extend these arguments to cases with more time-varying covariates), a natural motivating model is

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

which implies that

$$\Delta Y_{it^*}(0) = g_{t^*}(Z_i, X_{it^*}(0)) - g_{t^*-1}(Z_i, X_{it^*-1}) + \Delta v_{it^*}$$

Moreover, the same sorts of arguments as above imply that Assumption 2 holds in this model. Similar to the previous case, the main practical challenge is that $g_t(z, x_t(0))$ is likely to be challenging to estimate. Like the previous case, the natural way to parameterize this model is

$$Y_{it}(0) = Z_i'\delta_t + X_{it}(0)\beta_t + \eta_i + v_{it}$$

[1] 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.

[2] This is closely related to regression adjustment estimators (see, for example, Heckman et al., 1998, Imbens and Wooldridge, 2009, Sant'Anna and Zhao, 2020 for related discussion). An alternative strategy would be to estimate the ATT by $n_1^{-1}\sum_{i=1,D_i=1}^{n}(Y_{it^*} - \hat{Y}_{it^*}(0))$ where n_1 is the number of treated observations and $\hat{Y}_{it^*}(0)$ is an imputed untreated potential outcome given by $\hat{Y}_{it^*}(0) = Y_{it^*-1} + Z_i'\hat{\delta}^*_{t^*}$. This imputation estimator is numerically equal to the regression adjustment estimator, but the imputation formulation is convenient particularly in the case with multiple periods and variation in treatment timing; see Remark 6 below for more details.

which implies that

$$\Delta Y_{it^*}(0) = Z_i'\delta^*_{t^*} + \Delta X_{it^*}(0)\beta_{t^*} + X_{it^*-1}(0)\beta^*_{t^*} + \Delta v_{it^*} \tag{4.2}$$

where $\beta^*_{t^*} := (\beta_{t^*} - \beta_{t^*-1})$. Notice that, because untreated potential outcomes and untreated potential covariates 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. The model in Equation (4.2) is conceptually appealing as (up to the parametric assumptions) it compares units with both the same initial level of the time-varying covariate and that have the same change in time-varying covariates over time.

Although it is straightforward to recover the parameters in Equation (4.2), recall that,

$$\begin{aligned} ATT &= \mathbb{E}[\Delta Y_{t^*}|D=1] - \mathbb{E}[\Delta Y_{t^*}(0)|D=1] \\ &= \mathbb{E}[\Delta Y_{t^*}|D=1] - \Big(\mathbb{E}[Z|D=1]'\delta^*_{t^*} + \mathbb{E}[\Delta X_{t^*}(0)|D=1]\beta_{t^*} + \mathbb{E}[X_{t^*-1}(0)|D=1]\beta^*_{t^*}\Big) \end{aligned}$$

Given that the parameters are identified, every term is identified in this expression except $\mathbb{E}[\Delta X_{t^*}(0)|D = 1]$ (because $X_{t^*}(0)$ is not observed for the treated group). We briefly consider six settings for recovering $\mathbb{E}[\Delta X_{t^*}(0)|D = 1]$ — three of these come from the assumptions we have already considered for untreated potential covariates and three involve parallel trends assumptions for untreated potential covariates. Several of these cases involve averaging over conditional expectations of $\Delta X_t(0)$. In this section we additionally impose linear models for these conditional expectations; under this extra condition, researchers are able to estimate ATT while potentially allowing for the treatment to affect time-varying covariates using only regressions and averaging.

Case 1: Assumption Cov-Exogeneity holds

Under Assumption Cov-Exogeneity, $\mathbb{E}[\Delta X_{t^*}(0)|D = 1] = \mathbb{E}[\Delta X_{t^*}(1)|D = 1] = \mathbb{E}[\Delta X_{t^*}|D = 1]$. That is, when covariates evolve exogenously with respect to the treatment, we can replace the average change in untreated potential covariates for the treated group with the average change in covariates actually experienced by the treated group.

Case 2: Assumption Cov-Unconfoundedness holds conditional on (Z, X_{t^*-1})

In this case, if we are willing to assume the following linear model for the change in untreated potential covariates

$$\Delta X_{it^*}(0) = Z_i'\gamma_{t^*} + X_{it^*-1}\lambda_{t^*} + u_{it^*}$$

then the Cov-Unconfoundedness assumption gives the following result:

$$\mathbb{E}[u_{t^*}|Z, X_{t^*-1}, D = d] = 0$$

for $d \in \{0, 1\}$.

Plugging this expression into Equation (4.2) implies that

$$\begin{aligned}
\Delta Y_{it^*}(0) &= Z_i'(\delta^*_{t^*} + \gamma_{t^*}\beta_{t^*}) + X_{it^*-1}(0)(\beta^*_{t^*} + \lambda_{t^*}\beta_{t^*}) + \beta_{t^*}u_{it^*} + \Delta v_{it^*} \\
&:= Z_i'\delta^*_{2,t^*} + X_{it^*-1}(0)\beta^*_{2,t^*} + v_{2,it^*}
\end{aligned}$$

where $\delta^*_{2,t^*} := \delta^*_{t^*} + \gamma_{t^*}\beta_{t^*}$, $\beta^*_{2,t^*} := \beta^*_{t^*} + \lambda_{t^*}\beta_{t^*}$ and $v_{2,it^*} := \beta_{t^*}u_{it^*} + \Delta v_{it^*}$. Further, notice that $\mathbb{E}[v_{2,t^*}|Z, X_{t^*-1}, D = d] = 0$ for $d \in \{0, 1\}$. Thus, in this case, one can estimate δ^*_{2,t^*} and β^*_{2,t^*} from a regression of the change in outcomes over time using the untreated group, and then estimate the ATT from the sample analogue of

$$ATT = \mathbb{E}[\Delta Y_{t^*}|D = 1] - \Big(\mathbb{E}[Z|D = 1]'\delta^*_{2,t^*} + \mathbb{E}[X_{t^*-1}|D = 1]\beta^*_{2,t^*}\Big)$$

Thus, this particular case bypasses the need for actually estimating a separate model for the change in the time-varying covariate over time. This is perhaps not surprising as these are the same conditions as in Chapter 2 where it was sufficient for the researcher to condition on the pre-treatment value of the covariates to recover the ATT.

Case 3: Assumption Cov-Unconfoundedness holds conditional on X_{t^*-1}, W_{t^*-1}, Z

In this case,

$$\begin{aligned}
\mathbb{E}[\Delta X_{t^*}(0)|D = 1] &= \mathbb{E}\left[\mathbb{E}[\Delta X_{t^*}(0)|Z, X_{t^*-1}, W_{t^*-1}, D = 1]|D = 1\right] \\
&= \mathbb{E}[Z|D = 1]'\gamma_{t^*} + \mathbb{E}[X_{t^*-1}|D = 1]\lambda_{t^*} + \mathbb{E}[W_{t^*-1}|D = 1]'\xi_{t^*}
\end{aligned} \tag{4.3}$$

where the first equality holds by the law of iterated expectations, and the second equality holds by Assumption Cov-Unconfoundedness and by assuming a linear model for the change in untreated covariates over time. This suggests estimating $\mathbb{E}[\Delta X_{t^*}(0)|D = 1]$ by running a regression of ΔX_{t^*} on Z, X_{t^*-1}, and W_{t^*-1} using only untreated observations in order to estimate the parameters γ_{t^*}, λ_{t^*}, and ξ_{t^*}, and then to estimate $\mathbb{E}[\Delta X_{t^*}(0)|D = 1]$ by using the sample analogue of the expression in Equation (4.3).[3]

[3] In the special case (and perhaps leading case) considered in Corollary 1 where W_{t^*-1} includes the lagged outcome, Y_{t^*-1} (in addition to all time-invariant covariates and the pre-treatment version of the covariates), one can follow this same procedure with Y_{t^*-1} substituting for W_{t^*-1}.

Case 4: Unconditional Parallel Trends holds for time-varying covariates

For this case, we assume that $\mathbb{E}[\Delta X_{t^*}(0)|D=1] = \mathbb{E}[\Delta X_{t^*}(0)|D=0]$. It immediately follows that

$$ATT = \mathbb{E}[\Delta Y_t|D=1] - \Big(\mathbb{E}[Z|D=1]'\delta^*_{t^*} + \mathbb{E}[\Delta X_{t^*}|D=0]\beta_{t^*} + \mathbb{E}[X_{t^*-1}(0)|D=1]\beta^*_{t^*}\Big)$$

This expression is very similar to the one in Case 1, except that one should use the change in untreated potential covariates *for the untreated group*.

Case 5: Conditional Parallel Trends holds for time-varying covariates

For this case, we assume that $\mathbb{E}[\Delta X_{t^*}(0)|Z, D=1] = \mathbb{E}[\Delta X_{t^*}(0)|Z, D=0]$. In this case,

$$\begin{aligned}\mathbb{E}[\Delta X_{t^*}(0)|D=1] &= \mathbb{E}\left[\mathbb{E}[\Delta X_{t^*}(0)|Z, D=1]\big|D=1\right] \\ &= \mathbb{E}\left[\mathbb{E}[\Delta X_{t^*}|Z, D=0]\big|D=1\right] \\ &= \mathbb{E}[Z|D=1]'\gamma_{t^*} \end{aligned} \tag{4.4}$$

where the first equality holds by the law of iterated expectations, the second equality holds under conditional parallel trends for time-varying covariates, and the last equality holds under a linearity assumption. This suggests estimating γ_{t^*} by running a regression of ΔX_{t^*} on Z using only untreated observations and then to estimate $\mathbb{E}[\Delta X_{t^*}(0)|D=1]$ from the sample analogue of Equation (4.4).

Case 6: Conditional Parallel Trends Holds under Generic Parallel Trends Assumption

For this case, we assume that $\mathbb{E}[\Delta X_{t^*}(0)|Z, W_{t^*-1}, D=1] = \mathbb{E}[\Delta X_{t^*}(0)|Z, W_{t^*-1}, D=0]$. In this case,

$$\begin{aligned}\mathbb{E}[\Delta X_{t^*}(0)|D=1] &= \mathbb{E}\left[\mathbb{E}[\Delta X_{t^*}(0)|Z, W_{t^*-1}, D=1]|D=1\right] \\ &= \mathbb{E}[Z|D=1]'\gamma_{t^*} + \mathbb{E}[W_{t^*-1}|D=1]'\xi_{t^*} \end{aligned} \tag{4.5}$$

where the first equality holds by the law of iterated expectations, and the second equality holds by the conditional parallel trends assumption used in this case and a linearity assumption. Similarly to above, this suggests running a regression of ΔX_{t^*} on Z and W_{t^*-1} using only untreated observations to estimate γ_{t^*} and ξ_{t^*} and then to estimate $\mathbb{E}[\Delta X_{t^*}(0)|D=1]$ from the sample analogue of Equation (4.5).

All of the approaches discussed in this section are substantially more robust than the TWFE regressions discussed in Chapter 3. In particular, unlike TWFE regressions, they allow for the path of untreated potential outcomes to depend on (i) time-invariant covariates, (ii) the pre-treatment level of time-varying covariates, and (iii) the change in time-varying covariates over time that would have occurred if the treatment had not taken place. Given any of a number of assumptions on the path of time-varying covariates in the absence of the treatment (as in Cases 1-6 above), they allow for the treatment to have an effect on time-varying covariates. They allow for general forms of treatment effect heterogeneity; for example, they do not require conditional ATTs to be linear in covariates (as in Assumption 6(a)) nor do they require any of the extra treatment effect homogeneity conditions for TWFE regressions in Proposition 2. Finally, they do not require any linearity conditions for the propensity score as in Assumption 7. Relative to the approach discussed in Chapter 2, the approaches considered in this section require linearity assumptions on the model for untreated potential outcomes and, in some cases, on a model for the change in untreated potential covariates over time. The two main advantages of this approaches in this section are (i) parallel trends assumptions for time varying covariates can be strong enough to identify the ATT, and (ii) the approaches in this section are also easy to implement, only requiring running regressions and computing averages.

Remark 5. Even in the case where $\beta_t = \beta$ (i.e., that the effect of time-varying covariates does not change over time) the strategies proposed in this section would still result in improved estimators relative to the TWFE regressions considered in Chapter 3 as they would still allow for parallel trends to depend on time invariant covariates, allow for general forms of treatment effect heterogeneity, and do not require assumptions on the propensity score. In the special case where $\beta_t = \beta$ *and* Assumption Cov-Exogeneity holds (so that covariates are not affected by the treatment), then the approaches proposed in this section coincide with the "imputation estimators" proposed in Borusyak et al., 2021, but, in general, our approach allows for the path of untreated potential outcomes to depend on both the level and change of time-varying covariates as well as for time-varying covariates to be affected by the treatment.

Remark 6. This section has continued to consider the case with exactly two periods, but it is straightforward to extend these arguments to multiple periods and variation in treatment timing by estimating models for untreated potential outcomes using all available untreated observations (these are observations both for units that do not participate in the treatment in any time period as well as pre-treatment time periods for units that become treated at some point). Once the model for untreated potential outcomes has been estimated, one can "impute" untreated potential outcomes for treated observations, and weighted averages of differences between observed treated potential outcomes and imputed untreated potential outcomes correspond to various treatment effect parameters, depending on the weights chosen by the researcher.

Remark 7. Compare to Imai et al., 2018, in our case weights can depend on Y_{t-1}.

Chapter 5

Empirical Application

To show how imputation techniques can be used in practice, we implement imputation to study the effect of stand your ground laws on homicide rates, using the same data set that was used in Cheng and Hoekstra, 2013. Stand your ground laws allow someone who is attacked in public to use lethal force against their attacker. The data includes information related to stand your ground laws for all 50 states. For the purposes of this paper, we will examine what effect, if any, stand your ground laws had on the number of homicides. First, we will briefly discuss the data. Then, we will explain the estimation strategy used as well as provide comparisons to the results from a TWFE regression and the technique proposed in Callaway and Sant'Anna, 2021.

5.1 Data

The data consists of one observation for each state for each year from 2000-2010 for a total of 550 observations. Each state is given a state id number, which ranges from 1 to 50 and is used to group states. In addition, the "post" variable indicates whether a state has previously passed a stand your ground law. For example, Florida was the first state to pass a stand your ground law in 2006. For the Florida observations for 2000, 2001, 2002, 2003, 2004, and 2005, $post = 0$. Because Florida passed the law in 2006, $post = 1$ for all of the Florida observations from 2006 or later. Although the data set contains over 100 variables, the main focus of this section is on population and homicide count. The outcome variable is the change in the number of homicides. The explanatory variable is whether a state passed a stand your ground law, and the covariates are the population in time period g-1 and the change in population from time period g-1 to t. When using the imputation model, the year 2000 was dropped because the estimation strategies involved using lagged population as a regressor. Summary statistics for treated and untreated states are presented in Table 1.

Table 5.1: Summary Statistics

	Treated		Untreated	
	Mean	Std. Dev.	Mean	Std. Dev.
Homicide Count	371.2	334.1	285.7	447.8
Population	5,957,985.3	5,501,486.9	5,742,704.4	7,040,260.5

5.2 Estimation Strategy

Compared to the two time period case discussed in the theory section, this data set presents additional challenges because states become treated in different years. 2007 was the most common year for stand your ground laws to be passed, but some states passed a version of the law in 2006, 2008, 2009, and 2010 as well. To overcome the issue, we group states by the year they were treated and solve for the ATT for each group using the following procedure:

- We filter out treated states that are not in the group we are currently examining.
- We run a regression of population on lagged population for untreated states, using the results to predict the untreated potential populations at time t for the treated states.
- Finally, we calculate the group's ATT using the states' population in the period before treatment occurred (period g-1) and the predicted change in population between periods t and g-1 as covariates.

For this paper, we estimate the ATT of passing a stand your ground law on the change in homicide count. We compare the results of the standard TWFE estimator, the imputation strategy described above, and an event study using the pte package based on Callaway and Sant'Anna, 2021. The regressor in the event study is the change in population. To calculate the standard errors for the imputation and Callaway and Sant'Anna, 2021 techniques, we use a standard nonparametric bootstrap with 1,000 repetitions.

Table 5.2: Estimated ATT from Different Methods

Statistic	Mean	St. Error
Imputation	22.77	30.66
TWFE	31.27	17.10
Event Study	21.30	22.75

5.3 Results and Analysis

The empirical results are presented in Table 2, which was created using the stargazer R package from Hlavac, 2018. The estimated ATT for imputation was approximately 22.77 with an estimated standard error of 30.66. The estimated ATT from a TWFE regression was 31.27 with an estimated standard error of 17.10. Finally, the estimated ATT from the event study was 21.30, and the estimated standard error was 22.75. All of these results imply that there is no evidence to suggest that passing a stand your ground law changes the number of homicies, as none of the results reject the null hypothesis at the 95% level.

Notably, the imputation estimator is farther from the TWFE estimator, which is known to be biased, than the event study estimator. Callaway and Sant'Anna, 2021 does not account for time varying covariates. However, it is unlikely that stand your ground laws will have any significant impact on a state's population, reducing the bias from time varying covariates. Therefore, the estimates of ATT from imputation and Callaway and Sant'Anna's estimator should indeed be very similar.

Chapter 6

Conclusion

Time-varying covariates pose issues for traditional DID and TWFE estimators. This paper provides conditions in which researchers can uncover the ATT when there are time-varying covariates, including when these covariates may be affected by the treatment. The paper also gives doubly robust estimators to overcome some misspecification and recover the ATT.

In addition, the paper explores when these assumptions are likely to hold. For example, Cov-Exogeneity may hold in the stand your ground application but is unlikely to be a believable assumption when estimating the effects of COVID-19 policies and job displacement. In the latter two cases, researchers may need to make a case that Cov-Unconfoundedness holds.

Finally, the paper discussed an alternative approach to dealing with time-varying covariates in imputation. If conditional parallel trends holds, then researchers can use the observed outcomes of untreated units to predict the untreated potential outcomes of treated units. In the empirical application, the imputation model and the event study predicted smaller effects from passing a stand your ground law than the TWFE model, which may be a result of these methods avoiding the bias present in TWFE models. Future research will focus on the other empirical applications and how researchers can apply these methods to them.

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, *72.1*. http://www.jstor.org/stable/3700681

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, *66*(2), 249–288. http://www.jstor.org/stable/2998558

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Arellano, M., & Honoré, B. (2001). Panel data models: Some recent developments. *Handbook of econometrics* (pp. 3229–3296). Elsevier.

Aronow, P. M., & Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, *60*(1), 250–267.

Athey, S., & Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497.

Barnette, J., Odongo, K., & Reynolds, C. L. (2021). Changes over time in the cost of job loss for young men and women. *The BE Journal of Economic Analysis & Policy*, *21*(1), 335–378.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*(1), 105–134.

Blackwell, M., & Glynn, A. N. (2018). How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review*, *112*(4), 1067–1082.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, *44*(3), 565–640.

Bonhomme, S., & Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, *93*(2), 479–494.

Borusyak, K., Jaravel, X., & Spiess, J. (2021). *Revisiting event study designs: Robust and efficient estimation* [Working Paper].

Brand, J. E. (2006). The effects of job displacement on job quality: Findings from the wisconsin longitudinal study. *Research in Social Stratification and Mobility*, *24*(3), 275–298.

Callaway, B., & Li, T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, *10*(4), 1579–1618.

Callaway, B., & Li, T. (2021). *Policy evaluation during a pandemic* [Working Paper].

Callaway, B., Li, T., & Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, *206*(2), 395–413.

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

Chabé-Ferret, S. (2017). *Should we combine difference in differences with conditioning on pre-treatment outcomes?* [Working paper].

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, *23*(2), 177–191.

Cheng, C., & Hoekstra, M. (2013). Does strengthening self-defense law deter crime or escalate violence? evidence from expansions to castle doctrine. *Journal of Human Resources*, *48*(3), 821–854.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chetty, R., Friedman, J., Hendren, N., & Stepner, M. (2020). The economic impacts of covid-19: Evidence from a new public database built from private sector data. *Opportunity Insights*.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996.

Flores, C. A., & Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness.

Gardner, J. (2021). *Two-stage difference in differences* [Working Paper].

Goldsmith-Pinkham, P., Hull, P., & Kolesár, M. (2021). *On estimating multiple treatment effects with regression* [Working Paper].

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

Gupta, S., Montenovo, L., Nguyen, T. D., Lozano-Rojas, F., Schmutte, I. M., Simon, K. I., Weinberg, B. A., & Wing, C. (2020). *Effects of social distancing policy on labor market outcomes* [Working Paper].

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, *66*(5), 1017–1098.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, *64*(4), 605–654.

Hlavac, M. (2018). Stargazer: Well-Formatted regression and summary statistics tables. https://CRAN.R-project.org/package=stargazer

Huber, M. (2020). Mediation analysis. In K. F. Zimmermann (Ed.), *Handbook of labor, human resources and population economics* (pp. 1–38). Springer International Publishing.

Imai, K., Kim, I. S., & Wang, E. (2018). *Matching methods for causal inference with time-series cross-section data* [Working Paper].

Imbens, G., & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Ishimaru, S. (2021). *Empirical decomposition of the iv-ols gap with heterogeneous and nonlinear effects* [Working Paper].

Ishimaru, S. (2022). *What do we get from a two-way fixed effects estimator? implications from a general numerical equivalence* [Working Paper].

Jacobson, L. S., LaLonde, R. J., & Sullivan, D. G. (1993). Earnings losses of displaced workers. *American Economic Review*, 685–709.

Lechner, M. (2008). A note on endogenous control variables in causal studies. *Statistics & Probability Letters*, *78*(2), 190–195.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165–224.

Liu, L., Wang, Y., & Xu, Y. (2021). *A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data* [Working Paper].

Melly, B., & Santangelo, G. (2015). *The changes-in-changes model with covariates* [Working Paper].

Robins, J. M. (1997). Causal inference from complex longitudinal data. *Latent variable modeling and applications to causality* (pp. 69–117). Springer.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 551.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, *147*(5), 656–666.

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

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Słoczyński, T. (2020). Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *The Review of Economics and Statistics*, 1–27.

Słoczyński, T., & Wooldridge, J. M. (2018). A general double robustness result for estimating average treatment effects. *Econometric Theory*, *34*(1), 112–133.

Stevens, A. H. (1997). Persistent effects of job displacement: The importance of multiple job losses. *Journal of Labor Economics*, 165–188.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, 613–641.

Topel, R. (1991). Specific capital, mobility, and wages: Wages rise with job seniority. *Journal of Political Economy*, *99*(1), 145–176.

Weill, J. A., Stigler, M., Deschenes, O., & Springborn, M. R. (2021). *Researchers' degrees-of-flexibility and the credibility of difference-in-differences estimates: Evidence from the pandemic policy evaluations* [Working Paper].

Zeldow, B., & Hatfield, L. A. (2021). Confounding and regression adjustment in difference-in-differences studies. *Health Services Research*.

Appendix A

A.1 Main Proofs

Proof of Theorem 1

Proof. To start, notice that

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

where the first equality is just the definition of ATT, the second equality holds by adding and subtracting $\mathbb{E}[Y_{t^*-1}(0)|D=1]$, and the third equality holds by writing potential outcomes in terms of their observed counterparts. For part (1), further notice that,

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

where the first equality holds by the law of iterated expectations, the second equality holds by Assumption 2, and the last equality holds because $\Delta Y_{t^*}(0)$ and $X_{t^*}(0)$ are observed for the untreated group and uses Assumption Cov-Exogeneity to integrate over the distribution of observed covariates (i.e., treated potential covariates) for the treated group. Combining this expression with the previous one for ATT completes the proof for part (1) of the result.

For part (2), notice that

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

where the first equality holds by the law of iterated expectations, the second equality holds by Assumption 2 (unlike part (1), this term is not immediately identified because we do not have an immediate analogue of the distribution of $X_{t^*}(0)$ in order to identify the outer expectation), the third equality holds by the law of iterated expectations, the fourth equality holds by Assumption Cov-Unconfoundedness (because, after conditioning on $(X_{t^*-1}, W_{t^*-1}, Z)$, the only randomness comes from $X_{t^*}(0)$), the fifth equality holds by writing potential outcomes in terms of their observed counterparts, and this term is identified because the distribution of $(X_{t^*-1}, W_{t^*-1}, Z)$ is identified for the treated group. □

Proof of Corollary 1

Proof. For part (1), the result holds immediately by the law of iterated expectations. For part (2), the result holds immediately from the expression in part (2) of Theorem 1 using $W_{t^*-1} = Y_{t^*-1}$. □

Proof of Corollary 2

Proof. For part (1), we omit the proof as, after invoking Assumption Cov-Exogeneity, this becomes the same case as with time invariant covariates — see, for example, Sant'Anna and Zhao (2020) for this sort of result in the case with time invariant covariates. Given the expression for the ATT in part (1) of Corollary 1, the proof of part (2) follows using the same arguments as for part (1). □

Proof of Proposition 1

We prove the result in several steps.

To start, consider the numerator in the expression for α in Equation (1.1). Notice that

$$\begin{aligned}\mathbb{E}[De] &= \mathbb{E}[D(\Delta Y_{t^*} - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*})] \\ &= \mathbb{E}[D(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0)])] + \mathbb{E}[D(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0)] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*})]\end{aligned} \tag{A.1}$$

We provide results for each of the terms in Equation (A.1) next.

Lemma 1. *Under Assumptions 1, 2, and 3(a),*

$$\mathbb{E}[D(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0)])] = \mathbb{E}\left[\mathbb{E}[D](1-p(X^{all}(0)))ATT_{X^{all}(0)}(X^{all}(0))\Big|D=1\right]$$

Proof.

$$\begin{aligned}&\mathbb{E}[D(\Delta Y_{t^*} - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0)])] \\ &\quad= \mathbb{E}\left[D\left(\Delta Y_{t^*} - (\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1]p(X^{all}(0))\right.\right. \\ &\qquad - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0](1-p(X^{all}(0))) \\ &\quad= \mathbb{E}\left[\mathbb{E}[D\Delta Y_{t^*}|X^{all}(0)]\right. \\ &\qquad -p(X^{all}(0))\left(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1]p(X^{all}(0))\right. \\ &\qquad - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0](1-p(X^{all}(0)) \\ &\quad= \mathbb{E}\left[\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1]p(X^{all}(0))\right. \\ &\qquad -p(X^{all}(0))\left(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1]p(X^{all}(0))\right. \\ &\qquad - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0](1-p(X^{all}(0)) \\ &\quad= \mathbb{E}\left[p(X^{all}(0))(1-p(X^{all}(0)))\left(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1] - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0]\right)\right] \\ &\quad= \mathbb{E}\left[\mathbb{E}[D](1-p(X^{all}(0)))\left(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1] - \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0]\right)\Big|D=1\right] \\ &\quad= \mathbb{E}\left[\mathbb{E}[D](1-p(X^{all}(0)))ATT_{X^{all}(0)}(X^{all}(0))\Big|D=1\right]\end{aligned}$$

where the first three equalities hold by repeatedly applying the law of iterated expectations, the fourth equality holds by rearranging terms, the fifth equality holds by integrating over the distribution of $X^{all}(0)$ conditional on $D=1$ and re-weighting, and the last equality holds under the conditional parallel trends assumption in Assumption 2. □

Next, we provide a result for the second term in Equation (A.1).

Lemma 2. *Under Assumptions 1, 2, and 3(a),*

$$\begin{aligned}
&\mathbb{E}[D(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0)] - L(\Delta Y_{t^*}|\Delta X_{t^*})] \\
&\quad = \mathbb{E}\Big[\mathbb{E}[D]\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1] - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\Big)p(X^{all}(0)) \\
&\qquad -\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0] - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\Big)(1-p(X^{all}(0)))\Big\}\Big|D=1\Big] \\
&\qquad + \mathbb{E}\left[\mathbb{E}[D]\left\{L(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\right.\right. \\
&\qquad - L(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)(p(X^{all}(0)) - L(D|\Delta X_{t^*}))\Big|D=1]
\end{aligned}$$

Proof. Notice that

$$\begin{aligned}
&\mathbb{E}[D(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0)] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*})] \\
&\quad = \mathbb{E}[p(X^{all}(0))(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0)] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*})] \\
&\quad = \mathbb{E}[p(X^{all}(0))\{\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1]p(X^{all}(0)) \\
&\qquad + \mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0](1-p(X^{all}(0))]\}] \\
&\qquad - \mathbb{E}[p(X^{all}(0))\{\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\mathrm{L}(D|\Delta X_{t^*}) \\
&\qquad + \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)(1-\mathrm{L}(D|\Delta X_{t^*})\}] \\
&\quad = \mathbb{E}\Big[p(X^{all}(0))\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\Big)p(X^{all}(0)) \\
&\qquad +\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\Big)(1-p(X^{all}(0)))\Big\}\Big] \\
&\quad + \mathbb{E}\left[p(X^{all}(0))\left\{\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\right.\right. \\
&\quad - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)(p(X^{all}(0)) - \mathrm{L}(D|\Delta X_{t^*})) \\
&\quad = \mathbb{E}\Big[\mathbb{E}[D]\Big\{\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=1] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)\Big)p(X^{all}(0)) \\
&\qquad +\Big(\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D=0] - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\Big)(1-p(X^{all}(0)))\Big\}\Big|D=1\Big] \\
&\qquad + \mathbb{E}\left[\mathbb{E}[D]\left\{\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1) - \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)\right\}\right. \\
&\qquad * (p(X^{all}(0)) - \mathrm{L}(D|\Delta X_{t^*}))\Big|D=1
\end{aligned}$$

where the first equality holds by applying the law of iterated expectations, the second equality holds by applying the law of iterated expectations and the law of iterated projections, the third equality holds by adding and subtracting $\mathbb{E}[\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=1)p(X^{all}(0))^2]$ and $\mathbb{E}[\mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D=0)p(X^{all}(0))(1-p(X^{all}(0)))]$ and rearranging terms, the fourth equality holds by applying the law of iterated expectations to each each term. This completes the proof. □

Next, we provide a result on decomposing differences between the conditional expectation of ΔY_{t^*} (conditional on the full vector $X^{all}(0)$) and the linear projection of ΔY_{t^*} on ΔX_{t^*}.

Lemma 3. *Under Assumptions 1, 2, and 3(a) and for* $d \in \{0, 1\}$,

$$
\begin{aligned}
&\mathbb{E}[\Delta Y_{t^*}|X^{all}(0), D = d] - L(\Delta Y_{t^*}|\Delta X_{t^*}, D = d) \\
&\quad = \mathbb{E}[\Delta Y_{t^*}|X_t(0), X_{t-1}, Z, D = d] - \mathbb{E}[\Delta Y_{t^*}|X_t, X_{t-1}, Z, D = d] && \text{(A)} \\
&\qquad + \mathbb{E}[\Delta Y_{t^*}|X_t, X_{t-1}, Z, D = d] - \mathbb{E}[\Delta Y_{t^*}|X_t, X_{t-1}, D = d] && \text{(B)} \\
&\qquad + \mathbb{E}[\Delta Y_{t^*}|X_t, X_{t-1}, D = d] - \mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D = d] && \text{(C)} \\
&\qquad + \mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D = d] - L(\Delta Y_{t^*}|\Delta X_{t^*}, D = d) && \text{(D)}
\end{aligned}
$$

Proof. The result holds immediately just by adding and subtracting terms. □

Next, we provide a useful result for the denominator in the expression for α in Equation (1.1).

Lemma 4. *Under Assumptions 1, 2, and 3(a),*

$$
\begin{aligned}
\mathbb{E}[u^2] &= \mathbb{E}[p(\Delta X_{t^*})(1 - L(D|\Delta X_{t^*}))] \\
&= \mathbb{E}\big[\mathbb{E}[D](1 - L(D|\Delta X_{t^*}))|D = 1\big]
\end{aligned}
$$

Proof. From the definition of u, it follows that

$$
\begin{aligned}
\mathbb{E}\left[u^2\right] &= \mathbb{E}\left[(D - \mathrm{L}(D|\Delta X_{t^*}))^2\right] \\
&= \mathbb{E}[D] - 2\mathbb{E}[D\mathrm{L}(D|\Delta X_{t^*})] + \mathbb{E}[\mathrm{L}(D|\Delta X_{t^*})^2] \\
&:= A_1 - 2A_2 + A_3 && \text{(A.2)}
\end{aligned}
$$

and we consider each of these in turn. Start with,

$$
\begin{aligned}
A_2 &= \mathbb{E}[D\mathrm{L}(D|\Delta X_{t^*})] \\
&= \mathbb{E}[D\Delta X_{t^*}']\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\mathbb{E}[\Delta X_{t^*}D] \\
&= \mathbb{E}[\Delta X_{t^*}D]'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\mathbb{E}[\Delta X_{t^*}D] && \text{(A.3)}
\end{aligned}
$$

where the first equality holds from the definition of $\mathrm{L}(D|\Delta X_{t^*})$ and the second equality holds immediately from the previous one. Next,

$$
\begin{aligned}
A_3 &= \mathbb{E}[\mathrm{L}(D|\Delta X_{t^*})^2] \\
&= \mathbb{E}\left[\mathbb{E}[\Delta X_{t^*}D]'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\Delta X_{t^*}\Delta X_{t^*}'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\mathbb{E}[\Delta X_{t^*}D]\right] \\
&= \mathbb{E}[\Delta X_{t^*}D]'\mathbb{E}[\Delta X_{t^*}\Delta X_{t^*}']^{-1}\mathbb{E}[\Delta X_{t^*}D] && \text{(A.4)}
\end{aligned}
$$

where the first equality holds by the definition of A_3, the second equality holds by the definition of $\mathrm{L}(D|\Delta X_{t^*})$, and the last equality holds by canceling terms. Plugging Equations (A.3) and (A.4) in Equation (A.2) implies that

$$\begin{aligned}\mathbb{E}[u^2] &= \mathbb{E}[D(1 - \mathrm{L}(D|\Delta X_{t^*}))] \\ &= \mathbb{E}[p(\Delta X_{t^*})(1 - \mathrm{L}(D|\Delta X_{t^*}))] \\ &= \mathbb{E}[\mathbb{E}[D](1 - \mathrm{L}(D|\Delta X_{t^*}))|D = 1]\end{aligned}$$

where the second and third equalities hold by the law of iterated expectations and which completes the proof. □

Proof of Proposition 1. The first part of the expression for α comes from Equation (A.1) and by Lemma 1 and Lemma 4. The second and third parts come from Equation (A.1) and by Lemmas 2 to 4. □

Proof of Proposition 2

To show the first part of the result, we show that each of Terms (A)-(D) in Proposition 1 are equal to zero under the extra conditions in this proposition.

Term (A): First, consider the expression in Term (A) for $d = 0$. Notice that,

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

which holds because untreated potential covariates are equal to observed covariates for the untreated group. Second, consider the case when $d = 1$. In this case,

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

holds immediately by Assumption Cov-Exogeneity. Thus, Term (A) is equal to zero under Assumption Cov-Exogeneity.

Term (B): First, consider the expression in Term (B) for $d = 0$. Notice that,

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

under Assumption 4(b) because, conditional on being in the treated group, the observed change in outcomes is equal to the change in untreated potential outcomes and the observed X_{t^*} is equal to $X_{t^*}(0)$.

Second, consider the case when $d = 1$, in this case

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

where the first equality holds by adding and subtracting terms, the second equality holds by Assumption Cov-Exogeneity and the definition of $ATT_{X^{all}(0)}$, by Assumption 2, and by Assumption 4(b), the third equality holds by Assumption 4(a), the fourth equality holds by adding and subtracting terms and by Assumption Cov-Exogeneity, and the last equality holds because

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

where the first equality holds by the law of iterated expectations, the second equality holds by Assumption 2, the third equality holds by Assumption 4(b), and the last equality holds because, conditional on $X_{t^*}(0)$, and X_{t^*-1}, the inside conditional expectation is non-random. Thus, Term (B) is equal to zero under Assumption Cov-Exogeneity and Assumption 4.

Term (C): First, consider the expression in Term (C) for $d = 0$. Notice that,

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

where the first equality holds by replacing observed counterparts with their corresponding potential outcomes, the second equality holds by Assumption 5(b), and the third equality holds by writing potential outcomes in terms of their observed counterparts. Second, consider the case when $d = 1$, so that

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

where the first equality holds by adding and subtracting terms, the second equality holds by Assumption Cov-Exogeneity, the definition of $ATT_{X_{t^*}(0), X_{t^*-1}}$, Assumption 5(b), and the middle term is equal to zero by the same arguments as were used for Term (B) above, the third equality holds by Assumption 5(a), the fourth equality holds by adding and subtracting terms and by Assumption Cov-Exogeneity, and the last equality holds because

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

where the first equality holds by the law of iterated expectations, the second equality holds by Assumption 2, the third equality holds by Assumption 4(b) and Assumption 5(b), and the last equality holds because, conditional on $\Delta X_{t^*}(0)$, the inside conditional expectation is nonrandom. Thus, Term (C) is equal to zero under Assumption Cov-Exogeneity and Assumptions 4 and 5.

Term (D): First, consider the expression in Term (D) for $d = 0$. Notice that

$$\begin{aligned}
\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D = 0] &= \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}(0), D = 0] \\
&= \Delta X_{t^*}(0)'\delta_0 \\
&= \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D = 0)
\end{aligned}$$

where the first equality holds by writing observed outcomes and time-varying covariates in terms of potential outcomes/covariates, the second equality holds by Assumption 6(b), and the last equality holds by the definition of linear projection. Next, consider the expression in Term (D) for $d = 1$,

$$\begin{aligned}
\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D = 1] &= \Big(\mathbb{E}[\Delta Y_{t^*}|\Delta X_{t^*}, D = 1] - \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}(0), D = 1]\Big) \\
&\quad + \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}(0), D = 1] \\
&= ATT_{\Delta X_{t^*}(0)}(\Delta X_{t^*}(0)) + \mathbb{E}[\Delta Y_{t^*}(0)|\Delta X_{t^*}(0), D = 1] \\
&= \Delta X_{t^*}(0)'(\delta_1 + \delta_0) \\
&= \mathrm{L}(\Delta Y_{t^*}|\Delta X_{t^*}, D = 1)
\end{aligned}$$

where the first equality holds by adding and subtracting terms, the second equality holds using similar arguments as for previous terms and uses Assumptions 2, 4 and 5 and Assumption Cov-Exogeneity, the third equality holds by Assumption 6, and the last equality holds by the definition of linear projection where the linear projection coefficient is given by $\delta_1 + \delta_0$.

This completes the first part of the proof. Next, we prove additional result (a) in Proposition 2. Toward this end, recall that,

$$\omega_e(X^{all}(0)) = \frac{(p(X^{all}(0)) - \mathrm{L}(D|\Delta X))}{\mathbb{E}[(1 - \mathrm{L}(D|\Delta X_{t^*}))|D = 1]}$$

Under Assumption 7, $p(X^{all}(0)) = \mathrm{L}(D|\Delta X_{t^*})$ which implies that $\omega_e(X^{all}(0)) = 0$. Next, recall that

$$\begin{aligned}
\omega_{ATT}(X^{all}(0)) &= \frac{1 - p(X^{all}(0))}{\mathbb{E}[(1 - \mathrm{L}(D|\Delta X))|D = 1]} \\
&= \frac{1 - p(X^{all}(0))}{\mathbb{E}[(1 - p(X^{all}(0))|D = 1]}
\end{aligned}$$

where the second equality holds under Assumption 7. It immediately follows that $\mathbb{E}[\omega_{ATT}(X^{all}(0))|D = 1] = 1$ This completes the proof of additional result (a). Now, we move to proving additional result (b) in Proposition 2. Under Assumptions 1 to 7 and Assumption Cov-Exogeneity, we have shown that

$$\begin{aligned}\alpha &= \mathbb{E}[\omega_{ATT}(X^{all}(0))ATT_{\Delta X_{t^*}(0)}(\Delta X_{t^*}(0))|D=1]\\ &= ATT\,\mathbb{E}[\omega_{ATT}(X^{all}(0))|D=1]\\ &= ATT\end{aligned}$$

where the second equality holds when $ATT_{\Delta X_{t^*}(0)}$ does not vary across different values of $\Delta X_{t^*}(0)$, and the last equality holds because the weights have mean one.

A.2 Proof of Imputation Asymptotic Normality

A.2.1 Asymptotic Normality in the Two Period Case

ATT is defined as the following:

$$\begin{aligned}ATT = E[\Delta Y_{t*}|D=1] - (&E[Z|D=1]'\delta^*_{t^*}\\ &+ E[\Delta X_{t^*}(0)|D=1]\tilde{\beta}_{t^*} + E[X_{t^*-1}(0)\beta_{t^*}|D=1])\end{aligned}$$

$\widehat{ATT}$ is the sample analog of ATT. Therefore:

$$\widehat{ATT} = \frac{1}{n}\sum_{i=1}^{n}\frac{Di}{p}\Delta Y_{it} - \frac{Di}{\hat{p}}(Z_i\hat{\delta}_t + \Delta X_{it}\hat{\tilde{\beta}}_{t^*} + X_{it-1}\hat{\beta}_t), \tag{A.5}$$

where p = E[D]. We can estimate p in the following way:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{A.6}$$

Note that since period t-1 occurs before treatment, $X_{it-1} = X_{it-1}(0)\ \forall i$. Thus,

$$
\begin{aligned}
\sqrt{n}(\widehat{ATT} - ATT) &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta Y_{it}) - E[\frac{D}{\hat{p}}\Delta Y_t]) \\
&\quad - (\sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}Z_i\hat{\delta}_t) - E[\frac{D}{p}Z\delta_t] \\
&\quad + \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta X_{it^*}\hat{\tilde{\beta}}_{t^*}) - E[\Delta X_{t^*}(0)|D=1]\tilde{\beta}_{t^*})) \\
&\quad + \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}X_{it-1}\hat{\beta}_t) - E[X_{t^*-1}|D=1]\beta_{t^*}) \\
&= A - B - C - D \\
&= \frac{1}{\sqrt{(n)}}\sum_{i=1}^{n}\Psi \\
&\to N(0,\Omega)
\end{aligned}
$$

We now examine, A, B, C, and D.

For A:

$$
\begin{aligned}
A &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}\Delta Y_{it}) - E[\frac{D}{p}\Delta Y_t]) \\
&= \sqrt{n}(E[\frac{D}{p}\Delta Y_t] + o_p(1) - E[\frac{D}{p}\Delta Y_t]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For B:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\delta} - \delta)$. We need to find an expression for $\hat{\delta}$. We do so using FWL theorem arguments. So, we begin by regressing Y on X_{t-1} and ΔX_t:

$$
\begin{aligned}
Y &= X_{t-1}\hat{\beta}_t + \Delta X_{t^*}\hat{\tilde{\beta}}_{t^*} + e \\
&= b + e \\
\to e &= Y - b
\end{aligned}
$$

Next, we will regress Z on the other explanatory variables:

$$
\begin{aligned}
Z &= b + \epsilon \\
\rightarrow \epsilon &= Z - b
\end{aligned}
$$

Finally, we regress e on ϵ:

$$
\begin{aligned}
e &= \epsilon\hat{\delta} + \zeta \\
\hat{\delta} &= [\epsilon'\epsilon]^{-1}\epsilon'(e - \zeta) \\
&= [(Z-b)'(Z-b)]^{-1}(Z-b)'(e-\zeta) \\
&= [(Z - X_{t-1}\hat{\beta}_t - \Delta X_{t^*}(0)\hat{\tilde{\beta}}_{t^*})'(Z - X_{t-1}\hat{\beta}_t - \Delta X_{t^*}(0)\hat{\tilde{\beta}}_{t^*})]^{-1} \\
&\quad * (Z - X_{t-1}\hat{\beta}_t - \Delta X_{t^*}(0)\hat{\tilde{\beta}}_{t^*})'(e-\zeta) \\
\rightarrow E[\hat{\delta}] &= [(Z - X_{t-1}\beta_t - \Delta X_{t^*}(0)\tilde{\beta}_{t^*})'(Z - X_{t-1}\beta_t - \Delta X_{t^*}(0)\tilde{\beta}_{t^*})]^{-1} \\
&\quad * (Z - X_{t-1}\beta_t - \Delta X_{t^*}(0)\tilde{\beta}_{t^*})'(e-\zeta) \\
&= \delta.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sqrt{n}\frac{1}{n}(\hat{\delta} - \delta) &= E[(Z-b)'(Z-b)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((Z_i - b_i)'Y + (Z_i - b_i)'\zeta_i - (Z_i - b_i)'Y) \\
&= E[(Z-b)'(Z-b)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((Z_i - b_i)'\zeta_i).
\end{aligned}
$$

Then, we can solve for B:

$$
\begin{aligned}
B &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}(Z_i - b_i)'\hat{\delta}_t) - E[\frac{D}{p}(Z-b)\delta_t)) \\
&= \sqrt{n}(E[\frac{D}{p}(Z-b)\delta_t] + o_p(1) - E[\frac{D}{p}(Z-b)\delta_t]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For C:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\tilde{\beta}}_t - \tilde{\beta}_t)$. Using similar arguments as before:

$$
\begin{aligned}
\hat{\hat{\beta}}_{t^*} &= [(\Delta X_{t^*}(0) - X_{t-1}\hat{\beta}_t - Z\hat{\delta}_t)'((\Delta X_{t^*}(0) - X_{t-1}\hat{\beta}_t - Z\hat{\delta}_t)]^{-1} \\
&\quad * (\Delta X_{t^*}(0) - X_{t-1}\hat{\beta}_t - Z\hat{\delta}_t)'(e - \zeta) \\
&= [(\Delta X_{t^*}(0) - c)'(\Delta X_{t^*}(0) - c)]^{-1}(\Delta X_{t^*}(0) - c)'(e - \zeta) \\
\rightarrow E[\hat{\hat{\beta}}_{t^*}] &= [(\Delta X_{t^*}(0) - X_{t-1}\beta_t - Z\delta_t)'(\Delta X_{t^*} - X_{t-1}\beta_t - Z\delta_t)]^{-1} \\
&\quad * (\Delta X_{t^*}(0) - X_{t-1}\beta_t - Z\delta_t)'e \\
&= \tilde{\beta}_{t^*},
\end{aligned}
$$

where e is the residuals from regressing the Z and X_{t-1} on Y. Then:

$$
\begin{aligned}
\sqrt{n}\frac{1}{n}(\hat{\hat{\beta}}_t - \tilde{\beta}_t) &= \sqrt{n}\frac{1}{n}(\hat{\hat{\beta}}_t - \tilde{\beta}_t) \\
&= E[(\Delta X_{t^*}(0) - c)'(\Delta X_{t^*}(0) - c)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((\Delta X_{it^*}(0) - c_i)'e_i) \\
&= E[\frac{1-D}{1-p}(\Delta X_{t^*} - c)'(\Delta X_{t^*} - c)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{1-D_i}{1-\hat{p}}(\Delta X_{it^*} - c_i)'e_i).
\end{aligned}
$$

Then, we can solve for C:

$$
\begin{aligned}
C &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta X_{it}(0)\hat{\hat{\beta}}_{t^*}) - E[\frac{D}{p}\Delta X_{t^*}(0)]\tilde{\beta}_{t^*}) \\
&= \sqrt{n}(E[\frac{D}{p}\Delta X_{t^*}(0)\tilde{\beta}] + o_p(1) - E[\frac{D}{p}\Delta X_{t^*}(0)\tilde{\beta}]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For D:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\beta}_t - \beta_t)$. Using similar arguments as before:

$$
\begin{aligned}
\hat{\beta}_t &= [(X_{t-1} - \Delta X_{t^*}(0)\hat{\hat{\beta}}_{t^*} - Z\hat{\delta}_t)'(X_{t-1} - \Delta X_{t^*}(0)\hat{\hat{\beta}}_{t^*} - Z\hat{\delta}_t)]^{-1} \\
&\quad * (X_{t-1} - \Delta X_{t^*}(0)\hat{\hat{\beta}}_{t^*} - Z\hat{\delta}_t)'(e - \zeta) \\
&= [(X_{t-1} - d)'(X_{t-1} - d)]^{-1}(X_{t-1} - d)'(e - \zeta) \\
\rightarrow E[\hat{\beta}_t] &= [(X_{t-1} - \Delta X_{t^*}(0)\tilde{\beta}_{t^*} - Z\delta_t)'(X_{t-1} - \Delta X_{t^*}(0)\tilde{\beta}_{t^*} - Z\delta_t)]^{-1} \\
&\quad * (X_{t-1} - \Delta X_{t^*}(0)\tilde{\beta}_{t^*} - Z\delta_t)'(e) \\
&= \beta_t,
\end{aligned}
$$

where e is the residuals from regressing the Z and $\Delta X_{t^*}(0)\tilde{\beta}_{t^*}$ on Y. Then, we can solve for D:

$$\begin{aligned} D &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}X_{it-1}\hat{\beta}_{t^*}) - E[\frac{D}{p}X_{t-1}]\beta_{t^*}) \\ &= \sqrt{n}(E[\frac{D}{p}X_{t-1}\beta_{t^*}] + o_p(1) - E[\frac{D}{p}X_{t-1}\beta_{t^*}]) \\ &= o_p(1), \end{aligned}$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

A.2.2 Asymptotic Normality in the Multiple-Period Case

Having proved asymptotic normality in the two-period case, it is straightforward to prove asymptotic normality in the multiple-period case. Instead of simply finding the ATT, we will find the $ATT_{(g,t)}$, where g denotes an initial treatment period group we are examining and t represents a time period such that $t \geq g$. The treatment group will consist of units that were treated in time period g, and the untreated group will consist of units that were not treated by time period t. Then, we can proceed as we did in the two-period case:

$$\begin{aligned} \sqrt{n}(\widehat{ATT}_{(g,t)} - ATT_{(g,t)}) &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta Y_{i,g,t}) - E[\frac{D}{\hat{p}}\Delta Y_{g,t}]) \\ &- (\sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}Z_i\hat{\delta}_{g,t}) - E[\frac{D}{p}Z\delta_{g,t}] \\ &+ \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta X_{i,g,t^*}\hat{\tilde{\beta}}_{g,t^*}) - E[\Delta X_{g,t^*}(0)|D=1]\tilde{\beta}_{g,t^*})) \\ &+ \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}X_{i,g-1,t}\hat{\beta}_t) - E[X_{g-1,t}|D=1]\beta_{g,t^*}) \\ &= A - B - C - D \\ &= \frac{1}{\sqrt{(n)}}\sum_{i=1}^{n}\Psi_{(g,t)} \\ &\to N(0,\Omega_{(g,t)}) \end{aligned}$$

We now examine, A, B, C, and D.

For A:

$$
\begin{aligned}
A &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}\Delta Y_{i,g,t}) - E[\frac{D}{p}\Delta Y_{g,t}]) \\
&= \sqrt{n}(E[\frac{D}{p}\Delta Y_{g,t}] + o_p(1) - E[\frac{D}{p}\Delta Y_{g,t}]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For B:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\delta} - \delta)$. We need to find an expression for $\hat{\delta}$. We do so using FWL theorem arguments. So, we will begin by regressing Y on $X_{g-1,t}$ and $\Delta X_{g,t}$:

$$
\begin{aligned}
Y_{g,t} &= X_{g-1,t}\hat{\beta_{g,t}} + \Delta X_{g,t^*}\hat{\hat{\beta}}_{g,t^*} + e \\
&= b_{g,t} + e_{g,t} \\
\rightarrow e_{g,t} &= Y_{g,t} - b_{g,t}
\end{aligned}
$$

Next, we regress $Z_{g,t}$ on the other explanatory variables:

$$
\begin{aligned}
Z_{g,t} &= b_{g,t} + \epsilon \\
\rightarrow \epsilon &= Z_{g,t} - b_{g,t}
\end{aligned}
$$

Finally, we regress e on ϵ:

$$
\begin{aligned}
e &= \epsilon\hat{\delta} + \zeta_{g,t} \\
\hat{\delta}_{g,t} &= [\epsilon'\epsilon]^{-1}\epsilon'(e - \zeta_{g,t}) \\
&= [(Z_{g,t} - b_{g,t})'(Z_{g,t} - b_{g,t})]^{-1}(Z_{g,t} - b_{g,t})'(e - \zeta_{g,t}) \\
&= [(Z_{g,t} - X_{g-1,t}\hat{\beta}_{g,t} - \Delta X_{g,t^*}(0)\hat{\hat{\beta}}_{g,t^*})'(Z_{g,t} - X_{g-1,t}\hat{\beta}_{g,t} - \Delta X_{g,t^*}(0)\hat{\hat{\beta}}_{g,t^*})]^{-1} \\
&\quad * (Z_{g,t} - X_{g-1,t}\hat{\beta}_t - \Delta X_{g,t^*}(0)\hat{\hat{\beta}}_{g,t^*})'(e - \zeta_{g,t}) \\
\rightarrow E[\hat{\delta}] &= [(Z_{g,t} - X_{g-1,t}\beta_{g,t} - \Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*})'(Z_{g,t} - X_{g-1,t}\beta_{g,t} - \Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*})]^{-1} \\
&\quad * (Z_{g,t} - X_{g-1,t}\beta_{g,t} - \Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*})'(e - \zeta_{g,t}) \\
&= \delta_{g,t}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sqrt{n}\frac{1}{n}(\hat{\delta}_{g,t} - \delta_{g,t}) &= E[(Z_{g,t} - b_{g,t})'(Z_{g,t} - b_{g,t})]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((Z_{i,g,t} - b_{i,g,t})'Y \\
&+ (Z_{i,g,t} - b_{i,g,t})'\zeta_{i,g,t} - (Z_{i,g,t} - b_{i,g,t})'Y) \\
&= E[(Z_{g,t} - b_{g,t})'(Z_{g,t} - b_{g,t})]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((Z_{i,g,t} - b_{i,g,t})'\zeta_i)
\end{aligned}
$$

Then, we can solve for B:

$$
\begin{aligned}
B &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}(Z_{i,g,t} - b_{i,g,t})'\hat{\delta}_{g,t}) - E[\frac{D}{p}(Z_{g,t} - b_{g,t})\delta_{g,t})) \\
&= \sqrt{n}(E[\frac{D}{p}(Z_{g,t} - b_{g,t})\delta_{g,t}] + o_p(1) - E[\frac{D}{p}(Z_{g,t} - b_{g,t})\delta_t]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For C:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\tilde{\beta}}_{g,t} - \tilde{\beta}_{g,t})$. Using similar arguments as before:

$$
\begin{aligned}
\hat{\tilde{\beta}}_{g,t^*} &= [(\Delta X_{g,t^*}(0) - X_{g-1,t}\hat{\beta_{g,t}} - Z\hat{\delta}_{g,t})'((\Delta X_{g,t^*}(0) - X_{g-1,t}\hat{\beta_{g,t}} - Z\hat{\delta}_{g,t})]^{-1} \\
&* (\Delta X_{g,t^*}(0) - X_{g-1,t}\hat{\beta_{g,t}} - Z\hat{\delta}_{g,t})'(e - \zeta) \\
&= [(\Delta X_{g,t^*}(0) - c)'(\Delta X_{g,t^*}(0) - c_{g,t})]^{-1}(\Delta X_{g,t^*}(0) - c_{g,t})'(e - \zeta) \\
\rightarrow E[\hat{\tilde{\beta}}_{g,t^*}] &= [(\Delta X_{g,t^*}(0) - X_{g-1,t}\beta_{g,t} - Z\delta_{g,t})'(\Delta X_{g,t^*} - X_{g-1,t}\beta_t - Z\delta_{g,t})]^{-1} \\
&* (\Delta X_{g,t^*}(0) - X_{g-1,t}\beta_t - Z\delta_{g,t})'e \\
&= \tilde{\beta}_{g,t^*},
\end{aligned}
$$

where e is the residuals from regressing the $Z_{g,t}$ and $X_{g-1,t}$ on Y. Then:

$$
\begin{aligned}
\sqrt{n}\frac{1}{n}(\hat{\tilde{\beta}}_{g,t}-\tilde{\beta}_{g,t}) &= \sqrt{n}\frac{1}{n}(\hat{\tilde{\beta}}_{g,t}-\tilde{\beta}_{g,t}) \\
&= E[(\Delta X_{g,t^*}(0)-c_{g,t})'(\Delta X_{g,t^*}(0)-c_{g,t})]^{-1} \\
&\quad * \frac{1}{\sqrt{n}}\sum_{i=1}^{n}((\Delta X_{i,g,t^*}(0)-c_{i,g,t})'e_{i,g,t}) \\
&= E[\frac{1-D}{1-p}(\Delta X_{g,t^*}-c_{g,t})'(\Delta X_{g,t^*}-c_{g,t})]^{-1} \\
&\quad * \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{1-D_{i,g,t}}{1-\hat{p}}(\Delta X_{i,g,t^*}-c_{i,g,t})'e_{i,g,t})
\end{aligned}
$$

Then, we can solve for C:

$$
\begin{aligned}
C &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}}\Delta X_{i,g,t}(0)\hat{\tilde{\beta}}_{g,t^*}) - E[\frac{D}{p}\Delta X_{t^*}(0)]\tilde{\beta}_{g,t^*}) \\
&= \sqrt{n}(E[\frac{D}{p}\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}] + o_p(1) - E[\frac{D}{p}\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

For D:

First, we find $\sqrt{n}\frac{1}{n}(\hat{\beta}_{g,t}-\beta_{g,t})$. Using similar arguments as before:

$$
\begin{aligned}
\hat{\beta}_{g,t} &= [(X_{g-1,t}-\Delta X_{g,t^*}(0)\hat{\tilde{\beta}}_{g,t^*}-Z\hat{\delta}_{g,t})'(X_{g-1,t}-\Delta X_{g,t^*}(0)\hat{\tilde{\beta}}_{g,t^*}-Z\hat{\delta}_{g,t})]^{-1} \\
&\quad * (X_{g-1,t}-\Delta X_{g,t^*}(0)\hat{\tilde{\beta}}_{g,t^*}-Z\hat{\delta}_{g,t})'(e-\zeta) \\
&= [(X_{g-1,t}-d)'(X_{g-1,t}-d)]^{-1}(X_{g-1,t}-d)'(e-\zeta) \\
\to E[\hat{\beta}_{g,t}] &= [(X_{g-1,t}-\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}-Z\delta_{g,t})'(X_{g-1,t}-\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}-Z\delta_{g,t})]^{-1} \\
&\quad * (X_{g-1,t}-\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}-Z\delta_{g,t})'(e) \\
&= \beta_{g,t},
\end{aligned}
$$

where e is the residuals from regressing the Z and $\Delta X_{g,t^*}(0)\tilde{\beta}_{g,t^*}$ on Y. Then, we can solve for D:

$$
\begin{aligned}
D &= \sqrt{n}((\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{p}X_{i,g-1,t}\hat{\beta}_{g,t^*}) - E[\frac{D}{p}X_{g-1,t}]\beta_{g,t^*}) \\
&= \sqrt{n}(E[\frac{D}{p}X_{g-1,t}\beta_{g,t^*}] + o_p(1) - E[\frac{D}{p}X_{g-1,t}\beta_{g,t^*}]) \\
&= o_p(1),
\end{aligned}
$$

where the second equality holds by the Weak Law of Large Numbers and the Central Mapping Theorem.

Therefore, $ATT_{g,t}$ is asymptotically normal for all $g \in G$ and $g \leq t \leq \tau$.

To estimate the aggregate Ψ, we can find the average of $\Psi_{g,t}$, weighting by group size and total time treated:

$$\Psi = \sum_{g\in G}\sum_{t=g}^{\tau} w(g,t)\Psi_{g,t} \tag{A.7}$$

where w(g,t) are weights given based on group size g and time treated t. The formula for these weights can be determined by the researcher.

Because each of the estimated ATTs for each combination of g and t are asymptotically normal, the average of them is also asymptotically normal. Thus, we can conclude that:

$$
\begin{aligned}
\sqrt{n}(\widehat{ATT}_{(g,t)} - ATT_{(g,t)}) &= \frac{1}{\sqrt{(n)}}\sum_{i=1}^{n}\Psi \\
&\to N(0, \Omega_{g,t})
\end{aligned}
$$