Partially Conditional Average Treatment Effect on Treated for Difference-in-Differences*†

Jiale Yang[1], Zongwu Cai[2], Qixian Zhong[1,†]

April 16, 2026

[1] Department of Statistics and Data Science, Xiamen University, Xiamen, Fujian, China

[2]Department of Economics, University of Kansas, Lawrence, KS 66045, USA.

Abstract

Causal inference has been widely applied across various fields, with Difference-in-Differences (DiD) emerging as one of the most popular methods for estimating the causal effect of a treatment or policy change in settings where treatment timing is staggered or occurs at a specific point in time. However, in many scenarios, not all covariates are of primary interest. This paper studies the partially conditional average treatment effect on the treated and extends to the multi-period DiD settings under the assumption of conditional parallel trends. The proposed approach is a two-stage doubly robust estimator, which allows us to both feature the heterogeneity and enhance the interpretability of treatment effects by focusing on a subset of covariates of interest. The property of double robustness ensures a consistent estimator as long as either the outcome regression model or the propensity score model is correctly specified. Beyond the proposed new methodology, we derive asymptotic theories by establishing the convergence rate and asymptotically normality of the nonparametric estimator and further examine the double robustness and practical applicability through several simulation studies. Finally, the proposed methods are used to study the impact of the Deferred Action for Childhood Arrivals program for non-citizen immigrants in the US.

Keywords: Double robustness; Heterogeneity; Kernel smoothing; Local linear regression; Multi-period DiD.

*The research was supported by the National Key R&D Program of China (grant no. 2024YFA1000095) and National Natural Science Foundation of China (grant nos. 12571293, 12271286, 71988101, 72133002, 72033008, 72595870, 72595875).

†Corresponding author. E-mail: qxzhong@xmu.edu.cn (Q. Zhong)

1 Introduction

Causal inference is a field focused on identifying causal relationships between two or more variables, rather than merely examining their correlations. Recently, it has been widely applied in various domains, including medicine, public health, social studies, and economics. The primary goal of causal inference is to evaluate the effect of a treatment or intervention on specific outcomes. For instance, in economics, assessing the effectiveness of policy interventions is of particular interest, such as evaluating the impact of increasing the minimum income on employment; see, e.g., Abadie et al. (2010) and Grimmer et al. (2017) for more discussions.

Treatment effects encompass not only the average treatment effect (ATE) and the average treatment effect on the treated (ATT), but also the conditional average treatment effect (CATE) for individuals (Hernán and Robins, 2010; Imbens and Rubin, 2015; Ding, 2024). Both ATE and ATT focus on average effects within the population, reflecting treatment effects from a global perspective. In contrast, CATE estimates the heterogeneous treatment effects at the local level, allowing for the analysis of specific subgroups of interest. However, researchers may be more interested in the treatment effects of specified characters rather than the complete set of covariates. Consequently, this paper examines the partially conditional average treatment effect on the treated (PCATT), which is a conditional ATE based on a subset of covariates. Particularly, when this subset is univariate, the estimated CATE takes the form of a curve, providing a visualization of the treatment effect. This can facilitate more accurate decision support and personalized own treatments.

The study of partially conditional treatment effect has gained increasing attention in recent years. For example, Abrevaya et al. (2015) and Zhou and Zhu (2021) employed an inverse probability weighting (IPW)-based method to estimate average treatment effects conditioned on a subset of covariates. Recently, Cai et al. (2021), Zhou et al. (2022), and Cai

et al. (2025) utilized quantile regression to examine the partially conditional quantile treatment effect (PCQTE) by doubly robust method, characterizing heterogeneity conditional on predetermined covariates. Additionally, Lee et al. (2017) and Fan et al. (2022) proposed a two-step estimation procedure to construct a doubly robust estimator or uniform confidence of the partially conditional average treatment effect. Although these approaches provide useful tools to analysis PCATE, they are not amenable to time-dependent outcome.

This paper examines the PCATT within the framework of difference-in-differences (DiD) analysis. DiD is one of the most popular methods in economics and social sciences for estimating causal effects of time-dependent outcomes. The target estimand in DiD is a version of ATT. The approach can be applied in both two-period and multiple-period settings, depending on when units are treated as addressed in Athey and Imbens (2022) and Baker et al. (2022). For the two-period DiD, it is often referred to as "canonical" DiD with data observed from two time periods: one group of units receives the treatment starting in the second period, while a comparison group does not receive the treatment at any point. In contrast, the multiple-period DiD, commonly known as multi-period DiD, involves more than two time periods and variation in treatment timing across different units.

One of the common approaches to implement DiD strategies is the two-way fixed effects (TWFE) regression method. However, this model often performs poorly due to issues with negative weights and treatment effect heterogeneity; see, for instance, (Wooldridge, 2025) and Roth et al. (2023). Later, in response to these shortcomings, several alternative methods have been proposed for staggered adoption designs, including but not limited to the papers by De Chaisemartin and d'Haultfoeuille (2020), Sun and Abraham (2021), Goodman-Bacon (2021), Borusyak et al. (2024), and references therein. For example, De Chaisemartin and d'Haultfoeuille (2020) developed an alternative DiD estimator that identifies the treatment effect for groups that switch treatment, at the time when they

switch. Their setup is more adaptable and is able to accommodate more general designs. Additional methods under the DiD framework for studying ATT have also emerged, such as the double/debiased machine learning method as in Chang (2020) and matching methods as in Imai et al. (2023). However, the aforementioned approaches are often sensitive to the specification for nuisance parameters or functions, particularly the propensity score function. Recently, doubly robust estimators, which can alleviate the sensitivity of the nuisance functions, have gained increasing attention. For instance, Sant'Anna and Zhao (2020) and Callaway and Sant'Anna (2021) proposed doubly robust estimators for ATT in staggered DiD designs. They decompose multiple-period setup into a series of two-period DiD cases by defining groups based on the timing of first treatment, which effectively avoid the limitations of TWFE. Despite their improvement in TWFE model, the current approaches often lack interpretability regarding the underlying heterogeneity of a specific variable. Our estimator complements these methods as we focus on doubly robust estimation and inference for PCATT conditional on a subset of the covariates.

Under this framework, our work departs from the current approach of Callaway and Sant'Anna (2021) in the following two aspects. We propose a two-stage nonparametric estimation procedure. In the first stage, flexible machine learning methods are used to estimate nuisance functions. In the second stage, the local linear regression is employed to smooth the pseudo-outcome on a subset of covariates. Second, our primary focus is on estimating PCATT rather than ATT, allowing us to capture the heterogeneity of treatment effects across specific subsets of covariates and visualize the results as curves rather than single-point estimates. Taking our real data application as a motivated example, we study the treatment effect of the Deferred Action for Childhood Arrivals (DACA) policy, with an interest in studying whether this policy has a heterogeneous effect on the income across different age groups; see, Figure 3 in Section 5 for details. In this case, studying the

PCATT provides more interpretable insights into the role of specific covariates.

To summarize, the major contributions of this paper are three-fold.

1. We propose a two-stage estimation for PCATT under multi-period adoption designs. The estimator exhibits a doubly robust property, allowing either the propensity score model or the outcome regression model to be misspecified, which contributes to the literature on doubly robust estimation and DiD.

2. We derive the theoretical properties of the proposed PCATT estimators, demonstrating that they converge at an optimal nonparametric rate, provided that the nuisance functions converge at moderate rates. Additionally, we establish the asymptotic normality of the estimators under various DiD settings and data types, which further facilitates inference.

3. While ATE and ATT are widely used to measure the effects of policy interventions, they often fail to capture the heterogeneity of treatment effects across specific covariates of interest. In contrast, we introduce PCATT to enhance interpretability by displaying the results in a graph. Moreover, PCATT provides more precise intervention strategies and supports personalized decision-making.

The rest of the paper proceeds as follows. Section 2 introduce the identification and estimation of PCATT in canonical and multi-period DiD settings. Section 3 discusses the asymptotic properties of the estimators. Section 4 provides simulation studies to verify the double robustness of the method and Section 5 provides data applications on estimating the effect of Deferred Action for Childhood Arrivals program on income and weekly working hours. Section 6 concludes the paper. The detailed mathematical proofs are presented in Supplementary Materials.

2 Methodology

In this section, we introduce the estimation procedure of PCATT for both canonical and multi-period DiD cases.

2.1 Canonical DiD

We first consider the canonical DiD setup where there are two treatment periods and two treatment groups. For unit i, let Y_i^0 and Y_i^1 be the outcomes of interest in a pre-treatment period $t = 0$ and in a post-treatment period $t = 1$, respectively. Let D_{it} be a binary variable being one if unit i is exposed to the treatment in period t and zero otherwise. We assume that units are only exposed to treatment after period 0, i.e., $D_{i0} = 0$ for all i. We then denote $D_i = D_{i1}$ throughout the paper. Under Neyman-Rubin potential outcome framework (Neyman, 1923; Rubin, 1974), we denote $Y_i^t(0)$ and $Y_i^t(1)$ the control and treated potential outcomes of unit i in period t and suppose the observed outcome $Y_i^t = Y_i^t(D_{it})$. That is $Y_i^0 = Y_i^0(0)$ and $Y_i^1 = D_i Y_i^1(1) + (1 - D_i)Y_i^1(0)$. In addition, we also observe pre-treatment covariates $X_i \in \mathbb{R}^p$.

Let Z be a subset of X. We are of interest in estimating the partially conditional average treatment effect on treated, which is defined as the average treatment effect conditional on subset of covariates $Z = z$ in post-treatment period $t = 1$ on treated:

$$\tau(z) = \mathbb{E}[Y^1(1) - Y^1(0)|Z = z, D = 1].$$

When Z is a constant, the PCATT reduces to the ATT (Abadie, 2005; Chang, 2020; Sant'Anna and Zhao, 2020), while if $Z = X$ it becomes the CATT (Henderson et al., 2023). Since the potential outcomes $Y_{i1}(1)$ and $Y_{i1}(0)$ are not observed simultaneously, the PCATT $\tau(z)$ cannot be identified from observations without any restriction. To overcome it, we require the following standard assumptions (Assumptions 1 – 3) for DiD models (Heckman et al., 1997, 1998; Blundell et al., 2004; Abadie, 2005).

In this paper, we study either panel or repeated cross-section data. Following Abadie (2005) and Sant'Anna and Zhao (2020), we assume the observed data satisfies assumption for the DiD setup below. Here, we denote Y^0 and Y^1 as the observed outcome in pre-treatment and post-treatment period, respectively.

Assumption 1. *(a) When panel data (PD) are available, assume that* $\{(X_i, D_i, Y_i^0, Y_i^1) : i = 1, \dots, N\}$ *are independent and identically distributed (i.i.d.) copies of* (X, D, Y^0, Y^1)*; (b) When repeated cross-section data (RCD) are available, denote* $Y_i = T_i Y_i^1 + (1 - T_i) Y_i^0$*, where* T_i *is a binary variable that takes value zero if unit* i *is only observed in the pre-treatment period, and value one otherwise. Assume that covariates and treatment status are stationary and the pooled data* $\{(X_i, D_i, Y_i, T_i) : i = 1, \dots, N\}$ *independently draw from the distribution*

$$F(y, d, t, x) = \lambda t F_{Y^1, D, X|T}(y, d, x|T = 1) + (1 - \lambda)(1 - t) F_{Y^0, D, X|T}(y, d, x|T = 0),$$

where $\lambda = P(T = 1) \in (0, 1)$.

Assumption 1(a) supposes that each unit is independently observed with its outcomes in pre-treatment and post-treatment period being jointly available. Assumption 1(b) pertains to the scenario of repeated cross-section data in which the data are i.i.d. from the distribution of (X, D, Y^t) conditional on $T = t$ for $t = 0$, 1, respectively. The stationary condition in Assumption 1(b) implies that the joint distribution of (X, D) does not vary with T, i.e., (X, D) is independent of T. This is a common assumption in most existing literature as in Abadie (2005) and Chang (2020); Sant'Anna and Zhao (2020).

Assumption 2. $\mathbb{E}[Y^1(0) - Y^0(0)|D = 0, X] = \mathbb{E}[Y^1(0) - Y^0(0)|D = 1, X]$ *holds almost surely.*

The parallel trends assumption requires that, conditional on covariates $X = x$, the average trend of the control potential outcome would have been comparable between the treatment and control groups. It is an essential assumption for establishing a treatment effect in DiD

models. Without this assumption, it is difficult to attribute the changes in the outcome variable solely to the treatment effect, as differences in trend between the treatment and control groups may confound the treatment effect.

Assumption 3. *There exists a constant* $\delta_1 \in (0, 1/2)$*, such that the propensity scores* $\pi(x) = \mathbb{E}[D \mid X = x]$ *satisfy* $\pi(X) \in [\delta_1, 1 - \delta_1]$ *almost surely.*

Assumption 3 requires all units have a non-zero probability of treatment assignment, guaranteeing valid comparisons and reliable estimation across covariate strata. Without this assumption, it is difficult to attribute the changes in the outcome variable solely to the treatment effect, as differences in trend between the treatment and control groups may confound the treatment effect. Under the aforementioned assumptions, we are able to identify PCATT $\tau(z)$ as the following proposition.

Proposition 1. *Suppose that Assumptions 1 – 3 hold.*

(a) When panel data are available, we have

$$\tau(z) = \frac{1}{e(z)}\mathbb{E}\Big[\frac{D - \pi(X)}{1 - \pi(X)}\{\Delta Y - \mu^p(X)\}\Big|Z = z\Big], \tag{1}$$

where $e(z) = \mathbb{E}[D|Z = z]$*,* $\Delta Y = Y^1 - Y^0$ *and* $\mu^p(x) = \mathbb{E}[\Delta Y|D = 0, X = x]$*;*

(b) When repeated cross-section data are available, we have

$$\tau(z) = \frac{1}{e(z)}\mathbb{E}\Big[\frac{(T - \lambda)(D - \pi(X))}{\lambda(1 - \lambda)(1 - \pi(X))}\{Y - \mu^{rc}(T, X)\}\Big|Z = z\Big], \tag{2}$$

where $\mu^{rc}(T, X) = T\mathbb{E}[Y|D = 0, T = 1, X] + (1 - T)\mathbb{E}[Y|D = 0, T = 0, X]$*.*

Note that nuisance functions in (1) and (2), such as propensity score $\pi(x)$ and regression functions $\mu^p(x)$ and $\mu^{rc}(t, x)$ are estimable from the observed data, making $\tau(z)$ identifiable. Additionally, the proposed methods possess double robustness properties (see Proposition 2 below) that enable us to estimate correctly $\tau(z)$, even if either the propensity score $\pi(x)$ or the regression functions $\mu^p(x)$ and $\mu^{rc}(t, x)$ is incorrectly specified. Denote

$$\omega^p(D, X; h_0) = \frac{D - h_0(x)}{1 - h_0(x)}, \qquad H^p(Y^0, Y^1, X; \mu) = Y^1 - Y^0 - \mu(X),$$

$$\omega^{rc}(D,T,X;h_0)=\frac{(T-\lambda)(D-h_0(x))}{\lambda(1-\lambda)(1-h_0(x))} \quad \text{and} \quad H^{rc}(Y,T,X;\mu)=Y-\mu(T,X).$$

Then, we have the following proposition.

Proposition 2. *Suppose that Assumptions 1 – 3 hold.*

(a) When panel data are available, we have either

$$\tau(z)=\frac{1}{e(z)}\mathbb{E}[\omega^{p}(D,X;\pi)H^{p}(Y^0,Y^1,X;\mu)|Z=z] \text{ for any } \mu\in\mathcal{L}^2(P_X),$$

where $\mathcal{L}^2(P_X)=\{f(\cdot):\mathcal{X}\to\mathbb{R}:\mathbb{E}[f(X)]^2<\infty\}$*, or*

$$\tau(z)=\frac{1}{e(z)}\mathbb{E}[\omega^{p}(D,X;h_0)H^{p}(Y^0,Y^1,X;\mu^p)|Z=z] \text{ for any } h_0\in\mathcal{L}^2(P_X).$$

(b) When repeated cross-section data are available, we have either

$$\tau(z)=\frac{1}{e(z)}\mathbb{E}[\omega^{rc}(D,T,X;\pi)H^{rc}(Y,T,X;\mu)|Z=z] \text{ for any } \mu\in\mathcal{L}^2(P_{(T,X)}),$$

where $\mathcal{L}^2(P_{(T,X)})=\{f(\cdot):\mathcal{T}\times\mathcal{X}\to\mathbb{R}:\mathbb{E}[f(T,X)^2]<\infty\}$*, or*

$$\tau(z)=\frac{1}{e(z)}\mathbb{E}[\omega^{rc}(D,T,X;h_0)H^{rc}(Y,T,X;\mu^{rc})|Z=z] \text{ for any } h_0\in\mathcal{L}^2(P_X).$$

Propositions 1 and 2 suggest that under Assumptions 1 – 3, we are able to estimate $\tau(z)$ from the observed data. In this paper, we mainly focus on the estimation of the case that Z is univariate variable. To limit a potential overfitting bias that may result when estimating nuisance parameters/functions, we use the cross-fitting method as in Chernozhukov et al. (2018) to obtain the estimator of $\tau(z)$.

We randomly divide the sample $\{1,\dots,N\}$ into K disjoint subsets of approximately equal size, where K is pre-specified, which is often set to 5 or 10, and without loss of generality we assume that $n=N/K$ is an integer. For each $1\le k\le K$, we denote the k-th subset and its complement as $\mathcal{I}_k$ and $\mathcal{I}_k^c$, respectively. The subset $\mathcal{I}_k$ is used to construct the estimating equation by plugging in the nuisance parameter estimates obtained from the complement $\mathcal{I}_k^c$.

Specifically, one can use the data in $\mathcal{I}_k^c$ to estimate the propensity scores $\pi(x)$ and $e(z)$ by any classification methods that produce probability estimates, such as logistic regression or neural networks, and the regression function $\mu^p(x)$ or $\mu^{rc}(t,x)$ by methods like linear regression or neural networks. We respectively denote their estimators as $\hat{\pi}_k(x)$, $\hat{e}_k(z)$, $\hat{\mu}_k^p(x)$, and $\hat{\mu}_k^{rc}(t,x)$ for $1 \leq k \leq K$. For $i \in \mathcal{I}_k$, we set pseudo-outcome

$$W_i = \frac{D_i - \hat{\pi}_k(X_i)}{\hat{e}_k(Z_i)(1-\hat{\pi}_k(X_i))}\{Y_i^1 - Y_i^0 - \hat{\mu}_k^p(X_i)\},$$

when panel data $\{(X_i, D_i, Y_i^0, Y_i^1) : i = 1, \ldots, N\}$ are available, or

$$W_i = \frac{(T_i - \hat{\lambda}_k)(D_i - \hat{\pi}_k(X_i))}{\hat{\lambda}_k(1-\hat{\lambda}_k)(1-\hat{\pi}_k(X_i))\hat{e}_k(Z_i)}\{Y_i - \hat{\mu}_k^{rc}(T_i, X_i)\} \text{ with } \hat{\lambda}_k = \sum_{i \in \mathcal{I}_k^c} T_i/n,$$

when repeated cross-section data $\{(X_i, D_i, Y_i, T_i) : i = 1, \ldots, N\}$ are available. Now, we use the local linear regression to estimate $\tau(z)$ by $\hat{\tau}(z) = \hat{\alpha}_0$ that satisfies the following optimization problem (locally weighted least squares):

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \underset{\alpha_0, \alpha_1}{\arg\min} \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} K_h(Z_i - z)(W_i - \alpha_0 - \alpha_1(Z_i - z))^2, \tag{3}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ and h being a symmetric density function and bandwidth, respectively.

2.2 Multi-period DiD

Now, we turn to studying the multi-period DiD case with multiple time periods $\tilde{T}$ and variation in treatment timing. Denote a particular time period by t, where $t = 1, \ldots, \tilde{T}$, and let D_{it} be a binary variable being one, if unit i is exposed to the treatment in period t. For treated units, denote G as the time period that becomes treated and denote G_g as a binary variable being one, if a unit is first treated in period g. For units that belong to the "never-treated" group, denote by $G_i = \infty$ or a binary indicator $C_i = I\{G_i = \infty\}$. Let $\bar{g} = \max_{i=1,\ldots,n} G_i$ be the maximum G in the dataset and $\mathcal{G} = \text{supp}(G) \backslash \bar{g} \subseteq \{2, 3, \ldots, \tilde{T}\}$

denote the support of G excluding $\bar{g}$. For a generic $\delta \geq 0$, let $\mathcal{G}_\delta = \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \tilde{T}\}$. This setup follows the potential outcomes framework for staggered adoption discussed by Callaway and Sant'Anna (2021).

Next, we denote the nuisance parameters needed for identification. We define the generalized propensity score $p_{g,s}(X) = P(G_g = 1|X, G_g + (1-D_s)(1-G_g) = 1)$ as the probability of being treated if a unit belongs to group g (in this case, $G_g = 1$) or "not-yet-treated" group by time s (in this case, $(1-D_s)(1-G_g) = 1$). Define $p_g(X) = p_{g,\tilde{T}}(X) = P(G_g = 1|X, G_g + C = 1)$ as the probability of being treated if a unit belongs to group g or "never-treated" group. In addition, we also observe pre-treatment covariates $X \in R^p$ and denote $\mathcal{X} = \text{supp}(X) \subseteq R^p$ as the support of X. The observed outcome Y_{it} is generated from potential outcomes $Y_{it}(0)$ for never-treated units and $Y_{it}(g)$ for units first treated in period g, where $g \geq 2$,. Then, the observed outcome is defined as:

$$Y_{i,t} = Y_{it}(0) + \sum_{g=2}^{T} (Y_{it}(g) - Y_{it}(0)) \cdot G_{i,g}.$$

Of interest is to estimate the group-time partially conditional average treatment effect on the treated, which is a generalization of the PCATT with multiple treatment groups and multiple time periods. It is defined as the expected treatment effect of group g in period t conditional on subset of covariates $Z = z$ on treated:

$$\tau(z, g, t) = E[Y_t(g) - Y_t(0)|Z = z, G_g = 1].$$

Following Callaway and Sant'Anna (2021), we assume the observed data satisfies assumption below for multi-period DiD. In addition, we denote $Y_1, \ldots, Y_T$ as the observed outcome in multiple periods.

Assumption 4. *(a) When panel data are available, assume that* $\{(Y_{i1}, Y_{i2}, \ldots, Y_{i\tilde{T}}, X_i, D_{i1}, D_{i2}, \ldots, D_{i\tilde{T}}) : i = 1, \ldots, N\}$ *are i.i.d. copies of* $(Y_1, \ldots, Y_{\tilde{T}}, X, D_1, \ldots, D_{\tilde{T}})$*;*

(b) When repeated cross-section data are available, conditional of $\tilde{T} = t$*, the data are*

independent and identically distributed from the distribution of $(Y_t, G_2, \ldots, G_{\tilde{T}}, C, X)$*, for all* $t = 1, \ldots, \tilde{T}$*, with* $(G_2, \ldots, G_{\tilde{T}}, C, X)$ *being invariant to* $\tilde{T}$*. The sample are independently draw from the mixture distribution*

$$F(y, g_2, \ldots, g_{\tilde{T}}, c, t, x) = \sum_{t=1}^{\tilde{T}} \lambda_t \cdot F_{Y, G_2, \ldots, G_{\tilde{T}}, C, X|\tilde{T}}(y, g_2, \ldots g_{\tilde{T}}, c, x|t),$$

where $\lambda_t = P(\tilde{T} = t) \in (0, 1)$*.*

Assumption 5. *(a) (Irreversibility of Treatment).* $D_1 = 0$ *a.s.. For* $t = 2, \ldots, \tilde{T}$*,* $D_{t-1} = 1$ *implies that* $D_t = 1$ *a.s.*

(b) (Limited Treatment Anticipation). There is a known $\delta \geq 0$ *such that for all* $g \in G, t \in \{1, \ldots, \tilde{T}\}$ *and* $t < g - \delta$*,*

$$E[Y_t(g)|X, G_g = 1] = E[Y_t(0)|X, G_g = 1] \text{ a.s..}$$

Assumption 5(a) illustrates two issues. First, no one is treated at time $t = 1$. Second, once a unit becomes treated, that unit will remain treated in the following period. When $\delta = 0$, Assumption 5 ensures that individuals will not be affected by a policy before it is implemented, i.e., no-anticipation assumption. When $\delta > 0$, this assumption implies that the pre-treatment periods becomes $t \leq g - \delta$ rather than $t \leq g$. (Malani and Reif, 2015; Callaway and Sant'Anna, 2021).

Assumption 6. *(Conditional Parallel Trends Assumption). Let* δ *be as defined in Assumption 5(b). (a) ("Never-Treated" Group) For each* $g \in G$ *and* $t \in \{2, \ldots, \tilde{T}\}$ *such that* $t \geq g - \delta$*,*

$$E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, C = 1] \text{ holds almost surely.}$$

(b) ("Not-yet-Treated" Group) For each $g \in G$ *and each* $(s, t) \in \{2, \ldots, \tilde{T}\} \times \{2, \ldots, \tilde{T}\}$ *such that* $t \geq g - \delta$ *and* $t + \delta \leq s < \bar{g}$*,*

$$E[Y_t(0) - Y_{t-1}(0)|X, G_g = 1] = E[Y_t(0) - Y_{t-1}(0)|X, D_s = 1, G_g = 0] \text{ holds almost surely.}$$

Assumption 6 is the parallel trends assumption of the multiple time periods (Callaway and Sant'Anna, 2021). When we use "never-treated" group as the control group, the number of observations in this group needs to be comparable to the treated group. If the difference between the numbers of the two groups is large, we can use the "not-yet-treated" group as the control group. It implies that in the post-treatment periods $t \geq g - \delta$, within the covariate strata $X = x$, the potential outcome of the control group on average will move in a similar direction with similar magnitudes for the treatment group g.

Assumption 7. *(Overlap). For each $t \in \{2, \ldots, \tilde{T}\}, g \in G$, there exist some $\epsilon > 0$ such that $P(G_g = 1) > \epsilon$ and $p_{g,t}(X) < 1 - \epsilon$ almost surely.*

Under the aforementioned assumptions, we can reorganize the pre-treatment and post-treatment periods based on g and t by choosing "never-treated" or "not-yet-treated" groups as the comparison group. This allows us to estimate the PCATT in multi-period DiD setup.

Proposition 3. *Suppose that Assumptions 4-7 hold.For each g and t such that $g \in G_\delta, t \in \{2, \ldots, \tilde{T} - \delta\}$ and $t \geq g - \delta$,*

(a)When panel data are available,

(i) For "never-treated" group, we have

$$\tau(z, g, t; \delta) = \frac{1}{e_g(z)} \mathbb{E}\Big[\{G_g - \frac{p_g(X)C}{1 - p_g(X)}\}\{\Delta Y - \mu^p_{nev}(X)\}\Big| Z = z\Big], \tag{4}$$

where $e_g(z) = \mathbb{E}[G_g|Z = z], \Delta Y = Y_t - Y_{g-\delta-1}$ and $\mu^p_{nev}(X) = \mathbb{E}[\Delta Y|C = 1, X]$;

(ii) For "not-yet-treated" group, we have

$$\tau(z, g, t; \delta) = \frac{1}{e_g(z)} \mathbb{E}\Big[\{G_g - \frac{p_{g,t+\delta}(X)(1 - G_g)(1 - D_{t+\delta})}{1 - p_{g,t+\delta}(X)}\}\{\Delta Y - \mu^p_{nyt}(X)\}\Big| Z = z\Big], \tag{5}$$

where $\mu^p_{nyt}(X) = \mathbb{E}[\Delta Y|G_g = 0, D_{t+\delta} = 0, X]$.

(b) When repeated cross-section data are available,

(i) For "never-treated" group, we have

$$\tau(z, g, t; \delta) = \frac{1}{e_g(z)} \mathbb{E}\Big[\{G_g - \frac{p_g(X)C}{1 - p_g(X)}\}\frac{\tilde{T}_t - \lambda_t}{\lambda_t(1 - \lambda_t)}\{Y - \mu^{rc}_{nev}(\tilde{T}, X)\}\Big| Z = z\Big], \tag{6}$$

where $\tilde{T}_t$ is a binary variable that takes value one if a unit is only observed at period t, and value zero otherwise. $\lambda_t = P(\tilde{T}_t = 1|G_g + C = 1), e_g(z) = \mathbb{E}[G_g|Z = z, G_g + C = 1]$ *and* $\mu^{rc}_{nyt}(\tilde{T}, X) = \tilde{T}_t\mathbb{E}[Y|C = 1, \tilde{T}_t = 1, X] + (1 - \tilde{T}_t)\mathbb{E}[Y|C = 1, \tilde{T}_t = 0, X]$;

(ii) For "not-yet-treated" group, we have

$$\tau(z, g, t; \delta) = \frac{1}{e_g(z)}\mathbb{E}\Big[\{G_g - \frac{p_{g,t+\delta}(X)(1 - G_g)(1 - D_{t+\delta})}{1 - p_{g,t+\delta}(X)}\}\frac{\tilde{T}_t - \lambda_t}{\lambda_t(1 - \lambda_t)}\{Y - \mu^{rc}_{nyt}(\tilde{T}_t, X)\}\Big|Z = z\Big], \tag{7}$$

where $\lambda_t = P(\tilde{T}_t = 1|G_g + (1 - G_g)(1 - D_{t+\delta}) = 1), e_g(z) = \mathbb{E}[G_g|Z = z, G_g + (1 - G_g)(1 - D_{t+\delta}) = 1]$ *and* $\mu^{rc}_{nyt}(\tilde{T}_t, X) = \tilde{T}_t\mathbb{E}[Y|G_g = 0, D_{t+\delta} = 0, \tilde{T}_t = 1, X] + (1 - \tilde{T}_t)\mathbb{E}[Y|G_g = 0, D_{t+\delta} = 0, \tilde{T}_t = 0, X]$.

Similarly, we need to estimate the nuisance parameters/functions in (4) – (7), such as propensity score $p_g(x)$ and regression functions $\mu^p_{nev}(x)$, $\mu^p_{nyt}(t, x)$, $\mu^{rc}_{nev}(x)$, $\mu^{rc}_{nyt}(t, x)$ to make $\tau(z, g, t)$ identifiable. Additionally, the proposed methods possess double robustness properties (see Proposition 2) that enable us to estimate correctly $\tau(z, g, t)$, even if either the propensity score $p_g(x)$ or the regression functions $\mu_{nev}(x)$, $\mu_{nyt}(t, x)$ is misspecified. To this end, denote

$$\omega^p_{nev}(D, X; h_0) = G_g - \frac{h_0(X)C}{1 - h_0(X)}, \quad \omega^{rc}_{nev}(D, X; h_0) = \left[G_g - \frac{h_0(X)C}{1 - h_0(X)}\right]\frac{\tilde{T}_t - \lambda_t}{\lambda_t(1 - \lambda_t)},$$

$$\omega^p_{nyt}(G_g, D_{t+\delta}, X; h_0) = G_g - \frac{h_0(X)(1 - G_g)(1 - D_{t+\delta})}{1 - h_0(X)},$$

$$\omega^{rc}_{nyt}(G_g, D_{t+\delta}, X; h_0) = \left[G_g - \frac{h_0(X)(1 - G_g)(1 - D_{t+\delta})}{1 - h_0(X)}\right]\frac{\tilde{T}_t - \lambda_t}{\lambda_t(1 - \lambda_t)},$$

$H^p_{nyt}(Y_t, Y_{g-\delta-1}, X; \mu) = Y_t - Y_{g-\delta-1} - \mu(X)$, $H^{rc}_{nev}(Y, \tilde{T}, X; \mu) = Y - \mu(\tilde{T}, X)$, $H^p_{nev}(Y_t, Y_{g-\delta-1}, X; \mu) = Y_t - Y_{g-\delta-1} - \mu(X)$, and $H^{rc}_{nyt}(Y, \tilde{T}, X; \mu) = Y - \mu(\tilde{T}, X)$.

Proposition 4. *Suppose that Assumptions 5-7 hold. For each g and t such that $g \in G_\delta, t \in \{2, \ldots, \tilde{T} - \delta\}$ and $t \geq g - \delta$,*

(a) When panel data are available. For "never-treated" group, we have either

$$\tau(z, g, t; \delta) = \frac{1}{e(z)}\mathbb{E}[\omega^p_{nev}(D, X; p_g)H^p_{nev}(Y_t, Y_{g-\delta-1}, X; \mu)|Z = z], \text{ for any } \mu \in \mathcal{L}^2(P_X),$$

or

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^p_{nyt}(D,X;h_0)H^p(Y_t,Y_{g-\delta-1},X;\mu^p_{nev})|Z=z] \text{ for any } h_0 \in \mathcal{L}^2(P_X).$$

For "not-yet-treated" group, we have either

$$\tau(z,g,t;\delta) = \frac{1}{e(z)}\mathbb{E}[\omega^p_{nev}(D,X;p_g)H^p_{nev}(Y_t,Y_{g-\delta-1},X;\mu)|Z=z] \text{ for any } \mu \in \mathcal{L}^2(P_X),$$

or

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^p_{nyt}(G_g,D_{t+\delta},X;h_0)H^p_{nyt}(Y_t,Y_{g-\delta-1},X;\mu^p_{nyt})|Z=z] \text{ for any } h_0 \in \mathcal{L}^2(P_X).$$

(b) When repeated cross-section data are available. For "never-treated" group, we have either

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^{rc}_{nyt}(G_g,D_{t+\delta},X;p_{g,t+\delta})H^{rc}_{nyt}(Y,\tilde{T}_t,X;\mu)|Z=z] \text{ for any } \mu \in \mathcal{L}^2(P_X),$$

or

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^{rc}_{nyt}(D,X;h_0)H^{rc}_{nyt}(Y,\tilde{T}_t,X;\mu^{rc}_{nev})|Z=z] \text{ for any } h_0 \in \mathcal{L}^2(P_X).$$

For "not-yet-treated" group, we have either

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^{rc}_{nyt}(G_g,D_{t+\delta},X;p_{g,t+\delta})H^{rc}_{nyt}(Y,\tilde{T}_t,X;\mu)|Z=z] \text{ for any } \mu \in \mathcal{L}^2(P_X),$$

or

$$\tau(z,g,t;\delta) = \frac{1}{e_g(z)}\mathbb{E}[\omega^{rc}_{nyt}(G_g,D_{t+\delta},X;h_0)H^{rc}_{nyt}(Y,\tilde{T}_t,X;\mu^{rc}_{nyt})|Z=z] \text{ for any } h_0 \in \mathcal{L}^2(P_X).$$

Proposition 3 suggests that under Assumptions 4 – 7, we are able to estimate $\tau(z,g,t)$ from the observed data. Similar to the canonical DiD approach, we use the cross-fitting method to the get estimator of $\tau(z,g,t,\delta)$ as in Chernozhukov et al. (2018). Then, we can use the data in $\mathcal{J}^c_k$ to estimate the propensity scores and the regression functions. We respectively

denote their estimators as $\hat{p}_{g,k}(x), \hat{e}_{g,k}(z), \hat{\mu}^{p}_{nev,k}(x), \hat{\mu}^{p}_{nyt,k}(x), \hat{\mu}^{rc}_{nev,k}(x), \hat{\mu}^{rc}_{nyt,k}(x)$ for $1 \leq k \leq K$. For $i \in \mathcal{I}_k$, when panel data are available, we set pseudo-outcome

$$W_{i,g,t;\delta} = \frac{1}{\hat{e}_{g,k}(Z_i)}\{G_{i,g} - \frac{\hat{p}_{g,k}(X_i)C_i}{1-\hat{p}_{g,k}(X_i)}\}\{Y_{i,t} - Y_{i,g-\delta-1} - \hat{\mu}^{p}_{nev,k}(X_i)\},$$

when we use "never-treated" group as the control group, or

$$W_{i,g,t;\delta} = \frac{1}{\hat{e}_{g,k}(Z_i)}\{G_{i,g} - \frac{\hat{p}_{g,t+\delta,k}(X_i)(1-G_{i,g})(1-D_{i,t+\delta})}{1-\hat{p}_{g,t+\delta,k}(X_i)}\}\{Y_{i,t} - Y_{i,g-\delta-1} - \hat{\mu}^{p}_{nyt,k}(X_i)\},$$

when we use "not-yet-treated" group as the control group.

For $i \in \mathcal{I}_k$, when repeated cross-section data are available, we set pseudo-outcome

$$W_{i,g,t;\delta} = \frac{1}{\hat{e}_{g,k}(Z_i)}\{G_{i,g} - \frac{\hat{p_{g,k}}(X_i)C_i}{1-\hat{p}_{g,k}(X_i)}\}\frac{\tilde{T}_{i,t} - \hat{\lambda}_{t,k}}{\hat{\lambda}_{t,k}(1-\hat{\lambda}_{t,k})}\{Y_i - \hat{\mu}^{rc}_{nev,k}(\tilde{T}_{i,t}, X_i)\} \text{ with } \hat{\lambda}_{t,k} = \sum_{i \in \mathcal{I}_k^c} \frac{I\{\tilde{T}_{i,t}=1\}}{n_k},$$

when we use "never-treated" group as the control group, or

$$W_{i,g,t;\delta} = \frac{1}{\hat{e}_{g,k}(Z_i)}\{G_{i,g} - \frac{\hat{p}_{g,t+\delta,k}(X_i)(1-G_{i,g})(1-D_{i,t+\delta})}{1-\hat{p}_{g,t+\delta,k}(X_i)}\}\frac{\tilde{T}_{i,t} - \hat{\lambda}_{t,k}}{\hat{\lambda}_{t,k}(1-\hat{\lambda}_{t,k})}\{Y_i - \hat{\mu}^{rc}_{nyt,k}(\tilde{T}_t, X_i)\},$$

when we use "not-yet-treated" group as the control group. For group g in period t, similar to (3), we use the local linear regression to estimate $\tau(z,g,t)$ by $\hat{\tau}(z,g,t;\delta) = \hat{\alpha}_0$ that satisfies the following optimization problem:

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \arg\min_{\alpha_0,\alpha_1} \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in \mathcal{I}_k} K_h(Z_i - z)(W_{i,g,t;\delta} - \alpha_0 - \alpha_1(Z_i - z))^2,$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ and h being a symmetric density function and bandwidth, respectively.

3 Theoretical Properties

In this section, we investigate the asymptotic properties of the proposed estimators. Here, we only consider when Z is a one-dimensional variable, which is the common scenario in practical applications. For this case, the estimation results can be described by curve. Detailed proofs can be found in the Supplementary Material. Now, some necessary assumptions are needed, listed below.

Assumption 8. *Kernel function $K(\cdot)$ is a symmetric probability density function on $[-1,1]$ and*

$$\sigma_K^2 = \int [K(u)]^2 du < \infty, \int u^2 K(u) du < \infty$$

and there exists $0 < L < \infty$ such that

$$|K(u) - K(v)| \leq L|u - v|, \text{ for any } u, v \in [-1, 1].$$

Assumption 9. *The probability density function $p(\cdot)$ for continuous Z satisfies*

$$0 < C_1 < p(z) < C_2 < \infty,$$

and the second derivative $p^{(2)}(\cdot)$ is bounded.

Assumption 8 is a standard assumption for kernel functions, which can be satisfied by commonly used functions such as the Epanechnikov kernel. Assumption 9 imposes boundedness restrictions on density of covariate Z and the condition of the second derivative ensures the smoothness of the continuous distribution.

Assumption 10. *(a) There exists a constant $0 < C < \infty$, such that when PD is available, $\mathbb{E}[(Y^1 - Y^0 - \mu^p(X))^2|X] < C$, a.s. P_X or when RCD is available $\mathbb{E}[(Y - \mu^{rc}(T,X))^2|T,X] < C$, a.s. $P_{(T,X)}$.*

(b) There exists a constant $0 < C < \infty$, such that when PD is available, $\mathbb{E}[(Y_t - Y_{g-\delta-1} - \mu^p_{nev}(X))^2|X] < C$ or $\mathbb{E}[(Y_t - Y_{g-\delta-1} - \mu^p_{nyt}(X))^2|X] < C$, a.s. P_X. When RCD is available $\mathbb{E}[(Y - \mu^{rc}_{nev}(\tilde{T}, X))^2|T,X] < C$ or $\mathbb{E}[(Y - \mu^{rc}_{nyt}(T,X))^2|T,X] < C$, a.s. $P_{(T,X)}$.

Assumption 11. *(a) For canonical DiD, the nuisance estimators $\hat{e}(z), \hat{\pi}(x), \hat{\mu}^p(x), \hat{\mu}^{rc}(t,x)$ are bounded almost surely and they satisfy: $\sup_z \mathbb{E}\{|\hat{e}(z) - e(z)|^2\} = o_p(N^{-4/5})$, $\sup_x \mathbb{E}\{|\hat{\pi}(x) - \pi(x)|^2\}$, $\sup_x \mathbb{E}\{|\hat{\mu}^p(x) - \mu^p(x)|^2\}$, and $\sup_{t,x} \mathbb{E}\{|\hat{\mu}^{rc}(t,x) - \mu^{rc}(t,x)|^2\} = o_p(N^{-2/5})$.*

(b) For multi-period DiD, the nuisance estimators $\hat{\mu}^p_{nev}(x)$, $\hat{\mu}^p_{nyt}(x)$,$\hat{\mu}^{rc}_{nev}(t,x)$, $\hat{\mu}^{rc}_{nyt}(t,x)$, $\hat{e}_g(z)$, and $\hat{p}_g(x)$, $\hat{p}_{g,t+\delta}(x)$ are bounded almost surely and satisfy:

$\sup_x \mathbb{E}\{|\hat{p}_{g,k}(x) - p_g(x)|^2\}$, $\sup_x \mathbb{E}\{|\hat{\mu}^p_{nev,k}(x) - \mu^p_{nev,k}(x)|^2\}$, $\sup_{t,x} \mathbb{E}\{|\hat{\mu}^{rc}_{nev,k}(t,x) - \mu^{rc}_{nev,k}(t,x)|^2\}$, $\sup_x \mathbb{E}\{|\hat{p}_{g,t+\delta,k}(x) - p_{g,t+\delta}(x)|^2\}$, $\sup_x \mathbb{E}\{|\hat{\mu}^p_{nyt,k}(x) - \mu^p_{nyt,k}(x)|^2\}$, *and* $\sup_{t,x} \mathbb{E}\{|\hat{\mu}^{rc}_{nyt,k}(t,x) - \mu^{rc}_{nyt,k}(t,x)|^2\}$ *are of the order of* $o_p(N^{-2/5})$, *and* $\sup_x \mathbb{E}\{|\hat{e}_{g,k}(z) - e_g(z)|^2\} = o_p(N^{-4/5})$,

Assumption 10 is usually used for the estimation of standard errors. These assumptions impose restrictions on the second conditional moment functions, which are analogous to those in Hirano et al. (2003), Abrevaya et al. (2015), and Cai et al. (2021), commonly used in the literature of treatment effect and nonparametric estimation. Assumption 11 imposes conditions on the convergence rate of estimators of nuisance functionals.

Theorem 1. *Let* $\tilde{W} = (D-\pi(X))(Y^1 - Y^0 - \mu^p(X))/\{e(Z)(1-\pi(X))\}$ *when panel data is available or* $\tilde{W} = (T-\lambda)(D-\pi(X))(Y-\mu^{rc}(X))/\{\lambda(1-\lambda)e(Z)(1-\pi(X))\}$ *when repeated cross-section data is available. Under Assumptions 1–3, 8, 9 and 10(i), 11(i), if* $h = O(N^{-1/5})$, *we have*

$$\sqrt{Nh}\left[\hat{\tau}(z) - \tau(z) - \frac{1}{2}\tau^{(2)}(z)h^2\sigma_K^2 + o_p(h^2)\right] \;\to\; \mathcal{N}(0, \sigma^2_\tau(z)),$$

where $\tau^{(2)}(\cdot)$ *is the second derivative of* $\tau(\cdot)$, $\|K\|^2 = \int K^2(t)dt$, $\sigma_K^2 = \int t^2 K(t)dt$, *and* $\sigma^2_\tau(z) = \|K\|^2\mathbb{E}[\tilde{W}^2|Z=z]/p(z)$.

Next, we embrace on presenting the asymptotic normality of the estimator $\hat{\tau}(z,g,t;\delta)$ for the multi-period DiD case.

Theorem 2. *Let* $\tilde{W} = \{G_g - p_g(X)C/(1-p_g(X))\}\{Y_t - Y_{g-\delta-1} - \mu^p_{nev}(X)\}/e_g(z)$ *or* $\tilde{W} = \{G_g - p_{g,t+\delta}(X)(1-G_g)(1-D_{t+\delta})/(1-p_{g,t+\delta}(X))\}\{Y_t - Y_{g-\delta-1} - \mu^p_{nyt}(X)\}/e_g(z)$, *when panel data is available or* $\tilde{W} = \{G_g - p_g(X)C/(1-p_g(X))\}\{Y - \mu^{rc}_{nev}(\tilde{T}_t, X)\}\{\tilde{T}_t - \lambda_t\}/\{e_g(z)$

$\lambda_t(1-\lambda_t)\}$ *or* $\tilde{W} = \{G_g - p_{g,t+\delta}(X)(1-G_g)(1-D_{t+\delta})/(1-p_{g,t+\delta}(X))\}\{Y - \mu^{rc}_{nyt}(\tilde{T}_t, X)\}\{\tilde{T}_t - \lambda_t\}/\{e_g(z)\lambda_t(1-\lambda_t)\}$ *when repeated cross-section data is available. Under Assumption 1–3, 8, 9 and 10(b), 11(b), if* $h = O(N^{-1/5})$, g *and* t *are fixed such that*

$g \in G_\delta, t \in \{2, \ldots, \tilde{T} - \delta\}$ *and* $t \geq g - \delta$*, we have*

$$\sqrt{Nh}\left[\hat{\tau}(z, g, t; \delta) - \tau(z, g, t) - \frac{1}{2}\tau^{(2)}(z, g, t)h^2\sigma_K^2 + o_p(h^2)\right] \quad \to \quad \mathcal{N}(0, \sigma_\tau^2(z)),$$

where $\tau^{(2)}(\cdot)$ *is the second partial derivative with respect to* z *of* $\tau(\cdot)$ *and* $\sigma_\tau^2(z)$ *is similar one in Theorem 1.*

Theorems 1 and 2 establish the asymptotic normality of the proposed estimators for panel data and repeated cross-section data. We generate a doubly robust pseudo-outcome $\tilde{W}$ in the first step and subsequently perform the local linear regressions with a bandwidth of order $h = O(N^{-1/5})$ to achieve the optimal $\sqrt{Nh}$ nonparametric rate as in Stone (1982). These optimal rates result from a trade-off between bias and variance, with bias stemming from higher-order terms in the Taylor series expansion. This aligns with standard theory in the local linear regression; see, for instance, Fan and Gijbels (1996) for details. To obtain these optimal rates, it is essential that the nuisance functional estimators $\hat{\pi}_k(\cdot)$, $\hat{\mu}_k^p(\cdot)$, $\hat{\mu}_k^{rc}(\cdot)$ or $\hat{p}_{g,k}(\cdot)$, $\hat{p}_{g,t+\delta,k}(\cdot)$, $\hat{\mu}_{nev,k}^p(\cdot)$, $\hat{\mu}_{nyt,k}^p(\cdot)$, $\hat{\mu}_{nev,k}^{rc}(\cdot)$, and $\hat{\mu}_{nyt,k}^{rc}(\cdot)$ converge at a rate of $o_p(N^{-1/5})$, which is slower than the rate of the PCATT estimator $\hat{\tau}(\cdot)$. This requirement is due to the double robust properties outlined in Propositions 2 and 4.

We adopt a model-agnostic approach to estimating the PCATT. While parametric methods are commonly employed to estimate nuisance parameters, they often rely on strong assumptions regarding the underlying data structure, which can lead to model misspecification. In contrast, the proposed two-stage doubly robust estimators are model-agnostic and entirely non-parametric. This means we do not need to assume specific model structures for the nuisance parameters, allowing us to implement them using any standard machine learning methods.

Next, we need to estimate the asymptotic variance of $\hat{\tau}(z)$ and construct the confidence interval. In the empirical application, we use a nonparametric bootstrap. Specifically, for $b = 1, \ldots, B$, we draw a bootstrap sample of size n with replacement from the original

data, re-estimate all nuisance functions in the first step, and then recompute $\hat{\tau}^{*(b)}(z)$ in the second step using the same procedure as in the original sample. The bootstrap standard deviation is used as the standard error estimate:

$$\widehat{\mathrm{se}}_{\mathrm{boot}}(z) = \mathrm{sd}\big(\hat{\tau}^{*(1)}(z), \ldots, \hat{\tau}^{*(B)}(z)\big)\,.$$

A pointwise $(1-\alpha)$ confidence interval is then constructed by the normal approximation:

$$\hat{\tau}(z) \;\pm\; z_{1-\alpha/2}\,\widehat{\mathrm{se}}_{\mathrm{boot}}(z).$$

4 Monte Carlo Simulation Studies

4.1 Simulations for Canonical DiD

We first implement simulations on canonical DiD setups to study the finite sample performance of the proposed method. Let $X_1, \ldots, X_5$ be mutual independent and uniformly distributed on $[-1, 1]$ and set $X = (X_1, \ldots, X_5)^\top$ and $Z = X_1$. For $x = (x_1, \ldots, x_5)^\top$, we denote

$$f_{reg}(x) = \cos(x_1 x_2 + 0.5) + \frac{x_2}{1 + e^{0.3 x_3}} + {x_3}^2 - x_3(x_4 + x_5)^2 + (x_4 x_5 - 0.8)^3.$$

We consider the following four data generating processes:

Case 1: $D|X = x \sim Bernoulli(\pi(x)), \pi(x) = \{1 + \exp(x_1 + x_1 \sin(x_2) - {x_3}^2 + \exp(x_2 x_3/5) - x_4(x_5+1)^2 - 0.6)\}^{-1}, Y^0(0) = f_{reg}(X) + \nu_0(X, D) + \epsilon_0, Y^1(d) = 2f_{reg}(X) + \nu_1(X, D) + d\tau(Z) + \epsilon_1(d), d = 0, 1,$ and $\tau(z) = 0$;

Case 2: $D|X = x \sim Bernoulli(\pi(x)), \pi(x) = \{1 + \exp(x_1 + x_1 \sin(x_2) - {x_3}^2 + \exp(x_2 x_3/5) - x_4(x_5+1)^2 - 0.6)\}^{-1}, Y^0(0) = f_{reg}(X) + \nu_0(X, D) + \epsilon_0, Y^1(d) = 2f_{reg}(X) + \nu_1(X, D) + d\tau(Z) + \epsilon_1(d), d = 0, 1,$ and $\tau(z) = 2z$;

Case 3: $D|X = x \sim Bernoulli(\pi(x)), \pi(x) = \{1 + \exp(x_1 + x_1 \sin(x_2) - {x_3}^2 + \exp(x_2 x_3/5) - x_4(x_5+1)^2 - 0.6)\}^{-1}, Y^0(0) = f_{reg}(X) + \nu_0(X,D) + \epsilon_0, Y^1(d) = 2f_{reg}(X) + \nu_1(X,D) + d\tau(Z) + \epsilon_1(d), d = 0, 1$, and $\tau(z) = 2z^2 \sin(z)$;

Case 4: $D|X = x \sim Bernoulli(\pi(x))$, $\pi(x) = \max\{0.1, \min\{\Phi\Big(x_1/5 + \cos(x_4 x_5)^2 + \exp(\frac{x_1 x_5}{1+x_1+x_3})\Big), 0.9\}\}$, $Y^0(0) = f_{reg}(X) + \nu_0(X,D) + \epsilon_0, Y^1(d) = 2f_{reg}(X) + \nu_1(X,D) + d\tau(Z) + \epsilon_1(d)$ for $d = 0$ and 1, and $\tau(z) = 2z^2 \sin(z)$.

Here, $D|X = x \sim Bernoulli(\pi(x))$ means that conditional on $X = x$, treatment indicator D follows a Bernoulli distribution with success probability $\pi(x)$. Random variables $\nu_0(X,D)$ and $\nu_1(X,D)$ are independent normal with mean $D \cdot f_{reg}(X)$ and variance one, and errors $\epsilon_0, \epsilon_1(0), \epsilon_1(1)$ are independent standard normal random variables. In each case, we know that the PCATT $\mathbb{E}[Y^1(1) - Y^1(0)|Z = z, D = 1]$ equals to $\tau(z)$. The observed data is $\{(X_i, D_i, Y_i^0, Y_i^1) : i = 1, \ldots, n\}$, where $Y^0 = Y^0(0)$ and $Y^1 = DY^1(1) + (1-D)Y^1(0)$.

We use different methods to model the two nuisance parameters, i.e., outcome regression (OR) for $\mu(x)$ and propensity scores (PS) for $\pi(x)$ and $e(z)$. For outcome regression, we conduct both linear regression or random forest, while for the propensity score model, we apply logistic regression or random forest. Specifically, we have the following four approaches.

M1: Outcome regression $\mu(x)$ and propensity score $\pi(x)$ are both estimated using the random forest method.

M2: Outcome regression $\mu(x)$ is estimated using linear regression, while propensity score $\pi(x)$ is estimated using the random forest method.

M3: Outcome regression $\mu(x)$ is estimated using the random forest method, while propensity score $\pi(x)$ is estimated using logistic regression.

M4: Outcome regression $\mu(x)$ and propensity score $\pi(x)$ are estimated using linear regres-

sion and logistic regression, respectively.

Among these four estimation methods for the nuisance parameters, $\mu(x)$ and $\pi(x)$ are considered correctly specified if we use random forest, while they are deemed misspecified if we use linear regression or logistic regression. Once we have the estimations of the nuisance functions, we utilize the cross-fitting (say, $K = 5$) and the local linear regression methods described in Section 2.1 to obtain the estimation of PCATT, $\tau(x)$. For each case, we perform 1000 Monte Carlo experiments and use the mean squared error (MSE) to assess the performance of the estimator, which is defined as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{\tau}(z_i) - \tau(z_i)\right]^2,$$

where $\{z_i\}_{i=1}^N$ are grid points with the rang of Z. Tables 1 – 4 in Sections 4.1 and 4.2, respectively, report the median and standard deviation in parentheses for 1,000 MSEs for panel data and repeated cross-sectional data under the four cases and four estimation methods for $\hat{\tau}(z)$.

In general, the estimation errors decrease steadily as the sample size increases from 500 to 4000. As expected, the best performance occurs when both the outcome regression $\mu(x)$ and propensity score $\pi(x)$ are modeled correctly (i.e., M1), while M4 performs the worst when these nuisance parameters are misspecified. Furthermore, when either $\mu(x)$ or $\pi(x)$ is misspecified (i.e., M2 or M3), their performances are similar, being worse than M1 but better than M4. This demonstrates the effectiveness of the proposed method.

4.2 Simulations of Multi-period DiD

Next, we conduct simulation studies to evaluate the finite-sample performance of proposed method in multi-period DiD setups. We set $X = (X_1, \ldots, X_5)^\top$ with $X_1, \ldots, X_5$ are mutually independent and normally distributed with mean zero and variance one, and set $Z = X_1$. Here, we use the "not-yet-treated" group as the control group. Assume that

Table 1: Median and standard deviation of 1,000 MSE values for panel data in canonical DiD.

	Case 1				Case 2				Case 3				Case 4			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
500	0.192	0.332	0.291	0.523	0.177	0.361	0.268	0.539	0.210	0.377	0.297	0.575	0.133	0.175	0.151	0.196
	(0.184)	(0.260)	(0.238)	(0.336)	(0.208)	(0.366)	(0.303)	(0.449)	(0.220)	(0.336)	(0.269)	(0.492)	(0.148)	(0.183)	(0.164)	(0.212)
1000	0.098	0.195	0.181	0.421	0.098	0.220	0.180	0.449	0.115	0.223	0.197	0.454	0.071	0.107	0.084	0.124
	(0.079)	(0.124)	(0.116)	(0.198)	(0.102)	(0.169)	(0.134)	(0.231)	(0.100)	(0.141)	(0.126)	(0.216)	4(0.067)	(0.088)	(0.085)	(0.111)
2000	0.050	0.112	0.120	0.373	0.049	0.122	0.114	0.378	0.061	0.132	0.129	0.394	0.039	0.071	0.047	0.080
	(0.044)	(0.062)	(0.067)	(0.120)	(0.045)	(0.070)	(0.066)	(0.127)	(0.043)	(0.073)	(0.066)	(0.139)	((0.036)	(0.049)	(0.043)	(0.061)
4000	0.039	0.089	0.093	0.364	0.041	0.097	0.090	0.367	0.034	0.080	0.087	0.364	0.022	0.053	0.026	0.061
	(0.036)	(0.052)	(0.055)	(0.105)	(0.038)	(0.054)	(0.053)	(0.105)	(0.024)	(0.035)	(0.038)	(0.080)	(0.017)	(0.028)	(0.021)	(0.036)

Table 2: Median and standard deviation of 1,000 MSE values for for repeated cross-section data in canonical DiD.

	Case 1				Case 2				Case 3				Case 4			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
500	1.109	1.516	1.476	1.914	0.263	0.478	0.353	0.642	0.300	0.491	0.380	0.664	0.786	0.890	0.997	1.122
	(1.720)	(3.099)	(2.350)	(6.303)	(0.292)	(0.426)	(0.361)	(0.500)	(0.346)	(0.452)	(0.381)	(0.654)	(1.272)	(1.487)	(1.430)	(1.639)
1000	0.550	0.795	0.742	1.062	0.140	0.269	0.209	0.489	0.159	0.283	0.235	0.495	0.382	0.430	0.507	0.560
	(0.703)	(0.819)	(0.887)	(1.208)	(0.137)	(0.225)	(0.175)	(0.291)	(0.146)	(0.202)	(0.185)	(0.264)	(0.511)	(0.569)	(0.639)	(0.709)
2000	0.270	0.487	0.407	0.683	0.073	0.159	0.137	0.407	0.089	0.166	0.151	0.414	0.188	0.238	0.262	0.308
	(0.317)	(0.393)	(0.442)	(0.609)	(0.079)	(0.109)	(0.096)	(0.167)	(0.080)	(0.105)	(0.096)	(0.162)	(0.238)	(0.258)	(0.339)	(0.382)
4000	0.133	0.290	0.225	0.497	0.042	0.099	0.089	0.371	0.052	0.105	0.103	0.381	0.098	0.135	0.137	0.178
	(0.158)	(0.201)	(0.258)	(0.346)	(0.04)	(0.059)	(0.059)	(0.115)	(0.043)	(0.060)	(0.060)	(0.114)	(0.102)	(0.130)	(0.170)	(0.195)

there is no treatment effect anticipation, i.e., $\delta = 0$. For $x = (x_1, ..., x_5)^\top$, we further set

$$G_g|X = x \sim Bernoulli(p_g(x)),$$

$$p_g(x) = P(G = g|X = x) = \frac{\gamma_g(x_1 + \cos(\pi x_2/2) + 3\cos(2x_1x_4) + 2\sin(3x_3x_4x_5))}{\sum_{g\in G}\gamma_g(x_1 + \cos(\pi x_2/2) + 3\cos(2x_1x_4) + 2\sin(3x_3x_4x_5))},$$

$$Y_{it}(0) = \theta_t + \eta_i + f_{it}(X) + u_{it},$$

$$Y_{it}(g) = Y_{it}(0) + I\{g \le t\}\tau(z, g, t) + \epsilon_{it} - u_{it},$$

and

$$f_{it}(X) = X_i'\beta_t, \beta_t = (t, t, 2t, 4t, 3t)^T,$$

where $\gamma_g = (0.5g+0.5)/g$, and errors $u_t, \epsilon_t \sim N(0, 1)$ in all time periods. Here, G represents the time period when a unit is first treated and G_g follows a Bernoulli distribution with success probability $p_g(x)$, indicating whether one is first treated in period g. We set the untreated potential outcome for individuals as $Y_{it}(0)$, where the time-fixed effects $\theta_t = t$ and the individual-fixed effects $\eta|(G, X) \sim N(G, 1)$, and the treated potential outcome, $Y_{it}(g)$. This implies that the PCATT equals to $\tau(z, g, t)$ for $g \le t$. We further consider the following three cases for $\tau(z, g, t)$:

Case 5: $\tau(z, g, t) = 0$;

Case 6: $\tau(z, g, t) = 2z + t - g + 1$;

Case 7: $\tau(z, g, t) = 2z^2\sin(z) + t - g + 1$.

We estimate the PCATT $\tau(z, g, t)$ with $g = 2, t = 3$. Under the settings of four time periods $T = 1, 2, 3, 4$ and group structure $G \in \mathcal{G} = \{0, 1, 2, 3, 4\}$, we generate data of sample sizes {1250, 2500, 5000, 10000} and then drop units who received treatment in the first period. We use different methods to model the two nuisance parameters. For outcome regression, we conduct linear regression, while for the propensity score model, we apply logistic regression or random forest. Specifically, we have the following four approaches.

M1: Outcome regression $\mu(x)$ and propensity score $\pi(x)$ are estimated using linear regression and random forest, respectively.

M2: Outcome regression $\mu(x)$ is estimated using linear regression with only the first three covariates (X_1, X_2, X_3), omitting X_4 and X_5, while propensity score $\pi(x)$ is estimated using the random forest method.

M3: Outcome regression $\mu(x)$ and propensity score $\pi(x)$ are estimated using linear regression and logistic regression, respectively.

M4: Outcome regression $\mu(x)$ is estimated using linear regression with only the first three covariates (X_1, X_2, X_3), omitting X_4 and X_5, while propensity score $\pi(x)$ is estimated using logistic regression.

Among these four estimation methods for the nuisance functions, $\mu(x)$ and $\pi(x)$ are considered correctly specified if we use linear regression with full set of covariates and random forest, respectively. Conversely, they are deemed mis-specified if we use linear regression with omitting the last two covariates or logistic regression. The results, based on 1,000 Monte Carlo experiments, are displayed in Tables 3 and 4. They show the similar pattern with the results in canonical DiD case.

5 Reanalysis of DACA Data

Deferred Action for Childhood Arrivals (DACA) was announced by President Obama on June 15, 2012. This program allowed the Department of Homeland Security to accept applications from undocumented immigrants who had arrived in the U.S. as children (under 16) and were under 31 years of age on that date. One feature of DACA is that qualified individuals would receive deportation deferral for two years and become eligible for work authorization.

Table 3: Median and standard deviation of MSE for panel data in multi-peroid DiD.

	Case 5				Case 6				Case 7			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
1250	0.039	0.841	0.067	1.870	0.161	0.798	0.617	1.823	0.635	1.612	0.919	2.155
	(0.125)	(3.673)	(0.372)	(5.840)	(0.545)	(3.523)	(0.738)	(6.100)	(0.473)	(4.508)	(0.591)	(6.805)
2500	0.015	0.137	0.044	0.301	0.093	0.280	0.503	0.940	0.350	1.034	0.591	1.458
	(0.055)	(0.475)	(0.191)	(0.964)	(0.254)	(1.712)	(0.395)	(3.152)	(0.228)	(1.563)	(0.307)	(2.973)
5000	0.006	0.054	0.078	0.139	0.072	0.107	0.375	0.499	0.162	0.767	0.384	1.032
	(0.018)	(0.148)	(0.121)	(0.491)	(0.127)	(0.523)	(0.228)	(1.498)	(0.098)	(0.562)	(0.173)	(1.390)
10000	0.003	0.037	0.112	0.197	0.034	0.049	0.305	0.298	0.082	0.571	0.304	0.676
	(0.007)	(0.119)	(0.088)	(0.479)	(0.057)	(0.238)	(0.144)	(0.822)	(0.047)	(0.322)	(0.110)	(0.679)

Table 4: Median and standard deviation of MSE for repeated cross-section data in multi-peroid DiD.

	Case 5				Case 6				Case 7			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
1250	0.440	1.278	0.597	1.325	0.671	5.068	1.111	6.811	11.977	2.317	2.547	3.047
	(1.609)	(4.350)	(2.070)	(6.410)	(2.283)	(17.080)	(2.798)	(22.178)	(2.945)	(4.537)	(2.994)	(4.626)
2500	0.193	0.544	0.283	0.581	0.293	2.060	0.626	2.789	1.132	1.356	1.575	2.104
	(0.647)	(1.935)	(0.922)	(2.351)	(1.012)	(6.740)	(1.387)	(9.247)	(1.295)	(2.532)	(1.333)	(3.160)
5000	0.085	0.260	0.142	0.277	0.164	0.946	0.504	1.443	0.085	0.404	0.425	0.719
	(0.263)	(0.694)	(0.488)	(1.236)	(0.497)	(3.138)	(0.684)	(5.176)	(0.228)	(1.179)	(0.390)	(2.030)
10000	0.038	0.101	0.107	0.145	0.074	0.143	0.416	0.287	0.322	0.454	0.610	0.722
	(0.119)	(0.278)	(0.295)	(0.579)	(0.255)	(0.565)	(0.417)	(1.006)	(0.253)	(0.436)	(0.340)	(0.635)

A growing literature examines the effects of DACA across education, employment, income, and health. In education, studies such as Kuka et al. (2020) and Henderson et al. (2023) found that DACA significantly increased high school enrollment and completion rates while the program also appears to have induced a trade-off between schooling and work for some individuals. For example, Pope (2016) showed that DACA increased employment primarily by raising labor force participation and reducing unemployment among eligible immigrants. Beyond socioeconomic outcomes, DACA has been shown to improve health outcomes. Using Oregon Medicaid claims data, Hainmueller et al. (2017) found improvements in children's mental-health diagnoses associated with maternal DACA eligibility.

We apply our method to study the DACA dataset from the American Community Survey (ACS), which is publicly available in Pope (2016). The sample consists of 82619 noncitizens aged 16–32 observed in 2006 and 2013. Of these, 24059 are DACA-eligible (the treatment group) and 58560 are non-eligible (the control group). We consider two employment outcomes: Y_1 (log income plus one), Y_2 (weekly working hours) and five covariates: X_1 (gender), X_2 (married), X_3 (ethnicity), X_4 (age), X_5 (state-level unemployment rate). Y_1 measures the total amount of income an individual receives from all sources in the last 12 months and Y_2 measures the number of hours worked per week. Summary statistics are presented in Table 5. Compared with non-eligible individuals, the treated group is younger, more likely to be Hispanic, and less likely to be married.

Table 5: Summary statistics

Variable	Untreated	Treated	Diff.	P_val on diff.
X_1(Gender)	0.54	0.52	-0.02	0.00
X_2(Married)	0.54	0.52	-0.02	0.00
X_3(Ethnicity)	0.55	0.65	0.10	0.00

Variable	Untreated	Treated	Diff.	P_val on diff.
X_4(Age)	24.57	21.23	-3.34	0.00
X_5(State unemployment rate)	6.55	6.32	-0.23	0.00
No. of observations	24059	58560		

We are interested in studying how the effect of DACA vary for different subsamples of gender, married, ethnicity, and age. The estimation procedure of PCATT on discrete covariates is given in Appendix. Following the estimation procedures and cross-fitting approach with $K = 5$ as in Section 2, we estimate the nuisance functions and then get PCATT for each subgroup. The estimates along with their 95% point-wise confidence intervals obtained from 1,000 bootstrap samples are displayed in Figures 1 – 3. The results

Figure 1: Effect of DACA on income by discrete covariates

on the income and weekly working hours are consistent in both magnitude and sign. The estimated effects of DACA on gender fluctuate around zero, indicating little heterogeneity. The middle panels of Figure 1 and 2 show positive effect among unmarried individuals, whereas the effect for the married group is close to zero. This disparity may be explained by the fact that unmarried individuals are typically younger and more geographically mobile, enabling them to take advantage of employment opportunities once employment restrictions are relaxed. In contrast, married individuals' especially those with children' are frequently

Figure 2: Effect of DACA on the weekly working hours by discrete covariates

constrained by their family responsibilities and childcare demands. The right panels present the estimates by ethnicity and exhibit a statistically positive impact of DACA for Hispanic, which is the target group of DACA. Figure 3 presents the estimated effects of DACA on

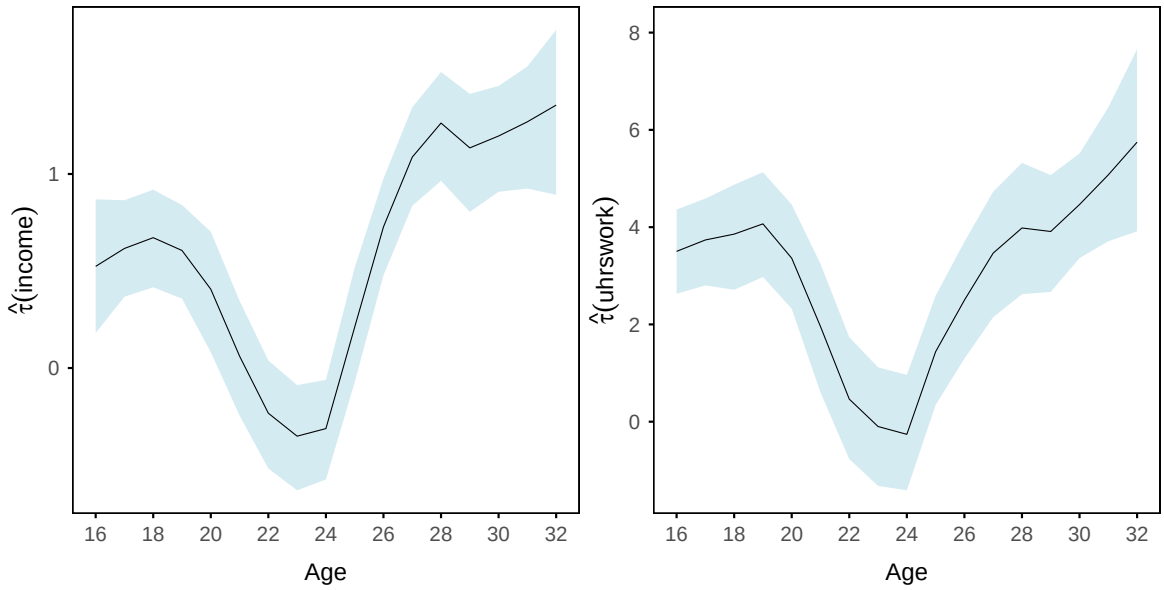

Figure 3: Effect of DACA on income and weekly working hours by age

income and weekly working hours by age group. The effects are positive for both outcomes among individuals aged 16-21 and 26-32. However, the positive effect gradually weakens within the 16-21 age group, while it strengthens across the 26-32 group. Notably, a negative effect around ages 22-25 is observed on income, while no significant effect is detected on weekly working hours in that range.

Undocumented youth aged 16-18 had minimal formal income or recorded working hours prior to DACA. Legal work authorization enabled them to enter the labor market immediately, demonstrating a significant positive impact. The subsequent weakening trend among 18-21 year-olds and negative income effect might suggest a shift from work toward an increment on educational investments, which may moderate the growth trajectory of their early-career earnings and hours. This interpretation aligns with Kuka et al. (2020), which find that DACA raised high school attendance and graduation rates among students on the margin of completion or further from the typical high school enrollment age. In contrast, the strengthening effect among those in the 26-32 age group might stem from formal employment opportunities, completed education and better job matches over time.

Therefore, the heterogeneity across age in our results shed new light and further enrich the understanding of how DACA's benefits affect individuals and the labor market.

6 Conclusion

In this paper, we study the partially conditional average treatment effect for DiD settings. This approach can capture the heterogeneous effects of treatment on the response and provide a visualization of results as curves. We further estimate the PCATT using a doubly robust method and establish the asymptotic normality for the estimator. Our methods constitute valuable additions to the DiD literature and are likely to improve their applications. Based on this work, it needs to investigate whether there exists heterogeneity in treatment effects for covariate Z. To this end, one might consider to test $\tau(z)$ is a constant or not, as Cai et al. (2025).

Yet there are other different forms of the treatment variable for future work. While DiD has been well-developed for binary treatment form, comprehensive studies on continuous or multi-valued discrete treatment form under more general DiD designs still remains to

be conducted. For example, one could explore doubly robust estimators to compare effects between two treatment doses by imposing a stronger parallel trends assumption as in Kennedy et al. (2017) and Callaway et al. (2024).

Another avenue for future work is extending the methodology to accommodate complex covariate structures. For instance, researchers may encounter high-dimensional or time-varying covariates, whereas previous studies predominantly focus on time-invariant covariates within the treated group under low-dimensional framework. Our methodology could be adapted by making some additional functional form assumptions for untreated potential outcomes and leveraging dimensionality reduction techniques, such as LASSO-type penalties as in Zimmert (2018), Caetano et al. (2022), Haddad et al. (2024), and references therein.

7 Data Availability Statement

Deidentified data have been made available at http://dx.doi.org/10.1016/j.jpubeco.2016.08.014.

SUPPLEMENTARY MATERIAL

Title: Proof of main theoretical results. (Supplement.pdf)

R-code: R-code containing code to perform the simulation and the application described in the article. (R-code.zip)

DACA data set: Data set used in Section 5. (DACA_data.zip)

References

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

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

Abrevaya, J., Hsu, Y.-C. and Lieli, R. P. (2015), 'Estimating conditional average treatment effects', *Journal of Business & Economic Statistics* **33**(4), 485–505.

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

Baker, A. C., Larcker, D. F. and Wang, C. C. (2022), 'How much should we trust staggered difference-in-differences estimates?', *Journal of Financial Economics* **144**(2), 370–395.

Blundell, R., Dias, M. C., Meghir, C. and Van Reenen, J. (2004), 'Evaluating the employment impact of a mandatory job search program', *Journal of the European Economic Association* **2**(4), 569–606.

Borusyak, K., Jaravel, X. and Spiess, J. (2024), 'Revisiting event-study designs: robust and efficient estimation', *Review of Economic Studies* **91**(6), 3253–3285.

Caetano, C., Callaway, B., Payne, S. and Rodrigues, H. S. (2022), 'Difference in differences with time-varying covariates', *arXiv preprint arXiv:2202.02903* .

Cai, Z., Fang, Y., Lin, M. and Tang, S. (2021), 'Estimation of partially conditional quantile treatment effect model', *China Journal of Econometrics* **1**(4), 741–762.

Cai, Z., Fang, Y., Lin, M. and Tang, S. (2025), 'A nonparametric test of heterogeneity in conditional quantile treatment effects', *Econometric Theory* **41**(3), 660–687.

Callaway, B., Goodman-Bacon, A. and Sant'Anna, P. H. (2024), Difference-in-differences with a continuous treatment, Technical report, National Bureau of Economic Research.

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

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

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

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

Ding, P. (2024), *A First Course in Causal Inference*, CRC Press, Boca Raton, FL.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman & Hall, London.

Fan, Q., Hsu, Y.-C., Lieli, R. P. and Zhang, Y. (2022), 'Estimation of conditional average treatment effects with high-dimensional data', *Journal of Business & Economic Statistics* **40**(1), 313–327.

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

Grimmer, J., Messing, S. and Westwood, S. J. (2017), 'Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods', *Political Analysis* **25**(4), 413–434.

Haddad, M. F., Huber, M. and Zhang, L. Z. (2024), 'Difference-in-differences with time-

varying continuous treatments using double/debiased machine learning', *arXiv preprint arXiv:2410.21105* .

Hainmueller, J., Lawrence, D., Martén, L., Black, B., Figueroa, L., Hotard, M., Jiménez, T. R., Mendoza, F., Rodriguez, M. I., Swartz, J. J. et al. (2017), 'Protecting unauthorized immigrant mothers improves their children's mental health', *Science* **357**(6355), 1041–1044.

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

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

Henderson, D. J., Sperlich, S. et al. (2023), 'A complete framework for model-free difference-in-differences estimation', *Foundations and Trends® in Econometrics* **12**(3), 232–323.

Hernán, M. A. and Robins, J. M. (2010), *Causal Inference*, CRC Press, Boca Raton, FL.

Hirano, K., Imbens, G. W. and Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**(4), 1161–1189.

Imai, K., Kim, I. S. and Wang, E. H. (2023), 'Matching methods for causal inference with time-series cross-sectional data', *American Journal of Political Science* **67**(3), 587–605.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical sciences*, Cambridge University Press, New York.

Kennedy, E. H., Ma, Z., McHugh, M. D. and Small, D. S. (2017), 'Nonparametric methods for doubly robust estimation of continuous treatment effects', *Journal of the Royal Statistical Society, Series B* **79**(4), 1229–1245.

Kuka, E., Shenhav, N. and Shih, K. (2020), 'Do human capital decisions respond to the returns to education? Evidence from DACA', *American Economic Journal: Economic Policy* **12**(1), 293–324.

Lee, S., Okui, R. and Whang, Y.-J. (2017), 'Doubly robust uniform confidence band for the conditional average treatment effect function', *Journal of Applied Econometrics* **32**(7), 1207–1225.

Malani, A. and Reif, J. (2015), 'Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform', *Journal of Public Economics* **124**(C), 1–17.

Neyman, J. (1923), 'Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle.', *Statistical Science* **5**(1), 465–475. English translation of excerpts by D. Dabrowska and T. Speed.

Pope, N. G. (2016), 'The effects of DACAmentation: The impact of Deferred Action for Childhood Arrivals on unauthorized immigrants', *Journal of Public Economics* **143**, 98–114.

Roth, J., Sant'Anna, P. H., Bilinski, A. and Poe, J. (2023), 'What's trending in difference-in-differences? A synthesis of the recent econometrics literature', *Journal of Econometrics* **235**(2), 2218–2244.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of Educational Psychology* **66**(5), 688.

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

Stone, C. J. (1982), 'Optimal global rates of convergence for nonparametric regression', *Annals of Statistics* **10**(4), 1040–1053.

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

Wooldridge, J. M. (2025), 'Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators', *Empirical Economics* **69**(5), 2545–2587.

Zhou, N., Guo, X. and Zhu, L. (2022), 'The role of propensity score structure in asymptotic efficiency of estimated conditional quantile treatment effect', *Scandinavian Journal of Statistics* **49**(2), 718–743.

Zhou, N. and Zhu, L. (2021), 'On IPW-based estimation of conditional average treatment effects', *Journal of Statistical Planning and Inference* **215**(December), 1–22.

Zimmert, M. (2018), 'Efficient difference-in-differences estimation with high-dimensional common trend confounding', *arXiv preprint arXiv:1809.01643* .