Doubly Robust Difference-in-Differences Estimators

Pedro H. C. Sant'Anna* Jun B. Zhao†

Vanderbilt University Vanderbilt University

May 5, 2020

Abstract

This article proposes doubly robust estimators for the average treatment effect on the treated (ATT) in difference-in-differences (DID) research designs. In contrast to alternative DID estimators, the proposed estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified. We also derive the semiparametric efficiency bound for the ATT in DID designs when either panel or repeated cross-section data are available, and show that our proposed estimators attain the semiparametric efficiency bound when the working models are correctly specified. Furthermore, we quantify the potential efficiency gains of having access to panel data instead of repeated cross-section data. Finally, by paying particular attention to the estimation method used to estimate the nuisance parameters, we show that one can sometimes construct doubly robust DID estimators for the ATT that are also doubly robust for inference. Simulation studies and an empirical application illustrate the desirable finite-sample performance of the proposed estimators. Open-source software for implementing the proposed policy evaluation tools is available.

*Department of Economics, Vanderbilt University. E-mail: pedro.h.santanna@vanderbilt.edu.

†Department of Economics, Vanderbilt University. E-mail: jun.zhao@vanderbilt.edu.

1 Introduction

Difference-in-differences (DID) methods are among the most popular procedures practitioners adopted to conduct policy evaluation with observational data. In its canonical form, DID identifies the average treatment effect on the treated (ATT) by comparing the difference in pre and post-treatment outcomes of two groups: one that receives and one that does not receive the treatment (the treated and comparison group, respectively). In order to attach a causal interpretation to DID estimators, researchers routinely invoke the (unconditional) parallel trends assumption (PTA): in the absence of the treatment, the average outcome for the treatment and comparison groups would have followed parallel paths over time. Although the PTA is fundamentally untestable, its plausibility is usually questioned if the observed characteristics that are thought to be associated with the evolution of the outcome are not balanced between the treated and comparison group. In such cases, researchers usually deviate from the canonical DID setup and incorporate pre-treatment covariates into the DID analysis and assume that the PTA is satisfied only after conditioning on these covariates.

In this paper, we study the robustness and efficiency properties of DID estimators for the ATT when the PTA holds after conditioning on covariates. We consider both settings where panel data are available and settings where only repeated cross-section data are available. We contribute to the DID literature in different fronts. First, we derive doubly robust (DR) estimands for the ATT under DID settings and propose DR DID estimators for the ATT that are consistent when either a working (parametric) model for the propensity score or a working (parametric) model for the outcome evolution for the comparison group is correctly specified. The setting where only repeated cross-section data are available is particularly interesting. We propose two different DR DID estimators for the ATT that differ from each other depending on whether or not one models the outcome regression for the treated group in both pre and post-treatment periods. Nonetheless, we show DR property does not depend on such a choice.

Second, we derive the semiparametric efficiency bounds for the ATT under DID designs. The semiparametric efficiency bounds we derive are nonparametric in the sense that we do not assume researchers have additional knowledge about outcome regressions or the propensity score functional forms. As so, these bounds provide a standard against which one can compare the efficiency of any (regular) semiparametric DID estimator for the ATT. Here, it is also worth stressing that these semiparametric efficiency bounds explicitly incorporate all the restrictions implied by the invoked identification assumptions. Importantly, these restrictions differ depending on whether panel or repeated cross-section data are available. In both cases they involve the moment restrictions implied by the conditional PTA, though, when repeated cross-section data are available, they also include the restrictions implied by the identifying assumption that the joint distribution of covariates and treatment status is invariant to the sampling period (pre and post-treatment). We emphasize that failing to account for all these

implied restrictions can lead to discrepancies on the derived efficiency bound, which, in turn, may suggest that some estimator is semiparametrically efficient when, in fact, it is not.

With the semiparametric efficiency bounds at hand, we can answer several questions that one may have. For instance, one may wonder whether there are efficiency gains associated with having access to panel instead of repeated cross-section data. By directly comparing the efficiency bounds under these two setups, we not only show that the answer to the aforementioned question is yes, but also show that such gains tend to be larger when the sample sizes of the pre and post-treatment repeated cross-section data are more imbalanced.

Another natural question that arises is whether our proposed DR DID estimators can attain the semiparametric efficiency bound. We show that when the working models for the propensity score and for the outcome evolution for the comparison group are correctly specified, our proposed DR DID estimator for the panel data setup is locally efficient, though the DR DID estimators for the cross-section setup are not. In fact, when only repeated cross-section data are available, we show that our proposed DR DID estimator that relies on modelling the propensity score and the outcome evaluation of *both* the treated and comparison groups attains the semiparametric efficiency bound when all working models are correctly specified. We quantify the loss of efficiency associated with using the inefficient DR DID estimator instead of the locally efficient one, and illustrate via Monte Carlo simulations that such a loss can indeed be large.

Our proposed methodology accommodates linear and nonlinear working models for the nuisance functions. We establish $\sqrt{n}$-consistency and asymptotic normality of the proposed DR DID estimators when generic parametric working models are used for the nuisance functions. In doing so we emphasize that, in general, the DR property of our estimators is with respect to consistency and not to inference. In other words, the exact form of the asymptotic variance of our proposed estimators depends on whether the propensity score and/or the outcome regression models are correctly specified. Given that, in practice, one does not know a priori which models are correctly specified, one should consider the estimation effects from all first-step estimators when estimating the asymptotic variance. Failing to do so may lead to invalid inference procedures.

Motivated by this observation, a third contribution of this paper is to show that, by paying particular attention to the estimation method used for estimating the nuisance parameters, it is sometimes possible to construct computationally simple DID estimators for the ATT that are not only DR consistent and locally semiparametric efficient, but are also doubly robust for inference. These further improved DR DID estimators are particularly attractive and easy to implement when researchers are comfortable with a logistic working model for the propensity score and with linear regression working models for the outcome of interest.

Related literature: Our proposal builds on two branches of the causal inference literature. First, our methodological results are intrinsically related to other DID papers; for an overview, see e.g., Section 6.5 of Imbens and Wooldridge (2009) and references therein. Two leading contributions in this branch of literature

that are particularly relevant to this paper are Heckman et al. (1997), who propose kernel-based DID regression estimators, and Abadie (2005), who proposes (parametric and nonparametric) DID inverse probability weighted (IPW) estimators. We note that when the dimension of available covariates is high or even moderate, fully nonparametric procedures usually do not lead to informative inference because of the "curse of dimensionality". In these cases, researchers often adopt parametric methods. Our DR DID estimators fall in this latter category.

Second, our results are also directly related to the literature on doubly robust estimators, see Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Wooldridge (2007), Chen et al. (2008), Cattaneo (2010), Graham et al. (2012, 2016), Vermeulen and Vansteelandt (2015), Lee et al. (2017), Słoczyński and Wooldridge (2018), Rothe and Firpo (2018), Muris (2019), among many others; for an overview, see section 2 of Słoczyński and Wooldridge (2018), and Seaman and Vansteelandt (2018). Recently, DR estimators have also been playing an important role when one uses data-adaptive, "machine learning" estimators for the nuisance functions, see e.g., Belloni et al. (2014), Farrell (2015), Chernozhukov et al. (2017), Belloni et al. (2017), and Tan (2019). As so, these papers are also broadly related to our proposal, even though we use parametric first-step estimators. On the other hand, we note that the aforementioned papers focus on either the "selection on observables" or "IV/LATE" type assumptions, whereas we pay particular attention to the conditional DID design. Thus, our results complement theirs.

To derive the semiparametric efficiency bounds for the ATT under the DID framework, we build on Hahn (1998) and Chen et al. (2008). Although we follow the structure of semiparametric efficiency bound derivation of the aforementioned papers (which, in turn, follow Newey (1990)), our derived semiparametric efficiency bounds complement theirs as we focus on DID designs while Hahn (1998) and Chen et al. (2008) results rely on "selection on observables" type assumptions in cross-section setups.

Our results for the further improved DR DID estimators build on Vermeulen and Vansteelandt (2015), who propose estimators that are DR for inference in cross-section setups under selection on observables type assumptions. We extend Vermeulen and Vansteelandt (2015) proposal to DID settings with both panel and repeated cross-section data. Our further improved DR DID estimators also build on Graham et al. (2012), as their proposed propensity score estimator is one important component of our proposal.

Finally, in work related but independent from ours, Zimmert (2019) provides high-level conditions under which one can use "machine-learning" first-step estimators when estimating the ATT in DID setups. His results complement ours, though we note that his proposed estimators for the repeated cross-section case do not attain the semiparametric efficiency bound derived in this paper, and the loss of efficiency can be of first-order importance. We also note that Zimmert (2019) does not provide a detailed comparison between the panel and repeated cross-section data setups like we do, nor discusses DR inference procedures, which are particularly relevant under model misspecifications.

Organization of the paper: In the next section, we describe this paper's framework, briefly give an overview of the existing DID estimators and describe how we combine the strengths of each method to form our DR DID estimands. We also derive semiparametric efficiency bounds for the ATT in Section 2. In Section 3, we propose different DR DID estimators, derive their large sample properties, and show that we can get improved DR DID estimators by paying particular attention to the estimation method used for estimating the nuisance parameters. We examine the finite sample properties of our proposed methodology by means of a Monte Carlo study in Section 4, and provide an empirical illustration in Section 5. Section 6 concludes. Mathematical proofs are gathered in the Supplemental Appendix.[1] Finally, all proposed policy evaluation tools discussed in this article can be implemented via the open-source R package `DRDID`, which is freely available from GitHub (https://pedrohcgs.github.io/DRDID/).

2 Difference-in-differences

2.1 Background

We first introduce the notation we use throughout the article. We focus on the case where there are two treatment periods and two treatment groups. Let Y_{it} be the outcome of interest for unit i at time t. We assume that researchers have access to outcome data in a pre-treatment period $t=0$ and in a post-treatment period $t=1$. Let $D_{it}=1$ if unit i is treated before time t and $D_{it}=0$ otherwise. Note that $D_{i0}=0$ for every i, allowing us to write $D_i=D_{i1}$. Using the potential outcome notation, denote $Y_{it}(0)$ the outcome of unit i at time t if it does not receive treatment by time t and $Y_{it}(1)$ the outcome for the same unit if it receives treatment. Thus, the realized outcome for unit i at time t is $Y_{it}=D_iY_{it}(1)+(1-D_i)Y_{it}(0)$. A vector of pre-treatment covariates X_i is also available. Henceforth, we assume that the first element of X_i is a constant.

In the rest of the article, we assume that either panel or repeated cross-section data on (Y_{it},D_i,X_i), $t=0,1$ are available. When repeated cross-section data are available, we follow Abadie (2005) and assume that covariates and treatment status are stationary. We formalize these conditions in the following assumption. Let T_i be a dummy variable that takes value one if the observation i is only observed in the post-treatment period, and zero if observation i is only observed in the pre-treatment period. Define $Y_i=T_iY_{i1}+(1-T_i)Y_{i0}$, and let n_1 and n_0 be the sample sizes of the post-treatment and pre-treatment periods such that $n=n_1+n_0$. Finally, let $\lambda=\mathbb{P}(T=1)\in(0,1)$.

Assumption 1 *Assume that either* (a) *the data* $\{Y_{i0},Y_{i1},D_i,X_i\}_{i=1}^n$ *are independent and identically distributed* (iid)*; or* (b) *the pooled repeated cross-section data* $\{Y_i,D_i,X_i,T_i\}_{i=1}^n$ *consist of iid draws from the mixture*

1 The Supplemental Appendix is available at https://pedrohcgs.github.io/files/DR-DIDAppendix.pdf

distribution

$$\begin{aligned} P(Y \leq y, D = d, X \leq x, T = t) &= t \cdot \lambda \cdot P(Y_1 \leq y, D = d, X \leq x | T = 1) \\ &\quad + (1-t) \cdot (1-\lambda) P(Y_0 \leq y, D = d, X \leq x | T = 0), \end{aligned}$$

where $(y,d,x,t) \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^k \times \{0,1\}$*, with the joint distribution of* (D,X) *being invariant to* T*.*

Assumption 1(a) covers the case where panel data are available, whereas Assumption 1(b) covers the case where repeated cross-section data are available, and allows for different sampling schemes. For instance, it accommodates the binomial sampling scheme where an observation i is randomly drawn from either (Y_1, D, X) or (Y_0, D, X) with fixed probability λ (here, T is a non-degenerated random variable). It also accommodates the "conditional" sampling scheme where n_1 observations are sampled from (Y_1, D, X), n_0 observations are sampled from (Y_0, D, X) and $\lambda = n_1/n$ (here, T is treated as fixed). On the other hand, Assumption 1(b) rules out settings with compositional changes in (D,X), see e.g. Hong (2013) for a discussion.

The parameter of interest is the average treatment effect on the treated,

$$\tau = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) | D_i = 1].$$

As expectations are linear operators and $Y_{i1}(1) = Y_{i1}$ if $D_i = 1$, we can rewrite the ATT as[2]

$$\tau = \mathbb{E}[Y_1(1)|D=1] - \mathbb{E}[Y_1(0)|D=1] = \mathbb{E}[Y_1|D=1] - \mathbb{E}[Y_1(0)|D=1], \tag{2.1}$$

where we drop subscript i to ease notation; we follow this convention throughout the paper. From the above representation, it is clear that the main challenge in identifying the ATT is to compute $\mathbb{E}[Y_{i1}(0)|D_i = 1]$ from the observed data. To overcome this challenge, we invoke the following assumptions.

Assumption 2 $\mathbb{E}[Y_1(0) - Y_0(0)|D=1,X] = \mathbb{E}[Y_1(0) - Y_0(0)|D=0,X]$ *almost surely (a.s.).*

Assumption 3 *For some* $\varepsilon > 0$*,* $\mathbb{P}(D=1) > \varepsilon$ *and* $\mathbb{P}(D=1|X) \leq 1-\varepsilon$ *a.s..*

Assumption 2, which we refer to as the conditional PTA throughout the paper, states that in the absence of treatment, the average conditional outcome of the treated and the comparison groups would have evolved in parallel. Note that Assumption 2 allows for covariate-specific time trends, though it rules out unit specific trends. Assumption 3 is an overlap condition and states that at least a small fraction of the population is treated and that for every value of the covariates X, there is at least a small probability that the unit is not treated. These two assumptions are standard in conditional DID methods, see e.g. Heckman et al. (1997), Heckman et al. (1998), Blundell et al. (2004), Abadie (2005) and Bonhomme and Sauder (2011).

---

2 Throughout the rest of the paper, to ease the notation burden we denote $\mathbb{E}[\cdot]$ as generic expectations. In the case of panel data, such expectations are with respect to the distribution of (Y_0, Y_1, D, X). In the case of repeated cross-section data, the expectations are with respect to the mixture distribution $\sum_{t=0}^{1} \mathbb{P}(T=t) \cdot \mathbb{P}(Y_t \leq y, D=d, X \leq x | T=t)$.

Under Assumptions 1-3, there are two main flexible estimation procedures to estimate the ATT: the outcome regression (OR) approach, see e.g. Heckman et al. (1997), and the IPW approach, see e.g. Abadie (2005). The OR approach relies on researchers ability to model the outcome evolution. In such cases, under the aforementioned assumptions one can estimate the ATT using

$$\widehat{\tau}^{reg} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + n_{treat}^{-1} \sum_{i|D_i=1} \left(\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i) \right) \right], \tag{2.2}$$

where $\bar{Y}_{d,t} = \sum_{i|D_i=d,T_i=t} Y_{it}/n_{d,t}$ is the sample average outcome among units in treatment group d and time t, and $\widehat{\mu}_{d,t}(x)$ is an estimator of the true, unknown $m_{d,t}(x) \equiv \mathbb{E}[Y_t|D=d,X=x]$,[3] see e.g. Heckman et al. (1997).

The IPW approach proposed by Abadie (2005) avoids directly modelling the outcome evolution and exploits that, under Assumptions 1-3, the ATT can be expressed as

$$\tau = \frac{1}{\mathbb{E}[D]} \mathbb{E}\left[\frac{D - p(X)}{1 - p(X)} (Y_1 - Y_0) \right]$$

when panel data are available, and as

$$\tau = \frac{1}{\mathbb{E}[D]} \mathbb{E}\left[\frac{D - p(X)}{1 - p(X)} \frac{T - \lambda}{\lambda(1-\lambda)} Y \right] \tag{2.3}$$

when repeated cross-section data are available, where $p(X) \equiv \mathbb{P}(D=1|X)$ is the true, unknown propensity score. Abadie's identification results suggest simple two-step estimators for the ATT that do not involve outcome regressions. For instance, when panel data are available, Abadie (2005) proposes the following Horvitz and Thompson (1952) type IPW estimator,

$$\widehat{\tau}^{ipw,p} = \frac{1}{\mathbb{E}_n[D]} \mathbb{E}_n\left[\frac{D - \widehat{\pi}(X)}{1 - \widehat{\pi}(X)} (Y_1 - Y_0) \right], \tag{2.4}$$

where $\widehat{\pi}(x)$ is an estimator of the true, unknown $p(x)$, and for a generic random variable Z, $\mathbb{E}_n[Z] = n^{-1}\sum_{i=1}^n Z_i$; the estimator for the repeated cross-section case is formed using the analogous procedure.

It is important to emphasize that the reliability of ATT estimators based on the OR and the IPW approaches depends on different, non-nested conditions. For the OR approach, the consistency of the ATT estimator (2.2) relies on the estimators of $m_{d,t}(\cdot)$, $\widehat{\mu}_{d,t}(\cdot)$, being correctly specified, whereas the IPW estimator (2.4) relies on the propensity score estimator $\widehat{\pi}(\cdot)$ of $p(\cdot)$ being correctly specified. As so, in practice, it may be hard to "rank" these two approaches in terms of their robustness to model misspecification.

Remark 1 It is common to see practitioners adopting the two-way fixed effects linear regression model

$$Y_{it} = \alpha_1 + \alpha_2 T_i + \alpha_3 D_i + \tau^{fe} (T_i \cdot D_i) + \theta' X_i + \varepsilon_{it}, \tag{2.5}$$

and interpreting estimates of τ^{fe} as estimates of the ATT, see e.g. chapter 5.2 in Angrist and Pischke (2009). Although (2.5) may be perceived as a "natural" specification, it implicitly imposes additional restrictions on the data generating process beyond Assumptions 1-3. More specifically, (2.5) implicitly imposes that

3 In the repeated cross-section case, $m_{d,t}(x) = \mathbb{E}[Y|D=d,T=t,X=x]$. In the next section, we differentiate the notation for the panel data and repeated cross-section case to avoid potential confusions.

$(i)\,\mathbb{E}\left[Y_1(1)-Y_1(0)\,|X,D=1\right]=\tau^{fe}$ *a.s.*, i.e., it assumes homogeneous (in X) treatment effects, and (ii) for $d=0,1$, $\mathbb{E}\left[Y_1-Y_0|X,D=d\right]=\mathbb{E}\left[Y_1-Y_0|D=d\right]$ *a.s.*, i.e., it rules out X-specific trends in both treated and comparison groups.[4] When these additional restrictions are not satisfied, the estimand τ^{fe} is, in general, different from the ATT, and policy evaluation based on it may be misleading. We further illustrate this point using Monte Carlo simulations in Section 4; see also Słoczyński (2018) for related results.

2.2 Doubly robust difference-in-differences estimands

In this section, we argue that instead of choosing between the OR and the IPW approaches, one can combine them to form doubly robust (DR) moments/estimands for the ATT. Here, double robustness means that the resulting estimand identifies the ATT even if either (but not both) the propensity score model or the outcome regression models are misspecified. As so, the DR DID estimand for the ATT shares the strengths of each individual DID method and, at the same time, avoids some of their weaknesses.

Before describing how we exactly combine the OR and the IPW approaches to form our DR DID estimand, we need to introduce some additional notation. Let $\pi(X)$ be an arbitrary model for the true, unknown propensity score. When panel data are available, let $\Delta Y=Y_1-Y_0$ and define $\mu^p_{d,\Delta}(X)\equiv\mu^p_{d,1}(X)-\mu^p_{d,0}(X)$, $\mu^p_{d,t}(x)$ being a model for the true, unknown outcome regression $m^p_{d,t}(x)\equiv\mathbb{E}[Y_t|D=d,X=x]$, $d,t=0,1$. When only repeated cross-section data are available, let $\mu^{rc}_{d,t}(x)$ be an arbitrary model for the true, unknown regression $m^{rc}_{d,t}(x)\equiv\mathbb{E}[Y|D=d,T=t,X=x]$, $d,t=0,1$, and for, $d=0,1$, $\mu^{rc}_{d,Y}(T,X)\equiv T\cdot\mu^{rc}_{d,1}(X)+(1-T)\cdot\mu^{rc}_{d,0}(X)$, and $\mu^{rc}_{d,\Delta}(X)\equiv\mu^{rc}_{d,1}(X)-\mu^{rc}_{d,0}(X)$.

For the case in which panel data are available, we consider the estimand

$$\tau^{dr,p}=\mathbb{E}\left[\left(w^p_1(D)-w^p_0(D,X;\pi)\right)\left(\Delta Y-\mu^p_{0,\Delta}(X)\right)\right], \tag{2.6}$$

where, for a generic g,

$$w^p_1(D)=\frac{D}{\mathbb{E}[D]},\quad\text{and}\quad w^p_0(D,X;g)=\frac{g(X)(1-D)}{1-g(X)}\Big/\mathbb{E}\left[\frac{g(X)(1-D)}{1-g(X)}\right]. \tag{2.7}$$

For the repeated cross-section case, we consider two different estimands,

$$\tau^{dr,rc}_1=\mathbb{E}\left[\left(w^{rc}_1(D,T)-w^{rc}_0(D,T,X;\pi)\right)\left(Y-\mu^{rc}_{0,Y}(T,X)\right)\right], \tag{2.8}$$

and

$$\begin{aligned}\tau^{dr,rc}_2=\tau^{dr,rc}_1&+\left(\mathbb{E}\left[\mu^{rc}_{1,1}(X)-\mu^{rc}_{0,1}(X)\middle|D=1\right]-\mathbb{E}\left[\mu^{rc}_{1,1}(X)-\mu^{rc}_{0,1}(X)\middle|D=1,T=1\right]\right)\\&-\left(\mathbb{E}\left[\mu^{rc}_{1,0}(X)-\mu^{rc}_{0,0}(X)\middle|D=1\right]-\mathbb{E}\left[\mu^{rc}_{1,0}(X)-\mu^{rc}_{0,0}(X)\middle|D=1,T=0\right]\right),\end{aligned} \tag{2.9}$$

4 Note that under Assumptions 1-3, (2.5) suggests that, with probability one, $\mathbb{E}\left[Y_1(1)\,|X,D=1\right]=\alpha_1+\alpha_2+\alpha_3+\tau+\theta'X$, and $\mathbb{E}\left[Y_1(0)\,|X,D=1\right]=\mathbb{E}[Y_0|D=1,X]+(\mathbb{E}[Y_1|D=0,X]-\mathbb{E}[Y_0|D=0,X])=\alpha_1+\alpha_2+\alpha_3+\theta'X$. Point (i) now follows directly. Point (ii) follows from analogous arguments.

where, for a generic g,

$$w_1^{rc}(D,T) = w_{1,1}^{rc}(D,T) - w_{1,0}^{rc}(D,T), \quad \text{and} \quad w_0^{rc}(D,T,X;g) = w_{0,1}^{rc}(D,T,X;g) - w_{0,0}^{rc}(D,T,X;g), \quad (2.10)$$

and, for $t = 0,1$,

$$\begin{aligned} w_{1,t}^{rc}(D,T) &= \frac{D \cdot 1\{T=t\}}{\mathbb{E}[D \cdot 1\{T=t\}]}, \\ w_{0,t}^{rc}(D,T,X;g) &= \frac{g(X)(1-D) \cdot 1\{T=t\}}{1-g(X)} \Big/ \mathbb{E}\left[\frac{g(X)(1-D) \cdot 1\{T=t\}}{1-g(X)}\right]. \end{aligned}$$

Theorem 1 *Let Assumptions 1-3 hold. Then:*

(a) When panel data are available, $\tau^{dr,p} = \tau$ if either (but not necessarily both) $\pi(X) = p(X)$ a.s. or $\mu_\Delta^p(X) = m_{0,1}^p(X) - m_{0,0}^p(X)$ a.s.;

(b) When repeated cross-section data are available, $\tau_1^{dr,rc} = \tau_2^{dr,rc} = \tau$ if either (but not necessarily both) $\pi(X) = p(X)$ a.s. or $\mu_{0,\Delta}^{rc}(X) = m_{0,1}^{rc}(X) - m_{0,0}^{rc}(X)$ a.s..

Theorem 1 states that provided that at least one of the working nuisance models is correctly specified, we can recover the ATT with either panel or repeated cross-section data. Thus, our proposed DR DID estimands are "less demanding" in terms of the researchers' ability to correctly specify models for the nuisance functions than either the OR or the IPW approach.

Given that we consider two different estimands for the case of repeated cross-section, it is interesting to use Theorem 1 to compare them. Given that $\tau_1^{dr,rc}$ does not rely on OR models for the treated group but $\tau_2^{dr,rc}$ does, one could a priori expect that $\tau_1^{dr,rc}$ would be more robust against model misspecification than $\tau_2^{dr,rc}$. Nonetheless, Theorem 1 states that this is not the case as they identify the ATT under the same conditions. At this stage, one may wonder how this is possible. To answer such a query, it suffices to remember that, under the stationarity condition in Assumption 1(b), for any generic integrable and measurable function g, $\mathbb{E}[g(X)|D=1] = \mathbb{E}[g(X)|D=1,T=t]$, $t=0,1$. Given that this holds for any generic function g, it must also hold for $\mu_{1,t}^{rc}(\cdot) - \mu_{0,t}^{rc}(\cdot)$, $t=0,1$, even when $\mu_{d,t}^{rc}(\cdot)$ are misspecified models of $m_{d,t}^{rc}(\cdot)$. Such a result reveals that modeling the OR for the treat group can be "harmless" in terms of identification, provided that these additional models are incorporated into $\tau_1^{dr,rc}$ in an appropriate manner.

2.3 Semiparametric efficiency bound

In the previous subsection, we derived DR moment equations for the ATT under the DID framework and showed that the resulting estimands are more robust against model misspecifications than DID estimands based on either the OR or the IPW approach. In this subsection, we shift our attention from "robustness" to efficiency. More precisely, we calculate the semiparametric efficiency bound for the ATT under Assumptions 1-3 when either panel or repeated cross-section data are available. These results provide the semiparametric analog of the Cramér–Rao lower bound commonly used in fully parametric procedures. As so, they provide a benchmark that

researchers can use to assess whether any given (regular) semiparametric DID estimator for the ATT is fully exploiting the empirical content of Assumptions 1-3.

Let $m^{p}_{0,\Delta}(x) \equiv m^{p}_{0,1}(x) - m^{p}_{0,0}(x)$, and, for $d = 0,1$, $m^{rc}_{d,\Delta}(X) \equiv m^{rc}_{d,1}(X) - m^{rc}_{d,0}(X)$. Recall that $\lambda \equiv \mathbb{P}(T=1)$. Next proposition displays the semiparametric efficiency bound for the ATT when one has access to panel data and when one has access to repeated cross-section data. To simplify exposition, we abstract from additional technical discussions related to the conditions to guarantee quadratic mean differentiability and their implications for the precise definition of efficient influence function ; see, e.g., Chapter 3 of Bickel et al. (1998) for additional details.

Proposition 1 *Let Assumptions 1-3 hold. Then:*

(a) When panel data are available, the efficient influence function for the ATT is

$$\begin{aligned}\eta^{e,p}(Y_1,Y_0,D,X) = \; & w^{p}_{1}(D)\left(m^{p}_{1,\Delta}(X) - m^{p}_{0,\Delta}(X) - \tau\right) \\ & + w^{p}_{1}(D)\left(\Delta Y - m^{p}_{1,\Delta}(X)\right) - w^{p}_{0}(D,X;p)\left(\Delta Y - m^{p}_{0,\Delta}(X)\right), \quad (2.11)\end{aligned}$$

and the semiparametric efficiency bound for all regular estimators for the ATT is

$$\begin{aligned}\mathbb{E}\left[\eta^{e,p}(Y_1,Y_0,D,X)^2\right] = \frac{1}{\mathbb{E}[D]^2}\mathbb{E}\Bigg[& D\left(m^{p}_{1,\Delta}(X) - m^{p}_{0,\Delta}(X) - \tau\right)^2 \\ & + D\left(\Delta Y - m^{p}_{1,\Delta}(X)\right)^2 + \frac{(1-D)\,p(X)^2}{(1-p(X))^2}\left(\Delta Y - m^{p}_{0,\Delta}(X)\right)^2\Bigg]. \quad (2.12)\end{aligned}$$

(b) When only repeated cross-section data are available, the efficient influence function for the ATT is

$$\begin{aligned}\eta^{e,rc}(Y,D,T,X) = \; & \frac{D}{\mathbb{E}[D]}\left(m^{rc}_{1,\Delta}(X) - m^{rc}_{0,\Delta}(X) - \tau\right) \\ & + \left(w^{rc}_{1,1}(D,T)\left(Y - m^{rc}_{1,1}(X)\right) - w^{rc}_{1,0}(D,T)\left(Y - m^{rc}_{1,0}(X)\right)\right) \\ & - \left(w^{rc}_{0,1}(D,T,X;p)\left(Y - m^{rc}_{0,1}(X)\right) - w^{rc}_{0,0}(D,T,X;p)\left(Y - m^{rc}_{0,0}(X)\right)\right), \quad (2.13)\end{aligned}$$

and the semiparametric efficiency bound for all regular estimators for the ATT is

$$\begin{aligned}\mathbb{E}\left[\eta^{e,rc}(Y,D,T,X)^2\right] = \frac{1}{\mathbb{E}[D]^2}\mathbb{E}\Bigg[& D\left(m^{rc}_{1,\Delta}(X) - m^{rc}_{0,\Delta}(X) - \tau\right)^2 \\ & + \frac{DT}{\lambda^2}\left(Y - m^{rc}_{1,1}(X)\right)^2 + \frac{D(1-T)}{(1-\lambda)^2}\left(Y - m^{rc}_{1,0}(X)\right)^2 \\ & + \frac{(1-D)\,p(X)^2\,T}{(1-p(X))^2\,\lambda^2}\left(Y - m^{rc}_{0,1}(X)\right)^2 + \frac{(1-D)\,p(X)^2\,(1-T)}{(1-p(X))^2\,(1-\lambda)^2}\left(Y - m^{rc}_{0,0}(X)\right)^2\Bigg]. \quad (2.14)\end{aligned}$$

It is interesting to compare $\eta^{e,p}(D,X)$ with $\eta^{e,rc}(D,T,X)$. First, note that the first component of their efficient influence functions are analogous to each other, and depends on the true, unknown conditional ATT, $m_{1,\Delta}(X) - m_{0,\Delta}(X)$.[5] The second and third terms in (2.11) and (2.13) are more different from each other. For $\eta^{e,p}$, the availability of panel data implies that Y_1 and Y_0 are observed for all units, and, therefore, we can

5 To avoid excessive notational burden, we supress the "p" and "rc" superscripts unless their omission leads to confusion.

directly reweight $\Delta Y - m_{1,\Delta}(X)$ and $\Delta Y - m_{0,\Delta}(X)$. In contrast, when only repeated cross-section data are available, one observes Y_t only if $T = t$, $t = 0, 1$, and, therefore, the efficient influence function (2.13) depends on different weights for each pair $(D, T) \in \{0,1\}^2$. In this latter case, we also stress the importance of imposing the stationarity condition in Assumption 1(b) when deriving the efficient influence function (2.13) – failing to do so will suggest an "efficiency bound" that is wider than (2.14).

It is also worth mentioning that the efficient influence functions (2.11) and (2.13) depend on the true, unknown, outcome regression functions for the treated group, $m_{1,1}(\cdot)$ and $m_{1,0}(\cdot)$, in an asymmetric manner. On one hand, when panel data are available, by simple manipulation, we can rewrite $\eta^{e,p}$ as

$$\eta^{e,p}(Y_1, Y_0, D, X) = \left(w_1^p(D) - w_0^p(D, X; p)\right)(\Delta Y - m_{0,\Delta}(X)) - w_1^p(D) \cdot \tau,$$

emphasizing that the efficient influence function for the ATT when panel data are available *does not* depend on $m_{1,1}(\cdot)$ and $m_{1,0}(\cdot)$. This is in sharp contrast to the case where only repeated cross-section data are available.

Another interesting question raised by Proposition 1 is whether the semiparametric efficiency bound for the case of repeated cross-section data is larger than the one for the case of panel data. In order to answer this question, we consider the case where T is independent of (Y_1, Y_0, D, X), so that Assumptions $1(a)$ and $1(b)$ are compatible with each other.[6]

Corollary 1 *Let Assumptions 1-3 hold, and assume that T is independent of (Y_1, Y_0, D, X). Then,*

$$\begin{aligned}
&\mathbb{E}\left[\eta^{e,rc}(Y, D, T, X)^2\right] - \mathbb{E}\left[\eta^{e,p}(Y_1, Y_0, D, X)^2\right] \\
&\quad = \frac{1}{\mathbb{E}[D]^2}\mathbb{E}\left[D\left(\sqrt{\frac{1-\lambda}{\lambda}}(Y_1 - m_{1,1}(X)) + \sqrt{\frac{\lambda}{1-\lambda}}(Y_0 - m_{1,0}(X))\right)^2\right. \\
&\qquad \left. + \frac{(1-D)\,p(X)^2}{(1-p(X))^2}\left(\sqrt{\frac{1-\lambda}{\lambda}}(Y_1 - m_{0,1}(X)) + \sqrt{\frac{\lambda}{1-\lambda}}(Y_0 - m_{0,0}(X))\right)^2\right] \geq 0.
\end{aligned}$$

In other words, under the DID framework it is possible to form more efficient estimators for the ATT when panel data are available than when only repeated cross-section data are available. In addition, from Corollary 1, we can also see that the efficiency loss is convex in λ, implying that the loss of efficiency is bigger when the pre and post-treatment sample sizes are more imbalanced. In fact, when

$$\begin{aligned}
&\mathbb{E}\left[D(Y_0 - m_{1,0}(X))^2 + \frac{(1-D)\,p(X)^2}{(1-p(X))^2}(Y_0 - m_{0,0}(X))^2\right] = \\
&\qquad \mathbb{E}\left[D(Y_1 - m_{1,1}(X))^2 + \frac{(1-D)\,p(X)^2}{(1-p(X))^2}(Y_1 - m_{0,1}(X))^2\right], \quad (2.15)
\end{aligned}$$

6 This "restriction" does not affect the semiparametric efficiency bound for the case where only repeated cross-section data are available, as it does not impose additional restrictions on the observed data.

we can show that $\lambda = 0.5$ is optimal . However, when (2.15) does not hold, the optimal λ depends on the data in a more complicated manner, and is given by $\lambda = \tilde{\sigma}_1 / (\tilde{\sigma}_0 + \tilde{\sigma}_1)$, where, for $t = 0, 1$

$$\tilde{\sigma}_t^2 = \mathbb{E}\left[D\left(Y_t - m_{1,t}(X)\right)^2 + \frac{(1-D)\, p(X)^2}{(1-p(X))^2} \left(Y_t - m_{0,t}(X)\right)^2 \right].$$

These results suggest that, in principle, one may benefit from "oversampling" from either the pre or post-treatment period. However, it is, in general, not feasible to know the optimal λ during the design stage, i.e., at the pre-treatment period, since $\tilde{\sigma}_1^2$ depends on the outcome data from the post-treatment period. Thus, if one were to design the DID study with repeated cross-section units, it seems that setting $\lambda = 0.5$ would be a "reasonable" choice.

3 Estimation and inference

In this section, we build on the DR DID estimands in Theorem 1 and the semiparametric efficiency bounds in Proposition 1, and discuss estimation and inference procedures for the ATT in DID designs. Indeed, the moment equations (2.6), (2.8), and (2.9) suggest a simple two-step strategy to estimate the ATT. In the first step, one estimates the true, unknown $p(\cdot)$ using $\pi(\cdot)$, and the true, unknown $m_{d,t}^p(\cdot)$ $(m_{d,t}^{rc}(\cdot))$ using $\mu_{d,t}^p(\cdot)(\mu_{d,t}^{rc}(\cdot))$, $d,t = 0,1$, when panel data (repeated cross-section data) are available. In the second step, one plugs the fitted values of the estimated propensity score and regression models into the sample analogue of $\tau^{dr,p}$, $\tau_1^{dr,rc}$, or $\tau_2^{dr,rc}$.

Although, in principle, one can use semi/non-parametric estimators for both the outcome regressions and the propensity score, see e.g. Heckman et al. (1997), Abadie (2005), Chen et al. (2008) and Rothe and Firpo (2018), in what follows ,we focus our attention on generic parametric first-step estimators. More precisely, we assume that $\pi(x;\gamma^*)$ is a parametric model for $p(x)$, such that π is known up to the finite dimensional pseudo-true parameter γ^*. Analogously, for $d,t = 0,1$, $\mu_{d,t}^p\left(x;\beta_{d,t}^{*,p}\right)$ (and $\mu_{d,t}^{rc}\left(x;\beta_{d,t}^{*,rc}\right)$) is a parametric model for $m_{d,t}^p(x)$ $(m_{d,t}^{rc}(x))$, such that $\mu_{d,t}^p$ $(\mu_{d,t}^{rc})$ is known up to the finite dimensional pseudo-true parameter $\beta_{d,t}^{*,p}$ $(\beta_{d,t}^{*,rc})$. This is perhaps the most popular approach adopted by practitioners, particularly when the available sample size is moderate and/or the dimension of available covariates is high or even moderate, as the "curse of dimensionality" usually prevents one to adopt fully nonparametric procedures.[7]

In the case when panel data are available, our proposed DR DID estimator for the ATT is based on (2.6) and is given by

$$\widehat{\tau}^{dr,p} = \mathbb{E}_n\left[\left(\widehat{w}_1^p(D) - \widehat{w}_0^p(D,X;\widehat{\gamma}) \right) \left(\Delta Y - \mu_{0,\Delta}^p\left(X;\widehat{\beta}_{0,0}^p, \widehat{\beta}_{0,1}^p\right) \right) \right], \tag{3.1}$$

7 Let $g(x)$ be a generic notation for $p(x)$, $m_{d,t}^l(X)$, $d,t = 0,1$, $l = p, rc$. From Newey (1994), Chen et al. (2003), Ai and Chen (2003, 2007, 2012), and Chen et al. (2008), one can see that the use of nonparametric first-step estimators $\widehat{g}(x)$ of $g(x)$ is warranted provided that $\|\widehat{g}(x) - g(x)\|_{\mathscr{H}} = o_p\left(n^{-1/4}\right)$ for a pseudo-metric $\|\cdot\|_{\mathscr{H}}$, $\mathscr{H}$ being a vector space of functions. However, when the dimension of X is moderate or large, as is usually the case in many empirical applications, conditons ensuring that $\|\widehat{g}(x) - g(x)\|_{\mathscr{H}} = o_p\left(n^{-1/4}\right)$ can be rather stringent because of the "curse of dimensionality".

where

$$\widehat{w}_1^p(D) = \frac{D}{\mathbb{E}_n[D]}, \quad \text{and} \quad \widehat{w}_0^p(D,X;\gamma) = \frac{\pi(X;\gamma)(1-D)}{1-\pi(X;\gamma)} \Big/ \mathbb{E}_n\left[\frac{\pi(X;\gamma)(1-D)}{1-\pi(X;\gamma))}\right], \tag{3.2}$$

$\widehat{\gamma}$ is an estimator for the pseudo-true γ^*, $\widehat{\beta}_{0,t}^p$ is an estimator for pseudo-true $\beta_{0,t}^{*,p}$, $t=0,1$, and for a generic β_0 and β_1, $\mu_{0,\Delta}^p(\cdot;\beta_0,\beta_1) = \mu_{0,1}^p(\cdot;\beta_1) - \mu_{0,0}^p(\cdot;\beta_0)$.

When only repeated cross-section data are available, we propose two different DR DID estimators for the ATT. The first one, which is based on (2.8) and can be interpreted as the analogue of $\widehat{\tau}^{dr,p}$, is given by

$$\widehat{\tau}_1^{dr,rc} = \mathbb{E}_n\left[\left(\widehat{w}_1^{rc}(D,T) - \widehat{w}_0^{rc}(D,T,X;\widehat{\gamma})\right)\left(Y - \mu_{0,Y}^{rc}\left(T,X;\widehat{\beta}_{0,0}^{rc},\widehat{\beta}_{0,1}^{rc}\right)\right)\right], \tag{3.3}$$

where $\mu_{0,Y}^{rc}\left(T,\cdot;\beta_{0,0}^{rc},\beta_{0,1}^{rc}\right) = T\cdot\mu_{0,1}^{rc}\left(\cdot;\beta_{0,1}^{rc}\right) + (1-T)\cdot\mu_{0,0}^{rc}\left(\cdot;\beta_{0,0}^{rc}\right)$, $\widehat{\beta}_{d,t}^{rc}$ is an estimator for the pseudo-true $\beta_{d,t}^{*,rc}$, $d,t=0,1$, and the weights $\widehat{w}_1^{rc}(D,T)$ and $\widehat{w}_0^{rc}(D,T,X;\widehat{\gamma})$ are, respectively, defined as the sample analogues of $w_1^{rc}(D,T)$ and $w_0^{rc}(D,T,X;g)$ defined in (2.10), but with $\pi(x;\widehat{\gamma})$ playing the role of g.

The second DR DID estimator for the case of repeated cross-section builds on (2.9) and is given by

$$\begin{aligned}\widehat{\tau}_2^{dr,rc} = \widehat{\tau}_1^{dr,rc} &+ \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \widehat{w}_{1,1}^{rc}(D,T)\right)\left(\mu_{1,1}^{rc}\left(X;\widehat{\beta}_{1,1}^{rc}\right) - \mu_{0,1}^{rc}\left(X;\widehat{\beta}_{0,1}^{rc}\right)\right)\right]\right) \\ &- \left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]} - \widehat{w}_{1,0}^{rc}(D,T)\right)\left(\mu_{1,0}^{rc}\left(X;\widehat{\beta}_{1,0}^{rc}\right) - \mu_{0,0}^{rc}\left(X;\widehat{\beta}_{0,0}^{rc}\right)\right)\right]\right),\end{aligned} \tag{3.4}$$

where $\mu_{d,\Delta}^{rc}\left(\cdot;\beta_{d,1}^{rc},\beta_{d,0}^{rc}\right) = \mu_{d,1}^{rc}\left(\cdot;\beta_{d,1}^{rc}\right) - \mu_{d,0}^{rc}\left(\cdot;\beta_{d,0}^{rc}\right)$, and the weights $\widehat{w}_{1,t}^{rc}(D,T)$ and $\widehat{w}_{0,t}^{rc}(D,T,X;\widehat{\gamma})$ are, respectively, defined as the sample analogues of $w_{1,t}^{rc}(D,T)$ and $w_{0,t}^{rc}(D,T,X;g)$, $t=0,1$, defined below (2.10), but with $\pi(x;\widehat{\gamma})$ playing the role of g.

As we show in the Appendix A, it is relatively straightforward to derive the asymptotic properties of $\widehat{\tau}^{dr,p}$, $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$ using generic first-step estimators that satisfy some relatively weak, high-level conditions; see Theorems A.1 and A.2 in Appendix A. Indeed, Theorem A.1 indicates that $\widehat{\tau}^{dr,p}$ is doubly robust, and also locally semiparametrically efficient, i.e., its asymptotic variance achieves the semiparametric efficiency bound when the working models for the nuisance functions are correctly specified. Theorem A.2 also indicates that both $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$ are doubly robust when repeated cross-section data are available. However, Theorem A.2 also highlights that $\widehat{\tau}_2^{dr,rc}$ is locally semiparametrically efficient, whereas $\widehat{\tau}_1^{dr,rc}$ is not. In other words, when repeated cross-section data are available, $\widehat{\tau}_2^{dr,rc}$ tends to have more attractive properties than $\widehat{\tau}_1^{dr,rc}$, regardless of the first-step estimators used.

Although the results in Theorems A.1 and A.2 accommodate a variety of different first-step estimators, in practice, one still needs to choose a particular estimation procedure to be implemented. In what follows, we attempt to provide some guidance on the choice of first-step estimators with the goal of further improving the (generic) DR DID estimators. We are particularly interested in forming DR DID estimators that are not only doubly robust in terms of consistency—like described above—but also doubly robust for inference, i.e., their asymptotic linear representation is also doubly robust. The attractiveness of forming estimators that are

DR for inference is that there is no estimation effect from first-step estimators, which, in turn, implies that the asymptotic variance of the results DR DID estimator for the ATT is invariant to which working models for the nuisance functions are correctly specified. In practice, this usually translates to simpler and more stable inference procedures.

To derive these improved DR DID estimators, we focus on the case where a researcher is comfortable with linear regression working models for the outcome of interest, a logistic working model for the propensity score, and with covariates X entering all the nuisance models in a symmetric manner. Although these modelling conditions are more stringent than those allowed by our generic DR DID estimators discussed in Appendix A, they are much weaker than those implicitly imposed in the TWFE specification (2.5), and can be seen as the default choice in many applications. Hence, these extra assumptions can be seen as a reasonable compromise to get further improved DR DID estimators that are also computationally tractable and easy to implement in practice.

3.1 Improved DR DID estimators when panel data are available

As discussed above, we consider the following working models for the nuisance functions:

$$\pi(X,\gamma) = \Lambda\left(X'\gamma\right) \equiv \frac{\exp\left(X'\gamma\right)}{1+\exp\left(X'\gamma\right)}, \text{ and } \mu^{p}_{0,\Delta}\left(X;\beta^{p}_{0,1},\beta^{p}_{0,1}\right) = \mu^{lin,p}_{0,\Delta}\left(X;\beta^{p}_{0,\Delta}\right) \equiv X'\beta^{p}_{0,\Delta}. \tag{3.5}$$

Our proposed improved DR DID estimator is given by the three-step estimator

$$\widehat{\tau}^{dr,p}_{imp} = \mathbb{E}_n\left[\left(\widehat{w}^{p}_{1}(D) - \widehat{w}^{p}_{0}\left(D,X;\widehat{\gamma}^{ipt}\right)\right)\left(\Delta Y - \mu^{lin,p}_{0,\Delta}\left(X;\widehat{\beta}^{wls,p}_{0,\Delta}\right)\right)\right],$$

where the first two-steps consist of computing

$$\begin{aligned}
\widehat{\gamma}^{ipt} &= \arg\max_{\gamma\in\Gamma}\mathbb{E}_n\left[DX'\gamma - (1-D)\exp\left(X'\gamma\right)\right], \\
\widehat{\beta}^{wls,p}_{0,\Delta} &= \arg\min_{b\in\Theta}\mathbb{E}_n\left[\frac{\Lambda\left(X'\widehat{\gamma}^{ipt}\right)}{1-\Lambda\left(X'\widehat{\gamma}^{ipt}\right)}\left(\Delta Y - X'b\right)^2 \middle| D=0\right],
\end{aligned}$$

while in the third and last step, one plugs the fitted values of the working models (3.5) into the sample analogue of $\tau^{dr,p}$. Here, note that $\widehat{\gamma}^{ipt}$ is the inverse probability tilting estimator proposed by Graham et al. (2012) in a different context, while $\widehat{\beta}^{wls,p}_{0,\Delta}$ is simply the weighted least squares estimator for $\beta^{*,p}_{0,\Delta}$.

At this point, one may wonder why we use the estimators $\widehat{\gamma}^{ipt}$ and $\widehat{\beta}^{wls,p}_{0,\Delta}$ instead of other available alternatives. To answer such a query, recall that the main goal here is to propose DID estimators for the ATT that are not only DR consistent but also DR for inference, i.e., the exact form of their asymptotic variance does not depend on which working models for the nuisance functions are correctly specified. As it turns out, the key to obtain DID estimators for the ATT that are also DR for inference is to choose first-step estimators for the nuisance parameters, say $\widehat{\gamma}$ and $\widehat{\beta}$, such that the limiting distribution of the resulting DR DID estimator $\widehat{\tau}^{dr,p}$ is equivalent to that of the infeasible DR DID estimator that uses the pseudo-true values of $\widehat{\gamma}$ and $\widehat{\beta}$, say γ^* and β^*. In a more precise manner, in order to get DID estimators that are DR for inference, we need to guarantee that there will be

no estimation effect from the first stage.

In Appendix A, we show that the estimation effect associated with using generic first-step estimators $\widehat{\gamma}$ and $\widehat{\beta}$ is given by $\eta^p_{est}(W;\gamma^*,\beta^*)$ as defined in (A.2). By paying closer attention to the exact form of $\eta^p_{est}(W;\gamma^*,\beta^*)$, one can see that if

$$\begin{aligned}
\mathbb{E}\left[\left(w^p_1 - w^p_0(\gamma^*)\right)\cdot \dot{\mu}^p_{0,\Delta}(\beta^*)\right] &= 0,\\
\mathbb{E}\left[\frac{(1-D)}{(1-\pi(X;\gamma^*))^2}\left(\Delta Y - \mu^p_{0,\Delta}(\beta^*)\right)\cdot \dot{\pi}(\gamma^*)\right] &= 0,\\
\mathbb{E}\left[w^p_0(\gamma^*)\cdot\left(\Delta Y - \mu^p_{0,\Delta}(\beta^*)\right)\right] &= 0,
\end{aligned} \tag{3.6}$$

then there will be no estimation effect from the first stage. As the first component of X is assumed to be constant and we adopt the working models (3.5), it follows that (3.6) reduces to

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\exp(X'\gamma^*)(1-D)}{\mathbb{E}[\exp(X'\gamma^*)(1-D)]}\right)X\right] &= 0,\\
\mathbb{E}\left[\exp(X'\gamma^*)\left(\Delta Y - \mu^{lin,p}_{0,\Delta}(X;\beta^*_{0,\Delta})\right)X \,\middle|\, D=0\right] &= 0.
\end{aligned}$$

However, as $n\to\infty$, these two vectors of moment conditions follow from the first-order conditions of the optimization problems associated with $\widehat{\gamma}^{ipt}$ and $\widehat{\beta}^{wls,p}_{0,\Delta}$, respectively, even when these working models are misspecified. Hence, by using $\widehat{\gamma}^{ipt}$ and $\widehat{\beta}^{wls,p}_{0,\Delta}$, we guarantee that $\widehat{\tau}^{dr,p}_{imp}$ is doubly robust for inference as there is no estimation effect from replacing the pseudo-true parameters $\gamma^{*,ipt}$ and $\beta^{*,wls,p}_{0,\Delta}$ with their estimators $\widehat{\gamma}^{ipt}$ and $\widehat{\beta}^{wls,p}_{0,\Delta}$, respectively.

The next theorem formalizes this discussion. Define

$$\begin{aligned}
\widehat{\tau}^{dr,p}_{imp} &= \mathbb{E}_n\left[\left(w^p_1(D) - w^p_0(D,X;\widehat{\gamma}^{ipt})\right)\left(\Delta Y - \mu^{lin,p}_{0,\Delta}\left(X;\widehat{\beta}^{wls,p}_{0,\Delta}\right)\right)\right],\\
\tau^{dr,p}_{imp} &= \mathbb{E}\left[\left(w^p_1(D) - w^p_0(D,X;\gamma^{*,ipt})\right)\left(\Delta Y - \mu^{lin,p}_{0,\Delta}\left(X;\beta^{*,wls,p}_{0,\Delta}\right)\right)\right],
\end{aligned} \tag{3.7}$$

and let

$$\eta^{dr,p}_{imp}\left(W;\gamma^{*,ipt},\beta^{*,wls,p}_{0,\Delta},\tau^{dr,p}_{imp}\right) = \left(w^p_1(D) - w^p_0(D,X;\gamma^{*,ipt})\right)\left(\Delta Y - \mu^{lin,p}_{0,\Delta}\left(X;\beta^{*,wls,p}_{0,\Delta}\right)\right) - w^p_1(D)\cdot\tau^{dr,p}_{imp}.$$

Theorem 2 *Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix A hold, and that the working nuisance models (3.5) are adopted. Then,*

(a) If either $\Lambda\left(X'\gamma^{*,ipt}\right) = p(X)$ *a.s or* $X'\beta^{*,wls,p}_{0,\Delta} = m^p_{0,\Delta}(X)$ *a.s., then, as* $n\to\infty$,

$$\widehat{\tau}^{dr,p}_{imp} \xrightarrow{p} \tau,$$

and

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}^{dr,p}_{imp} - \tau^{dr,p}_{imp}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta^{dr,p}_{imp}\left(W;\gamma^{*,ipt},\beta^{*,wls,p}_{0,\Delta},\tau^{dr,p}_{imp}\right) + o_p(1)\\
&\xrightarrow{d} N\left(0,V^p_{imp}\right),
\end{aligned}$$

where $V^p_{imp} = \mathbb{E}\left[\eta^{dr,p}_{imp}\left(W;\gamma^{*,ipt},\beta^{*,wls,p}_{0,\Delta},\tau^{dr,p}_{imp}\right)^2\right]$.

(b) If both $\Lambda\left(X'\gamma^{*,ipt}\right) = p(X)$ *a.s and* $X'\beta_{0,\Delta}^{*,wls,p} = m_{0,\Delta}^{p}(X)$ *a.s., then* $\eta_{imp}^{dr,p}\left(W;\gamma^{*,ipt},\beta_{0,\Delta}^{*,wls,p},\tau_{imp}^{dr,p}\right) = \eta^{e,p}(Y_1,Y_0,D,X)$ *a.s. and* V_{imp}^{p} *is equal to the semiparametrically efficiency bound (2.12).*

Part (a) of Theorem 2 generalizes the cross-section results of Vermeulen and Vansteelandt (2015) to the DID framework. It states that the proposed DR DID estimators for the ATT, $\widehat{\tau}_{imp}^{dr,p}$, is doubly robust, $\sqrt{n}$-consistent and asymptotically normal. It also states that the exact form of V_{imp}^{p} does not depend on which working models are correctly specified, implying that $\widehat{\tau}_{imp}^{dr,p}$ is doubly robust not only in terms of consistency but also terms of inference. An important consequence of this DR-for-inference property is that it allows one to treat the summands of $\widehat{\tau}_{imp}^{dr,p}$ as if they were independent and identically distributed, and, therefore, estimate V_{imp}^{p} by

$$\widehat{V}_{imp}^{p} = \mathbb{E}_n\left[\eta_{imp}^{dr,p}\left(W;\widehat{\gamma}^{ipt},\widehat{\beta}_{0,\Delta}^{wls,p},\widehat{\tau}_{imp}^{dr,p}\right)^2\right].$$

Part (b) of Theorem 2 indicates that $\widehat{\tau}_{imp}^{dr,p}$ is semiparametrically efficient when the working model for the propensity score, and the working models for the outcome regression for the comparison units are correctly specified.

Remark 2 From the discussion above, it may be natural to directly use the moment conditions (3.6) to form (generic) nonlinear generalized method of moment (GMM) estimators for γ and β. However, it is important to emphasize that to justify the use of such estimation procedure, one must at least establish the local identification of the pseudo-true parameters, which, in turn, requires the matrix of derivatives of (3.6) having full column rank. Importantly, such a condition may not hold for some working models. This is particularly the case when one adopts the working models (3.5) and both specifications are correctly specified. Thus, care must be taken when one attempts to use alternative, more general estimation techniques to generalize the DR inference results discussed above.

Remark 3 As discussed in Appendix A of Graham et al. (2012), it is possible to use alternative specifications for the propensity score, e.g., a probit working model. However, when one deviates from the logit specification, the optimization algorithm involved to estimate the nuisance parameters γ tends to be more computationally demanding, as it involves numerical integration. As discussed above, $\widehat{\gamma}^{ipt}$ clearly avoids such complications.

3.2 Improved DR DID estimators when repeated cross-section data are available

In this section, we turn our attention to our proposed improved DR DID estimators for the ATT when only repeated cross-section data are available. Similar to the panel data case, we consider the case where a researcher is comfortable with the following specifications,

$$\pi(X,\gamma) = \Lambda\left(X'\gamma\right) \equiv \frac{\exp(X'\gamma)}{1+\exp(X'\gamma)}, \text{ and } \mu_{d,t}^{rc}\left(X;\beta_{d,t}^{rc}\right) = \mu_{d,t}^{lin,rc}\left(X;\beta_{d,t}^{rc}\right) \equiv X'\beta_{d,t}^{rc}. \tag{3.8}$$

We consider two improved DR DID estimators based on (2.8) and (2.9), namely

$$\widehat{\tau}_{1,imp}^{dr,rc} = \mathbb{E}_n\left[\left(\widehat{w}_1^{rc}(D,T) - \widehat{w}_0^{rc}\left(D,T,X;\widehat{\gamma}^{ipt}\right)\right)\left(Y - \mu_{0,Y}^{lin,rc}\left(X;\widehat{\beta}_{0,1}^{wls,rc},\widehat{\beta}_{0,0}^{wls,rc}\right)\right)\right], \tag{3.9}$$

and

$$\widehat{\tau}_{2,imp}^{dr,rc} = \widehat{\tau}_{1,imp}^{dr,rc} + \left(\mathbb{E}_n \left[\left(\frac{D}{\mathbb{E}_n[D]} - \widehat{w}_{1,1}^{rc}(D,T) \right) \left(\mu_{1,1}^{rc}\left(X;\widehat{\beta}_{1,1}^{ols,rc}\right) - \mu_{0,1}^{rc}\left(X;\widehat{\beta}_{0,1}^{wls,rc}\right) \right) \right] \right) - \left(\mathbb{E}_n \left[\left(\frac{D}{\mathbb{E}_n[D]} - \widehat{w}_{1,0}^{rc}(D,T) \right) \left(\mu_{1,0}^{rc}\left(X;\widehat{\beta}_{1,0}^{ols,rc}\right) - \mu_{0,0}^{rc}\left(X;\widehat{\beta}_{0,0}^{wls,rc}\right) \right) \right] \right), \quad (3.10)$$

where

$$\begin{aligned} \widehat{\gamma}^{ipt} &= \arg\max_{\gamma\in\Gamma} \mathbb{E}_n\left[DX'\gamma - (1-D)\exp\left(X'\gamma\right)\right], \\ \widehat{\beta}_{0,t}^{wls,rc} &= \arg\min_{b\in\Theta} \mathbb{E}_n\left[\frac{\Lambda\left(X'\widehat{\gamma}^{ipt}\right)}{1-\Lambda\left(X'\widehat{\gamma}^{ipt}\right)} \left(Y-X'b\right)^2 \middle| D=0, T=t \right], \\ \widehat{\beta}_{1,t}^{ols,rc} &= \arg\min_{b\in\Theta} \mathbb{E}_n\left[\left(Y-X'b\right)^2 \middle| D=1, T=t \right]. \end{aligned}$$

Here, note that $\widehat{\tau}_{1,imp}^{dr,rc}$ does not rely on OR models for the treated group while $\widehat{\tau}_{2,imp}^{dr,rc}$ does. In addition, when one compares $\widehat{\tau}_{1,imp}^{dr,rc}$ and $\widehat{\tau}_{2,imp}^{dr,rc}$ with $\widehat{\tau}_{imp}^{dr,p}$, it is evident that the latter relies on a single OR model since we observe Y_1 and Y_0 for all units; when only repeated cross-section data are available, one needs to model the OR in each time period (and each treatment group). Another interesting feature worth mentioning is that we estimate the OR parameters for the treated group via ordinary least squares, whereas we estimate the OR parameters for the control group with weighted least squares. This follows from the fact that estimating the pseudo-true parameters $\beta_{1,t}^{*,rc}$, $t=0,1$, does not lead to any estimation effect, and therefore one can choose her favorite estimation method. Given this observation and the linear specification in (3.8), we find it natural to estimate $\beta_{1,t}^{*,rc}$, $t=0,1$, via OLS as this is the most widespread estimation procedure adopted by practitioners.

Let

$$\tau_{imp}^{dr,rc} = \mathbb{E}\left[\left(w_1^{rc}(D,T) - w_0^{rc}\left(D,T,X;\gamma^{*,ipt}\right) \right) \left(Y - \mu_{0,Y}^{lin,rc}\left(T,X;\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}\right) \right) \right]$$

and for $\beta_{imp}^{*,rc} = \left(\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc},\beta_{1,1}^{*,ols,rc},\beta_{1,0}^{*,ols,rc}\right)$, define

$$\begin{aligned} \eta_{1,imp}^{dr,rc}\left(W;\gamma^{*,ipt},\beta_{imp}^{*,rc}\right) &= \eta_1^{rc,1}(W;\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}) - \eta_0^{rc,1}(W;\gamma^{*,ipt},\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}), \\ \eta_{2,imp}^{dr,rc}\left(W;\gamma^{*,ipt},\beta_{imp}^{*,rc}\right) &= \eta_1^{rc,2}(W;\beta_{imp}^{*,rc}) - \eta_0^{rc,2}(W;\gamma^{*,ipt},\beta_{0,1}^{*,wls,rc},\beta_{0,0}^{*,wls,rc}), \end{aligned}$$

where $\eta_1^{rc,1}$, $\eta_0^{rc,1}$, $\eta_1^{rc,2}$, and $\eta_0^{rc,2}$ are defined as in (B.1)-(B.4) in the Appendix B.

Next theorem states that $\widehat{\tau}_{1,imp}^{dr,rc}$ and $\widehat{\tau}_{2,imp}^{dr,rc}$ are not only doubly robust consistent but also doubly robust for inference. Furthermore, it states that $\widehat{\tau}_{2,imp}^{dr,rc}$ is locally semiparametrically efficient, whereas $\widehat{\tau}_{1,imp}^{dr,rc}$ is not.

Theorem 3 *Let $n = n_1 + n_0$, where n_1 and n_0 are the sample sizes of the post-treatment and pre-treatment periods, respectively. Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix A hold, that $n_1/n \xrightarrow{p} \lambda \in (0,1)$ as $n_0, n_1 \to \infty$, and that the working nuisance models (3.8) are adopted. Then,*

(a) If either $\Lambda\left(X'\gamma^{*,ipt}\right) = p(X)$ a.s or $X'\beta_{0,1}^{*,wls,rc} - X'\beta_{0,0}^{*,wls,rc} = m_{0,\Delta}^{rc}(X)$ a.s., then, for $j=1,2$, as $n\to\infty$,

$$\widehat{\tau}_{j,imp}^{dr,rc} \xrightarrow{p} \tau,$$

and

$$\begin{aligned}\sqrt{n}(\widehat{\tau}^{dr,rc}_{j,imp}-\tau) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta^{dr,rc}_{j,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right)+o_p(1)\\ &\xrightarrow{d} N\left(0,V^{rc}_{j,imp}\right),\end{aligned}$$

where $V^{rc}_{j,imp}=\mathbb{E}\left[\eta^{dr,rc}_{j,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right)^2\right]$.

(b) Suppose that $\Lambda\left(X'\gamma^{*,ipt}\right)=p(X)$ *a.s and, for all* $(d,t)\in\{0,1\}^2$, $X'\beta^{*,wls,rc}_{d,t}=m^{rc}_{d,t}(X)$ *a.s.. Then,* $\eta^{dr,rc}_{2,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right)=\eta^{e,rc}(Y,D,T,X)$ *a.s., and* $V^{rc}_{2,imp}$ *is equal to the semiparametrically efficiency bound (2.14). On the other hand,* $V^{rc}_{1,imp}$ *does not attain the semiparametric efficiency bound.*

In other words, Theorem 3 states that both $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ are doubly robust for the ATT, $\sqrt{n}$-consistent and asymptotically normal. Similar to the panel data case, the exact form of the $V^{rc}_{j,imp}$, $j=1,2$, does not depend on which working models are correctly specified, implying that both $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ are also doubly robust in terms of inference.

Part (b) of Theorem 3 indicates that $\widehat{\tau}^{dr,rc}_{2,imp}$ is semiparametrically efficient when the working model for the propensity score, and all working models for the outcome regressions, for both treated and comparison units, are correctly specified. When compared to Theorem 2(b), it is evident that such a requirement is much stronger than when panel data are available. Part (b) of Theorem 3 also indicates that, in general, $\widehat{\tau}^{dr,rc}_{1,imp}$ is not locally semiparametrically efficient. As so, we argue that, in practice, one should favor $\widehat{\tau}^{dr,rc}_{2,imp}$ with respect to $\widehat{\tau}^{dr,rc}_{1,imp}$, as both estimators are doubly robust in terms of consistency and inference, but the former is locally semiparametrically efficiency whereas the latter is not.

We conclude this section by providing a precise characterization of the efficiency loss associated with using $\widehat{\tau}^{dr,rc}_{1,imp}$ instead of $\widehat{\tau}^{dr,rc}_{2,imp}$ when all working models are correctly specified. Here, our main goal is to illustrate that by using an estimator that attempts to mimic the panel data setup and that does not explicitly exploit the stationarity condition in Assumption 1(b), one may incur in substantial efficiency loss. As so, we argue that, in practice, one should favor estimators based on the DR moment (2.9)—such as $\widehat{\tau}^{dr,rc}_{2,imp}$—with respect to estimators based on the DR moment (2.8)—such as $\widehat{\tau}^{dr,rc}_{1,imp}$.

Corollary 2 *Suppose the assumptions in Theorem 3 hold. Furthermore, assume that* $\Lambda\left(X'\gamma^{*,ipt}\right)=p(X)$ *a.s and, for all* $(d,t)\in\{0,1\}^2$, $X'\beta^{*,wls,rc}_{d,t}=m^{rc}_{d,t}(X)$ *a.s.. Then,*

$$V^{rc}_{1,imp}-V^{rc}_{2,imp}=\mathbb{E}[D]^{-1}\cdot Var\left[\sqrt{\frac{1-\lambda}{\lambda}}\left(m^{rc}_{1,1}(X)-m^{rc}_{0,1}(X)\right)+\sqrt{\frac{\lambda}{1-\lambda}}\left(m^{rc}_{1,0}(X)-m^{rc}_{0,0}(X)\right)\middle| D=1\right]\geq 0.$$

Remark 4 We stress that the result in Corollary 2 does not depend on the fact that one is using the specifications in (3.8). As we show in its proof, such a result remains true provided that the (generic) first-step estimators for the nuisance functions are correctly specified. Thus, Corollary 2 quantifies the loss of efficiency associated with

using estimators based on $\tau_1^{dr,rc}$ as defined in (2.8)—which includes the estimator proposed by Zimmert (2019)—instead of using estimators based on $\tau_2^{dr,rc}$ as defined in (2.9). Given that this loss of efficiency is usually strictly positive, estimators based on $\tau_1^{dr,rc}$ are not, in general, semiparametrically efficient. As we show in the next section via Monte Carlo simulations, this loss of efficiency can be large.

4 Monte Carlo simulation study

In this section, we conduct a series of Monte Carlo experiments in order to study the finite sample properties of our proposed DR DID estimators. When panel data are available, we compare our proposed DR DID estimators $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ given in (3.1) and (3.7), respectively, to the OR DID estimator (2.2), the Horvitz and Thompson (1952) type IPW estimator (2.4), and the TWFE regression model (2.5). Given that the weights of the IPW estimator (2.4) are not normalized to sum up to one, $\widehat{\tau}^{ipw,p}$ can be unstable particularly when propensity score estimates are relatively close to one. To assess the role played by the weights, we also consider the Hájek (1971) type IPW estimator for the ATT

$$\widehat{\tau}^{ipw,p}_{std} = \mathbb{E}_n\left[\left(\widehat{w}^p_1(D) - \widehat{w}^p_0(D,X;\widehat{\gamma})\right)(Y_1 - Y_0)\right], \tag{4.1}$$

where the weights $\widehat{w}^p_1(D)$ and $\widehat{w}^p_0(D,X;\widehat{\gamma})$ are given by (3.2) and are normalized to sum up to one.

When only repeated cross-section data are available, we compare our proposed DR DID estimators $\widehat{\tau}^{dr,rc}_1$ and $\widehat{\tau}^{dr,rc}_2$ given in (3.3) and (3.4), and their further improved versions $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ given in (3.9) and (3.10), to the OR DID estimator (2.2), the plug-in IPW estimator based on (2.3), and the TWFE regression model (2.5). As in the case of panel data, we also consider the Hájek (1971) type IPW estimator for the ATT

$$\widehat{\tau}^{ipw,rc}_{std} = \mathbb{E}_n\left[\left(\widehat{w}^{rc}_1(D,T) - \widehat{w}^{rc}_0(D,T,X;\widehat{\gamma})\right)Y\right], \tag{4.2}$$

where the weights are the same as those in $\widehat{\tau}^{dr,rc}_1$.

In all simulation exercises, we consider a logistic propensity score working model and a linear regression working model for the outcome evolution. All observed covariates enter the working models linearly. With the exception of $\widehat{\tau}^{dr,p}_{imp}$, $\widehat{\tau}^{dr,rc}_{j,imp}$, $j = 1,2$, where we use the estimation methods proposed in Section 3.1 and in Section 3.2, the OR models are estimated using ordinary least squares, and the propensity score working model is estimated using maximum likelihood estimation. When panel data are available, we consider OR models for ΔY instead of OR models for Y_0 and Y_1 separately.

We consider sample size n equal to 1000. For each design, we conduct 10,000 Monte Carlo simulations. We compare the various DID estimators for the ATT in terms of average bias, median bias, root mean square error (RMSE), empirical 95% coverage probability, the average length of a 95% confidence interval, and the average of their plug-in estimator for the asymptotic variance. The confidence intervals are based on the normal approximation, with the asymptotic variances being estimated by their sample analogues. We also compute the semiparametric efficiency bound under each design to allow one to assess the potential loss of efficiency/accuracy

associated with using inefficient DID estimators for the ATT.

4.1 Simulation 1: panel data are available

We first discuss the case where panel data are available. For a generic $W = (W_1, W_2, W_3, W_4)'$, let

$$\begin{aligned} f_{reg}(W) &= 210 + 27.4 \cdot W_1 + 13.7 \cdot (W_2 + W_3 + W_4), \\ f_{ps}(W) &= 0.75 \cdot (-W_1 + 0.5 \cdot W_2 - 0.25 \cdot W_3 - 0.1 \cdot W_4). \end{aligned}$$

Let $\mathbf{X} = (X_1, X_2, \ X_3, X_4)'$ be distributed as $N(0, I_4)$, and I_4 be the 4×4 identity matrix. For $j = 1,2,3,4$, let $Z_j = \left(\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right]\right) / \sqrt{Var\left(\tilde{Z}\right)}$, where $\tilde{Z}_1 = \exp(0.5X_1)$, $\tilde{Z}_2 = 10 + X_2/(1 + \exp(X_1))$, $\tilde{Z}_3 = \left(0.6 + X_1 X_3/25\right)^3$ and $\tilde{Z}_4 = (20 + X_2 + X_4)^2$.

Building on Kang and Schafer (2007), we consider the following data generating processes (DGPs):

$$\begin{aligned}
&DGP1. \quad Y_0(0) = f_{reg}(Z) + v(Z,D) + \varepsilon_0, && Y_1(d) = 2 \cdot f_{reg}(Z) + v(Z,D) + \varepsilon_1(d), \ d = 0,1, \\
&\qquad p(Z) = \frac{\exp(f_{ps}(Z))}{1 + \exp(f_{ps}(Z))}, && D = 1\{p(Z) \geq U\}; \\
&DGP2. \quad Y_0(0) = f_{reg}(Z) + v(Z,D) + \varepsilon_0, && Y_1(d) = 2 \cdot f_{reg}(Z) + v(Z,D) + \varepsilon_1(d), \ d = 0,1, \\
&\qquad p(X) = \frac{\exp(f_{ps}(X))}{1 + \exp(f_{ps}(X))}, && D = 1\{p(X) \geq U\}; \\
&DGP3. \quad Y_0(0) = f_{reg}(X) + v(X,D) + \varepsilon_0, && Y_1(d) = 2 \cdot f_{reg}(X) + v(X,D) + \varepsilon_1(d), \ d = 0,1, \\
&\qquad p(Z) = \frac{\exp(f_{ps}(Z))}{1 + \exp(f_{ps}(Z))}, && D = 1\{p(Z) \geq U\}; \\
&DGP4. \quad Y_0(0) = f_{reg}(X) + v(X,D) + \varepsilon_0, && Y_1(d) = 2 \cdot f_{reg}(X) + v(X,D) + \varepsilon_1(d), \ d = 0,1, \\
&\qquad p(X) = \frac{\exp(f_{ps}(X))}{1 + \exp(f_{ps}(X))}, && D = 1\{p(X) \geq U\},
\end{aligned}$$

where $\varepsilon_0, \varepsilon_1(d), d = 0,1$ are independent standard normal random variables, U is an independent standard uniform random variable, and for a generic W, $v(W,D)$ is an independent normal random variable with mean $D \cdot f_{reg}(W)$ and variance one. The available data are $\{Y_{0,i}, Y_{1,i}, D_i, Z_i\}_{i=1}^n$, where $Y_0 = Y_0(0)$, and $Y_1 = DY_1(1) + (1-D)Y_1(0)$. In the aforementioned DGPs, the true ATT is zero, and v plays the role of time-invariant unobserved heterogeneity.

Given that we focus on the empirically relevant setting where the observed covariates Z enter all working models linearly, it is clear that in $DPG1$, both propensity score (PS) and OR working models are correctly specified. In $DGP2$, only the OR working model is correctly specified, whereas in $DGP3$ only the PS working model is correctly specified. In $DGP4$, all working models are misspecified. The simulation results are presented in Table 1.

First, note that the TWFE estimator $\widehat{\tau}^{fe}$ is severely biased and its confidence interval for the ATT has almost zero coverage in all analyzed DGPs. These results should not be unexpected, because, as discussed in Remark 1, $\widehat{\tau}^{fe}$ implicitly rules out covariate-specific trends, and when these are relevant, like in the considered DGPs, the

Table 1: Monte Carlo results under designs $DGP1 - DGP4$ with panel data. Sample size $n = 1{,}000$.

	DGP1: OR correct, PS correct Semiparametric Efficiency Bound: 11.1						DGP2: OR correct, PS incorrect Semiparametric Efficiency Bound: 11.6					
	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL
$\widehat{\tau}^{fe}$	-20.952	-20.965	21.123	6392.2	0.000	9.906	-19.286	-19.287	19.468	6640.3	0.000	10.095
$\widehat{\tau}^{reg}$	-0.001	-0.001	0.100	10.2	0.950	0.396	-0.001	-0.001	0.100	10.1	0.949	0.394
$\widehat{\tau}^{ipw,p}$	0.026	0.195	2.774	8078.0	0.952	10.441	2.010	2.054	3.298	7048.3	0.838	9.819
$\widehat{\tau}^{ipw,p}_{std}$	0.008	-0.013	1.132	1286.4	0.948	4.309	-0.794	-0.798	1.225	891.7	0.856	3.623
$\widehat{\tau}^{dr,p}$	-0.001	0.000	0.106	11.1	0.947	0.412	-0.001	-0.002	0.104	10.7	0.947	0.404
$\widehat{\tau}^{dr,p}_{imp}$	-0.001	0.000	0.106	10.9	0.945	0.409	-0.001	-0.001	0.104	10.6	0.945	0.404
	DGP3: OR incorrect, PS correct Semiparametric Efficiency Bound: 11.1						DGP4: OR incorrect, PS incorrect Semiparametric Efficiency Bound: 11.6					
	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL
$\widehat{\tau}^{fe}$	-13.170	-13.194	13.364	12687.9	0.004	13.960	-16.385	-16.393	16.538	13160.7	0.000	14.217
$\widehat{\tau}^{reg}$	-1.384	-1.365	1.868	1514.4	0.800	4.816	-5.204	-5.171	5.364	1666.6	0.015	5.053
$\widehat{\tau}^{ipw,p}$	0.011	0.158	3.198	10062.5	0.947	11.777	-1.085	-1.017	2.656	6151.4	0.949	9.308
$\widehat{\tau}^{ipw,p}_{std}$	-0.030	-0.032	1.427	1988.0	0.945	5.484	-3.954	-3.949	4.215	2156.5	0.228	5.717
$\widehat{\tau}^{dr,p}$	-0.051	-0.046	1.214	1400.9	0.942	4.613	-3.188	-3.183	3.454	1704.9	0.308	5.075
$\widehat{\tau}^{dr,p}_{imp}$	-0.071	-0.064	1.015	971.2	0.942	3.858	-2.529	-2.514	2.720	970.1	0.274	3.856

Notes: Simulations based on 10,000 Monte Carlo experiments. $\widehat{\tau}^{fe}$ is the TWFE outcome regression estimator of τ^{fe} in (2.5), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,p}$ is the IPW DID estimator (2.4), $\widehat{\tau}^{ipw,p}_{std}$ is the standardized IPW DID estimator (4.1), $\widehat{\tau}^{dr,p}$ is our proposed DR DID estimator (3.1), and $\widehat{\tau}^{dr,p}_{imp}$ is our proposed DR DID estimator (3.7). We use a linear OR working model and a logistic PS working model, where the unknown parameters are estimated via OLS and maximum likelihood, respectively, except for $\widehat{\tau}^{dr,p}_{imp}$, where we use the estimation methods described in Section 3.1. Finally, "Av. Bias", "Med. Bias", "RMSE", "Asy. V", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

estimand associated with $\widehat{\tau}^{fe}$ is not the ATT. As so, policy evaluations based on $\widehat{\tau}^{fe}$ can be misleading.

The results in Table 1 also suggest that, when both the OR and PS working models are correctly specified, all semiparametric estimators for the ATT show little to no Monte Carlo bias, but $\widehat{\tau}^{reg}$, $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ dominate the IPW DID estimators $\widehat{\tau}^{ipw,p}$ and $\widehat{\tau}^{ipw,p}_{std}$on the basis of bias, root mean square error, asymptotic variance, and length of the confidence interval. Indeed, both IPW DID estimator seem to be substantially less efficient than $\widehat{\tau}^{reg}$, $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$. The performance of these last three estimators are very close, though $\widehat{\tau}^{reg}$ tends to be more efficient than the other two DR DID estimators. Given that $\widehat{\tau}^{reg}$ exploits additional assumptions when compared to $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$, such a result is not unexpected. Also note that the Hájek (1971) type IPW estimator $\widehat{\tau}^{ipw,rc}_{std}$ is more stable than the Horvitz and Thompson (1952) type IPW estimator $\widehat{\tau}^{ipw,rc}$: the RMSE (the asymptotic variance) of $\widehat{\tau}^{ipw,rc}$ are more than two (four) times bigger than that of $\widehat{\tau}^{ipw,rc}_{std}$. Such a finding highlights the practical importance of using weights that are normalized to sum up to one.

When only the OR working model is correctly specified, our proposed DR DID estimators $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ are competitive with the OR DID estimator $\widehat{\tau}^{reg}$, while the IPW DID estimators are biased, as one should expect. On the other hand, when only the PS working model is correctly specified, the IPW and DR estimators show little to no bias, while $\widehat{\tau}^{reg}$ displays non-negligible bias. Here, it is worth emphasizing that $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ drastically outperform $\widehat{\tau}^{ipw,p}$ and $\widehat{\tau}^{ipw,p}_{std}$, with $\widehat{\tau}^{dr,p}_{imp}$ also showing substantial improvements with respect to both $\widehat{\tau}^{dr,p}$ and

$\widehat{\tau}^{ipw,p}_{std}$. When one compares the two IPW estimators, the role played by the normalized weights is again clear, as $\widehat{\tau}^{ipw,p}_{std}$ is again much more "stable" than $\widehat{\tau}^{ipw,p}$.

When both OR and PS working models are misspecified, not unexpectedly all estimators have non-negligible biases and inference procedures are, in general, misleading. In this scenario, our DR DID estimators have smaller biases and RMSE than the OR and the normalized IPW DID estimators, with $\widehat{\tau}^{dr,p}_{imp}$ strictly dominating $\widehat{\tau}^{dr,p}$. However, the Horvitz and Thompson (1952) IPW DID estimator $\widehat{\tau}^{ipw,p}$ seems to perform best in this DGP.

In terms of efficiency, the results in Table 1 show that the estimated asymptotic variance of $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ are very close to the semiparametric efficiency bound when both the PS and OR regression are correctly specified, which is in agreement with our locally efficiency results in Theorems 2 and A.1 (in the Appendix). When the PS is misspecified but the OR is not, the estimated asymptotic variances of $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ are still close to the semiparametric efficiency bound in this particular DGP, though we emphasize that this is not predicted by our results and can be a feature of this particular DGP. Finally, we note that when the OR is misspecified but the PS is not, the estimated asymptotic variances of our proposed DR DID estimators $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ are far from the semiparametric efficiency bound, with $\widehat{\tau}^{dr,p}_{imp}$ outperforming $\widehat{\tau}^{dr,p}$ in terms of efficiency in this particular DGP.

4.2 Simulation 2: repeated cross-section data are available

We now analyze the performance of the DID estimators for the ATT when one only observes repeated cross-section data. To do so, we consider the same DGPs as in the panel data framework, but instead of observing data on (Y_0, Y_1, D, Z), one observes data on (Y_0, D, Z) if $T = 0$, or on (Y_1, D, Z) if $T = 1$, where $T = 1\{U_T \leq \lambda\}$, and U_T is a standard uniform random variable, and $\lambda \in (0,1)$ a fixed constant.

Table 2 presents the simulation results with $\lambda = 0.5$ and with $n \equiv n_1 + n_0 = 1{,}000$.[8] Overall, the simulation exercise reveals that the efficiency bound, RMSE, asymptotic variance, and confidence interval length of the considered DID estimators are much larger when only repeated cross-section data are available than when panel data are available. In light of Corollary 1, such a result should be expected, though the magnitude of such loss of efficiency can be striking. In addition, the results in Table 2 reveal that: (i) the TWFE estimator $\widehat{\tau}^{fe}$ is severely biased for the ATT in all DGPs, just like in the panel data case; (ii) the IPW estimator with standardized weights $\widehat{\tau}^{ipw,rc}_{std}$ is much more stable and efficient than $\widehat{\tau}^{ipw,rc}$ in all DGPs, and, as one should expect, when the PS working model is misspecified, these IPW estimators display non-negligible biases; (iii) as one should expect, the OR DID estimator displays non-negligible bias when the OR working models are misspecified; (iv) all four DR DID estimators display little to no bias when one of the working models is correctly specified, but the locally efficient DR DID estimators $\widehat{\tau}^{dr,rc}_2$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ present important efficiency gains when compared to all other DID estimators, including $\widehat{\tau}^{dr,rc}_1$ and $\widehat{\tau}^{dr,rc}_{1,imp}$. These gains in efficiency are more pronounced when the OR models

8 Simulation results with $\lambda = 0.25$ and $\lambda = 0.75$ reached analogous conclusions to those discussed below and are available upon request.

Table 2: Monte Carlo results under designs $DGP1 - DGP4$ with repeated cross section data. Sample size $n = 1{,}000$, and $\lambda = 0.5$.

	DGP1: OR correct, PS correct Semiparametric Efficiency Bound: 44.4						DGP2: OR correct, PS incorrect Semiparametric Efficiency Bound: 46.4					
	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL
$\widehat{\tau}^{fe}$	-20.792	-20.741	21.099	12773.9	0.000	13.996	-19.178	-19.125	19.529	13240.7	0.001	14.247
$\widehat{\tau}^{reg}$	0.026	-0.030	7.588	57417.8	0.951	29.675	-0.024	-0.057	8.191	66577.5	0.948	31.945
$\widehat{\tau}^{ipw,rc}$	-0.662	-0.932	55.971	3090077.6	0.949	217.762	1.820	1.506	55.050	3023548.1	0.949	215.449
$\widehat{\tau}^{ipw,rc}_{std}$	-0.050	-0.125	9.648	92235.7	0.949	37.560	-0.812	-0.698	9.814	94343.0	0.946	38.031
$\widehat{\tau}^{dr,rc}_{1}$	0.013	-0.007	3.041	9222.0	0.950	11.893	-0.010	-0.022	3.281	10686.4	0.949	12.799
$\widehat{\tau}^{dr,rc}_{2}$	0.004	0.003	0.216	44.4	0.944	0.824	0.000	0.001	0.211	42.3	0.945	0.805
$\widehat{\tau}^{dr,rc}_{1,imp}$	0.014	-0.008	3.041	9220.1	0.951	11.892	-0.009	-0.022	3.282	10686.2	0.949	12.799
$\widehat{\tau}^{dr,rc}_{2,imp}$	0.005	0.002	0.216	42.1	0.937	0.803	0.000	0.001	0.213	41.3	0.940	0.796
	DGP3: OR incorrect, PS correct Semiparametric Efficiency Bound: 44.4						DGP4: OR incorrect, PS incorrect Semiparametric Efficiency Bound: 46.4					
	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL	Av. Bias	Med. Bias	RMSE	Asy. V	Cover	CIL
$\widehat{\tau}^{fe}$	-13.131	-13.092	14.058	25446.9	0.260	19.766	-16.330	-16.354	17.126	26347.3	0.114	20.112
$\widehat{\tau}^{reg}$	-1.376	-1.397	8.137	64143.7	0.942	31.378	-5.338	-5.437	9.977	72665.8	0.908	33.397
$\widehat{\tau}^{ipw,rc}$	-0.973	-1.452	57.262	3241967.3	0.947	223.050	-1.391	-0.980	55.178	3101777.5	0.952	218.233
$\widehat{\tau}^{ipw,rc}_{std}$	0.051	-0.011	9.428	86806.4	0.943	36.483	-4.149	-4.387	10.520	94034.1	0.930	37.971
$\widehat{\tau}^{dr,rc}_{1}$	-0.086	-0.083	5.692	31830.9	0.945	22.060	-3.342	-3.375	7.071	38663.1	0.916	24.290
$\widehat{\tau}^{dr,rc}_{2}$	-0.029	-0.022	4.742	21869.3	0.942	18.261	-3.275	-3.249	6.016	24194.2	0.886	19.159
$\widehat{\tau}^{dr,rc}_{1,imp}$	-0.119	-0.102	4.837	23038.9	0.945	18.804	-2.689	-2.708	5.564	23473.3	0.913	18.979
$\widehat{\tau}^{dr,rc}_{2,imp}$	-0.076	-0.081	4.062	15765.2	0.944	15.550	-2.614	-2.610	4.845	15769.1	0.892	15.552

Notes: Simulations based on 10,000 Monte Carlo experiments. $\widehat{\tau}^{fe}$ is the TWFE outcome regression estimator of τ^{fe} in (2.5), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,rc}$ is the IPW DID estimator based on the sample analogue of (2.3), $\widehat{\tau}^{ipw,rc}_{std}$ is the standardized IPW DID estimator (4.2), and $\widehat{\tau}^{dr,rc}_{1}$, $\widehat{\tau}^{dr,rc}_{2}$, $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$are our proposed DR DID estimators given in (3.3), (3.4), and in (3.9) and (3.10) in Section 3.2. We use a linear OR working model and a logistic PS working model, where the unknown parameters are estimated via OLS and maximum likelihood, respectively, except for $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ where we use the estimation methods described in Section 3.2. Finally, "Av. Bias","Med. Bias", "RMSE", "Asy. V", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, , average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

are correctly specified. The simulation results also show that (v) when one compares the performance of the further improved DR DID estimators $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ with the "traditional" DR DID estimators $\widehat{\tau}^{dr,rc}_{1}$ and $\widehat{\tau}^{dr,rc}_{2}$, it is clear that appropriately choosing the estimation methods for the nuisance parameters can have practical consequences, especially when the outcome regression working models are misspecified.

In terms of efficiency, the results in Table 2 highlight that, when all working models are correctly specified, the estimated asymptotic variances of $\widehat{\tau}^{dr,rc}_{2}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ are indeed close to the semiparametric efficiency bound, but the asymptotic variances $\widehat{\tau}^{dr,rc}_{1}$ and $\widehat{\tau}^{dr,rc}_{1,imp}$ are substantially higher than the semiparametric efficiency bound; these findings are in agreement with our locally efficiency results in Theorems 3 and A.2 (in the Appendix). Similarly to the panel data case, we find that, in this specific DGP, the estimated asymptotic variances of $\widehat{\tau}^{dr,rc}_{2}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ are still close to the semiparametric efficiency bound when the outcome regressions are correctly specified but the PS is not, but not when the PS is correctly specified but the outcome regressions are not.

5 Empirical illustration: the effect of job training on earnings

In a very influential study, LaLonde (1986) analyzes whether different treatment effect estimators based on observational data are able to replicate the experimental findings of the NSW job training program on post treatment earnings. His negative results led to an increased awareness of the potential pitfalls of observational data and helped spur the use of randomized controlled trials among economists. In addition, alternative policy evaluation tools arose to overcome "LaLonde's critique" of observational estimators. Two prominent examples are the propensity score matching (PSM), see e.g. Dehejia and Wahba (1999, 2002) (henceforth DW) and the difference-in-differences matching, see e.g. Heckman et al. (1997) and Smith and Todd (2005) (henceforth ST). For instance, DW show that PSM can replicate the experimental benchmark of the NSW for a particular subsample of the original data. ST, on the other hand, cast doubt on the "generalizability" of DW PSM results to a larger population and argue that the conclusions may be sensitive to the propensity score specification. ST also argue that for the NSW data, difference-in-differences matching estimators may be more suitable than cross-section PSM, as they can account for time-invariant unobserved confounding factors.

Motivated by ST findings, in what follows, we focus on DID estimators and evaluate whether our proposed DR DID estimators can better reduce the selection bias when compared to other DID estimation procedures. We analyze three different experimental samples — the original LaLonde experimental sample, the DW sample, and the "early random assignment" (early RA) subsample of the DW sample considered by ST — and consider data from the Current Population Survey (CPS) to form a non-experimental comparison group. The pre-treatment covariates in the data include age, years of education, real earnings in 1974, and dummy variables for high school dropout, married, black, and Hispanic. The outcome of interest is real earnings in 1978. We also observe real earnings in 1975, which we use as the pre-treatment outcome Y_0. The experimental benchmark for the ATT is equal to \$886 (s.e. \$488), \$1794 (s.e. \$671), and \$2748 (s.e. \$1005) for the LaLonde, DW, and early RA sample, respectively. For additional description and summary statistics for each sample, see Smith and Todd (2005).

Following ST, we focus on estimating the average "evaluation bias" of different DID estimators. This is only made possible given the availability of experimental data. First, randomization ensures that both "treatment" groups are comparable in terms of self selection. Second, given that randomized-out individuals did not receive training via NSW, the impact of NSW is known to be zero in this group. Thus, applying different DID estimators to data from randomized-out individuals (our pseudo treated group in this exercise) and nonexperimental CPS comparison observations (our comparison group in this exercise) should produce an estimated ATT equal to

zero, if these DID estimators are consistent. Deviations from zero are what we call evaluation bias.[9]

Like in the Monte Carlo simulation exercises, we compare our proposed DR DID estimators $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}^{dr,p}_{imp}$ with the TWFE estimator $\widehat{\tau}^{fe}$ based on (2.5), the OR DID estimator $\widehat{\tau}^{reg}$ as defined in (2.2), and the Horvitz and Thompson (1952) type IPW DID estimator proposed by Abadie (2005), $\widehat{\tau}^{ipw,p}$, as defined in (2.4). We also consider the Hájek (1971) type IPW estimator $\widehat{\tau}^{ipw,p}_{std}$ as defined in (4.1). We assume that the outcome models are linear in parameters and that the propensity score follows a logistic specification. The unknown parameters are estimated using ordinary least squares (OLS) and maximum likelihood, respectively, except in $\widehat{\tau}^{dr,p}_{imp}$, where we use the estimation methods described in Section 3.1.

In order to assess the sensitivity of the findings with respect to the model specifications, we consider three different specifications for how covariates enter into each model: (i) a linear specification where all covariates enter the models linearly; (ii) a specification in the spirit of DW, which adds to the linear specification a dummy for zero earnings in 1974, age squared, age cubed divided by 1000, years of schooling squared, and an interaction term between years of schooling and real earnings in 1974; and (iii) an "augmented DW" specification, which adds to the "DW" specification the interactions between married and real earnings in 1974, and between married and zero earnings in 1974 — these two interaction terms were used in Firpo (2007).

Table 3 summarizes the results. Standard errors are reported in parentheses and the estimated evaluation biases relative to the experimental ATT benchmark are reported in brackets. As argued by ST, these "relative biases" are useful for comparing DID estimators within each sample, but as the experimental benchmark estimates for the ATT vary substantially among the three experimental samples, they should not be used for comparing DID estimators across samples.

Table 3 highlights some interesting patterns. First, estimators based on two-way fixed effect regression models tend to be very stable across specifications, but usually display large positive and statistically significant evaluation biases. Second, DID estimators based on the regression approach tend to lead to the most precise estimates. However, for the LaLonde sample, point estimates are severely biased downward, leading to statistically significant evaluation biases. Abadie's IPW estimators $\widehat{\tau}^{ipw,p}$ for the ATT tend to have the largest standard errors across all considered estimators, but their evaluation biases are relatively small. Like in our Monte Carlo simulation results, considering normalized weights as in $\widehat{\tau}^{ipw,p}_{std}$ can improve the stability of the IPW estimators $\widehat{\tau}^{ipw,p}$. Finally, note that our proposed DR DID estimators share the favorable bias properties of Abadie's IPW estimator, but at the same time, have smaller standard errors than IPW estimators. When we compare $\widehat{\tau}^{dr,p}$ with

9 An alternative way to estimate "evaluation bias" is to compare the ATT using the experimental data with ATT using data from randomized-in and nonexperimental comparison units. This is the approach taken by LaLonde (1986) and Dehejia and Wahba (1999, 2002). A disadvantage of this approach compared to the one we and Smith and Todd (2005) use is that experimental ATT estimates are also random and may differ from the "true" ATT. Thus, the computation of "true" evaluation biases is much more challenging if not impossible. In any case, results treating the experimental ATT as true effects lead to similar conclusions and are available upon request.

Table 3: Evaluation bias of different difference-in-differences estimators for the effect of of training on real earnings in 1978. NSW data with CPS comparison group.

	Results for Lalonde sample Evaluation Bias: ATT= 0						Results for DW sample Evaluation Bias: ATT= 0						Results for Early RA sample Evaluation Bias: ATT= 0					
Spec.	$\widehat{\tau}^{dr,p}$	$\widehat{\tau}^{dr,p}_{imp}$	$\widehat{\tau}^{reg}$	$\widehat{\tau}^{ipw,p}$	$\widehat{\tau}^{ipw,p}_{std}$	$\widehat{\tau}^{fe}$	$\widehat{\tau}^{dr,p}$	$\widehat{\tau}^{dr,p}_{imp}$	$\widehat{\tau}^{reg}$	$\widehat{\tau}^{ipw,p}$	$\widehat{\tau}^{ipw,p}_{std}$	$\widehat{\tau}^{fe}$	$\widehat{\tau}^{dr,p}$	$\widehat{\tau}^{dr,p}_{imp}$	$\widehat{\tau}^{reg}$	$\widehat{\tau}^{ipw,p}$	$\widehat{\tau}^{ipw,p}_{std}$	$\widehat{\tau}^{fe}$
Lin.	-871	-901	-1301	-1108	-1022	868	253	253	-230	188	155	2092	-434	-441	-831	-516	-515	1136
	(396)	(394)	(350)	(409)	(398)	(353)	(451)	(452)	(408)	(459)	(452)	(459)	(605)	(607)	(583)	(611)	(607)	(730)
	[-98%]	[-102%]	[-147%]	[-125%]	[-115%]	[98%]	[14%]	[14%]	[-13%]	[10%]	[9%]	[117%]	[-16%]	[-16%]	[-30%]	[-19%]	[-19%]	[41%]
DW	-626	-591	-830	-732	-564	868	408	520	402	-34	481	2092	-246	-176	-264	-495	-223	1136
	(496)	(467)	(360)	(534)	(487)	(359)	(691)	(588)	(426)	(845)	(672)	(471)	(724)	(683)	(596)	(781)	(718)	(751)
	[-71%]	[-67%]	[-94%]	[-83%]	[-64%]	[98%]	[23%]	[29%]	[22%]	[-2%]	[27%]	[117%]	[-9%]	[-6%]	[-10%]	[-18%]	[-8%]	[41%]
ADW	-597	-599	-1041	-685	-558	868	514	524	27	97	502	2092	-148	-144	-498	-337	-165	1136
	(491)	(470)	(358)	(523)	(485)	(352)	(663)	(582)	(428)	(793)	(653)	(458)	(701)	(677)	(591)	(740)	(700)	(728)
	[-67%]	[-68%]	[-118%]	[-77%]	[-63%]	[98%]	[29%]	[29%]	[2%]	[5%]	[28%]	[117%]	[-5%]	[-5%]	[-18%]	[-12%]	[-6%]	[41%]

Notes: The results (standard errors are in parentheses) represent the estimated average effect of being in the experimental sample (i.e. the estimated evaluation bias) on the 1978 earnings where the experimental control group is compared with untreated non-experimental CPS sample. The estimated evaluation biases relative to the experimental ATT benchmark, in percentage terms, are reported in brackets. $\widehat{\tau}^{fe}$ is the TWFE outcome regression estimator of τ^{fe} in (2.5), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,p}$ is the IPW DID estimator (2.4), $\widehat{\tau}^{dr,p}_{std}$ is the standardized IPW DID estimator (4.1), $\widehat{\tau}^{dr,p}$ is our proposed DR DID estimator (3.1), and $\widehat{\tau}^{dr,p}_{imp}$ is our proposed DR DID estimator (3.7). We use a linear OR working model and a logistic PS working model, where the unknown parameters are estimated via OLS and maximum likelihood, respectively, except for $\widehat{\tau}^{dr,p}_{imp}$, where we use the estimation methods described in Section 3.1. For each DID estimator, we report three different specifications depending on how covariates are included: "lin." specification, where all covariates enter the model linearly; "DW" specification, which adds to the linear specification a dummy for zero earnings in 1974, age squared, age cubed divided by 1000, years of schooling squared, and an interaction term between years of schooling and real earnings in 1974; and the "ADW" specification, which adds to the "DW" specification the interactions between married with real earnings in 1974, and between married and zero earnings in 1974.

$\widehat{\tau}^{dr,p}_{imp}$, we note that the further improved DR DID estimator $\widehat{\tau}^{dr,p}_{imp}$ tends to have smaller standard errors, particularly when one adopts the "DW" or the "augmented DW" specifications. Taken together, the results using the NSW job training data suggest that our proposed DR DID estimators are an attractive alternative to existing DID procedures.

6 Concluding remarks

In this article, we proposed doubly robust estimators for the ATT in difference-in-differences settings where the parallel trends assumption holds only after conditioning on a vector of pre-treatment covariates. Our proposed estimators remain consistent for the ATT when either (but not necessarily both) a propensity score model or outcome regression models are correctly specified, and achieve the semiparametric efficiency bound when the working models for the nuisance functions are correctly specified. We derived the large sample properties of the proposed estimators in situations where either panel data or repeated cross-section data are available, and showed that by paying particular attention to the estimation methods used to estimate the nuisance parameters, one can form DID estimators for the ATT that are not only DR consistent and locally semiparametric efficient, but also DR for inference. We illustrated the attractiveness of our proposed causal inference tools via a simulation exercise and with an empirical application.

Our results can be extended to other situations of practical interest. A leading case is when researchers are interested in understanding treatment effect heterogeneity with respect to continuous covariates X_1, where X_1 is a (strict) subset of available covariates X. Here, the parameter of interest is the conditional average treatment effect on the treated $\text{CATT}(X_1) \equiv \mathbb{E}[Y(1) - Y(0) | X_1, D = 1]$ and because of its infinite dimensional nature, the estimation and inference tools proposed in this paper are not directly applicable. However, by combining the DR DID formulation proposed in this paper with the methodology put forward by Chen and Christensen (2018), one can propose uniformly valid inference procedures not only for the CATT but also for possibly nonlinear functionals of the CATT such as (higher order) partial derivatives, conditional average (higher order) partial derivatives, and partial derivatives of its log.

Another interesting extension is when researchers want to adopt data-adaptive, "machine learning" first-step estimators instead of the parametric models discussed in this paper. Here, the main challenge is to derive the influence function of the DR DID estimator for the ATT, as "machine learning" estimators are, in general, in a non-Donsker classes of functions. We envision that one can bypass such technical complications by combining the results derived in this paper with those in Chernozhukov et al. (2017), Belloni et al. (2017), and Tan (2019), for example; see e.g. Zimmert (2019) for some recent results in this direction. We leave the detailed analysis of these extensions to future work.

A Appendix A: Asymptotic Properties of the DR DID estimators based on generic first-step estimators

Let $g(x)$ be a generic notation for $\pi(x)$, $\mu^p_{d,t}(x)$ and $\mu^{rc}_{d,t}(x)$, $d,t=0,1$. Analogously and with some abuse of notation, let $g(x;\theta)$ be a generic notation for $\pi(x;\gamma)$, $\mu^p_{d,t}\left(x,\beta^p_{d,t}\right)$ and $\mu^{rc}_{d,t}\left(x,\beta^{rc}_{d,t}\right)$, $d,t=0,1$. Let $W=(Y_0,Y_1,D,X)$ in the panel data case and $W=(Y,T,D,X)$ in the repeated cross-section data case. Denote the support of X by $\mathscr{X}$ and for a generic Z, let $\|Z\|=\sqrt{trace\left(Z'Z\right)}$ denote the Euclidean norm of Z.

Let

$$\begin{aligned}
h^p\left(W;\kappa^p\right) &= \left(w^p_1(D)-w^p_0(D,X;\gamma)\right)\left(\Delta Y-\mu^p_{0,\Delta}(X;\beta^p_{0,0},\beta^p_{0,1})\right),\\
h^{rc\ 1}\left(W;\kappa^{rc\ 1}\right) &= \left(w^{rc}_1(D,T)-w^{rc}_0(D,T,X;\gamma)\right)\left(Y-\mu^{rc}_{0,Y}\left(T,X;\beta^{rc}_{0,0},\beta^{rc}_{0,1}\right)\right),\\
h^{rc\ 2}\left(W;\kappa^{rc\ 2}\right) &= \left(D/\mathbb{E}[D]\right)\cdot\left(\mu^{rc}_{1,\Delta}\left(X;\beta^{rc}_{1,1},\beta^{rc}_{1,0}\right)-\mu^{rc}_{0,\Delta}\left(X;\beta^{rc}_{0,1},\beta^{rc}_{0,0}\right)\right)\\
&\quad +w^{rc}_{1,1}(D,T)\left(Y-\mu^{rc}_{1,1}\left(X;\beta^{rc}_{1,1}\right)\right)-w^{rc}_{1,0}(D,T)\left(Y-\mu^{rc}_{1,0}\left(X;\beta^{rc}_{1,0}\right)\right)\\
&\quad -\left(w^{rc}_{0,1}(D,T,X;\gamma)\left(Y-\mu^{rc}_{0,1}\left(X;\beta^{rc}_{0,1}\right)\right)-w^{rc}_{0,0}(D,T,X;\gamma)\left(Y-\mu^{rc}_{0,0}\left(X;\beta^{rc}_{0,0}\right)\right)\right)
\end{aligned}$$

where $\kappa^p=\left(\gamma',\beta^{p\prime}_{0,0},\beta^{p\prime}_{0,1}\right)'$, $\kappa^{rc\ 1}=\left(\gamma',\beta^{rc\prime}_{0,0},\beta^{rc\prime}_{0,1}\right)'$ and $\kappa^{rc\ 2}=\left(\gamma',\beta^{rc\prime}_{0,0},\beta^{rc\prime}_{0,1},\beta^{rc\prime}_{1,1},\beta^{rc\prime}_{1,0}\right)'$. In obvious notation, the vector of pseudo-true parameter[10] is given by $\kappa^{*,p}$, $\kappa^{*,rc\ 1}_0$, and $\kappa^{*,rc\ 2}$. Let $\dot{h}^p\left(W;\kappa^p\right)=\partial\, h^p\left(W;\kappa^p\right)/\partial\,\kappa^p$ and define $\dot{h}^{rc\ j}\left(W;\kappa^{rc\ j}\right)$, $j=0,1$, analogously.

Assumption A.1 *(i) $g(x)=g(x;\theta)$ is a parametric model, where $\theta\in\Theta\subset\mathbb{R}^k$, Θ being compact; (ii) $g(X;\theta)$ is a.s. continuous at each $\theta\in\Theta$; (iii) there exists a unique pseudo-true parameter $\theta^*\in int(\Theta)$; (iv) $g(X;\theta)$ is a.s. twice continuously differentiable in a neighborhood of θ^*, $\Theta^*\subset\Theta$; (v) the estimator $\widehat{\theta}$ is strongly consistent for the θ^* and satisfies the following linear expansion:*

$$\sqrt{n}\left(\widehat{\theta}-\theta^*\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_g\left(W_i;\theta^*\right)+o_p(1),$$

where $l_g(\cdot;\theta)$ is such that $\mathbb{E}\left[l_g\left(W;\theta^*\right)\right]=0$, $\mathbb{E}\left[l_g\left(W;\theta^*\right)l_g\left(W;\theta^*\right)'\right]$ exists and is positive definite and $\lim_{\delta\to 0}\mathbb{E}\left[\sup_{\theta\in\Theta^*:\|\theta-\theta^*\|\le\delta}\left\|l_g(W;\theta)-l_g\left(W;\theta^*\right)\right\|^2\right]=0$. In addition, (vi) for some $\varepsilon>0$, $0<\pi(X;\gamma)\le 1-\varepsilon$ a.s., for all $\gamma\in int\left(\Theta^{ps}\right)$, where Θ^{ps} denotes the parameter space of γ.

Assumption A.2 *(i) When panel data are available, assume that $\mathbb{E}\left[\left\|h^p\left(W;\kappa^{*,p}\right)\right\|^2\right]<\infty$ and $\mathbb{E}\left[\sup_{\kappa\in\Gamma^{*,p}}\left|\dot{h}^p\left(W;\kappa\right)\right|\right]<\infty$, where $\Gamma^{*,p}$ is a small neighborhood of $\kappa^{*,p}$. (ii) When cross-section data are available, assume that, for $j=1,2$, $\mathbb{E}\left[\left\|h^{rc,j}\left(W;\kappa^{*,rc,j}\right)\right\|^2\right]<\infty$ and $\mathbb{E}\left[\sup_{\kappa\in\Gamma^{*,rc\ j}}\left|\dot{h}^{rc,j}\left(W;\kappa\right)\right|\right]<\infty$, where $\Gamma^{*,rc\ j}$ is a small neighborhood of $\kappa^{*,rc\ j}$.*

Assumptions A.1-A.2 are standard in the literature, see e.g. Abadie (2005), Wooldridge (2007), Bonhomme and Sauder (2011), Graham et al. (2012) and Callaway and Sant'Anna (2018). Assumption A.1 requires

10 Note that we allow for possible misspecification when we define pseudo-true parameters.

that the first-step estimators are based on smooth parametric models and that the estimated parameters admit $\sqrt{n}$-asymptotically linear representations, whereas Assumption A.2 imposes some weak integrability conditions. Under mild moment conditions, these requirements are fulfilled when one adopts linear/nonlinear outcome regressions or logit/probit models, for example, and estimates the unknown parameters by (nonlinear) least squares, quasi-maximum likelihood, or other alternative estimation methods, see e.g. Chapter 5 in van der Vaart (1998), Wooldridge (2007), Graham et al. (2012) and Sant'Anna et al. (2018).

Next, we derive the asymptotic properties of $\widehat{\tau}^{dr,p}$, $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$ using generic first-step estimators that satisfy Assumptions A.1 and A.2.

A.1 Panel data case

In this section, we discuss the asymptotic properties of $\widehat{\tau}^{dr,p}$. Define $\dot{\pi}(x;\gamma) \equiv \partial \pi(x;\gamma) / \partial \gamma$ and, for $t = 0,1$, define $\dot{\mu}_{0,t}^p\left(x;\beta_{0,t}^p\right)$ analogously. In what follows, we drop the dependence of the functionals on W to ease the notational burden. For example, we write $w_1^p = w_1^p(D)$, $w_0^p(\gamma) = w_0^p(D,X;\gamma)$, and so on and so forth.

For generic γ and $\beta_0 = \left(\beta_{0,1}', \beta_{0,0}'\right)'$, let

$$\eta^p(W;\gamma,\beta) = \eta_1^p(W;\beta_0) - \eta_0^p(W;\gamma,\beta_0) - \eta_{est}^p(W;\gamma,\beta_0), \tag{A.1}$$

where

$$\begin{aligned}
\eta_1^p(W;\beta_0) &= w_1^p \cdot \left[\left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right) - \mathbb{E}\left[w_1^p \cdot \left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right)\right]\right], \\
\eta_0^p(W;\gamma,\beta_0) &= w_0^p(\gamma) \cdot \left[\left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right) - \mathbb{E}\left[w_0^p(\gamma) \cdot \left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right)\right]\right],
\end{aligned}$$

and

$$\begin{aligned}
\eta_{est}^p(W;\gamma,\beta_0) = {} & l_{reg}(\beta_0)' \cdot \mathbb{E}\left[\left(w_1^p - w_0^p(\gamma)\right) \cdot \dot{\mu}_{0,\Delta}^p(\beta_0)\right] \\
& + l_{ps}(\gamma)' \cdot \mathbb{E}\left[\alpha_{ps}^p(\gamma)\left(\left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right) - \mathbb{E}\left[w_0^p(\gamma) \cdot \left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right)\right]\right) \cdot \dot{\pi}(\gamma)\right],
\end{aligned} \tag{A.2}$$

with $l_{reg}(\beta_0) = \left(l_{reg,0,1}(\beta_{0,1})', l_{reg,0,0}(\beta_{0,0})'\right)'$, where $l_{reg,d,t}(\cdot)$ is the asymptotic linear representation of the estimators for the outcome regression as described in Assumption A.1(iv), $l_{ps}(\cdot)$ is defined analogously, $\dot{\mu}_{0,\Delta}^p(\beta_0) = \left(\dot{\mu}_{0,1}^p(\beta_{0,1})',\ -\dot{\mu}_{0,0}^p(\beta_{0,0})\right)'$ and

$$\alpha_{ps}^p(\gamma) = \frac{(1-D)}{(1-\pi(X;\gamma))^2} \Bigg/ \mathbb{E}\left[\frac{\pi(X;\gamma)(1-D)}{1-\pi(X;\gamma)}\right].$$

For $d,t = 0,1$, let $\Theta_{d,t}^{reg}$ be the parameter space for the regression coefficient $\beta_{d,t}$, and Θ^{ps} be the parameter space for the propensity score coefficient γ. Consider the following claims:

$$\exists \gamma^* \in \Theta^{ps} : \mathbb{P}\left(\pi(X;\gamma^*) = p(X)\right) = 1, \tag{A.3}$$

$$\exists \left(\beta_{0,1}^{*,p}, \beta_{0,0}^{*,p}\right) \in \Theta_{0,1}^{reg} \times \Theta_{0,0}^{reg} : \mathbb{P}\left(\mu_{0,1}^p\left(X;\beta_{0,1}^{*,p}\right) - \mu_{0,0}^p\left(X;\beta_{0,0}^{*,p}\right) = m_{0,1}^p(X) - m_{0,0}^p(X)\right) = 1. \tag{A.4}$$

Now we are ready to state the large sample properties of $\widehat{\tau}^{dr,p}$.

Theorem A.1 *Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix A hold.*

(*a*) *Provided that either (A.3) or (A.4) is true, as* $n \to \infty$,

$$\widehat{\tau}^{dr,p} \xrightarrow{p} \tau.$$

Furthermore,

$$\begin{aligned}\sqrt{n}(\widehat{\tau}^{dr,p} - \tau^{dr,p}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta^{p}\left(W_i; \gamma^*, \beta_0^{*,p}\right) + o_p(1) \\ &\xrightarrow{d} N\left(0, V^{p}\right),\end{aligned}$$

where $V^{p} = \mathbb{E}[\eta^{p}\left(W; \gamma^*, \beta_0^{*,p}\right)^2]$.

(*b*) *When both (A.3) and (A.4) are true,* $\eta^{p}\left(W; \gamma^*, \beta_0^{*,p}\right) = \eta^{e,p}\left(Y_1, Y_0, D, X\right)$ *a.s. and* V^{p} *is equal to the semiparametrically efficiency bound (2.12).*

Theorem A.1 indicates that, provided that either the propensity score model or the model for the evolution of the outcome for the comparison group is correctly specified, $\widehat{\tau}^{dr,p}$ is consistent for the ATT, implying that our proposed estimator is indeed doubly robust. In addition, Theorem A.1 indicates that our proposed estimator admits an asymptotically linear representation and as a consequence, it is $\sqrt{n}$-consistent and asymptotically normal. When the models for the nuisance functions are correctly specified, our proposed DR DID estimator is semiparametrically efficient.

Theorem A.1 also suggests that one can use the analogy principle to estimate V^{p} and conduct asymptotically valid inference.[11] However, it is worth mentioning the fact that the exact form of V^{p} depends on which nuisance models are correctly specified, implying that our (generic) estimator $\widehat{\tau}^{dr,p}$ is doubly robust in terms of consistency but, in general, not doubly robust for inference. Given that in practice it is hard to know *a priori* which nuisance models are correctly specified, one should include all "correction" terms in η_{est}^{p} when estimating V^{p}. Failing to do so may lead to asymptotically invalid inference procedures.

A.2 Repeated cross-section data case

In this section, we turn our attention to our proposed DR DID estimators for the ATT when only repeated cross-section data are available. For generic γ and $\beta = \left(\beta_1', \beta_0'\right)'$, where, for $d = 0, 1$, $\beta_d = \left(\beta_{d,1}', \beta_{d,0}'\right)'$, let

$$\eta_j^{rc}(W; \gamma, \beta) = \eta_1^{rc,j}(W; \beta) - \eta_0^{rc,j}(W; \gamma, \beta) - \eta_{est}^{rc,j}(W; \gamma, \beta), \tag{A.5}$$

such that, for $j = 1, 2$,

$$\begin{aligned}\eta_1^{rc,j}(W; \beta) &= \eta_{1,1}^{rc,j}(W; \beta) - \eta_{1,0}^{rc,j}(W; \beta), \\ \eta_0^{rc,j}(W; \gamma, \beta) &= \eta_{0,1}^{rc,j}(W; \gamma, \beta) - \eta_{0,0}^{rc,j}(W; \gamma, \beta), \\ \eta_{est}^{rc,j}(W; \gamma, \beta) &= \eta_{est,reg}^{rc,j}(W; \gamma, \beta) + \eta_{est,ps}^{rc,j}(W; \gamma, \beta),\end{aligned}$$

11 It is easy to show that the plug-in estimator of V^{p} is consistent, see e.g. Lemma 4.3 in Newey and McFadden (1994) and Theorem 4.4 in Abadie (2005). We omit the detailed derivation of this result for the sake of brevity.

and the precise definitions of all these η^{rc} functions are deferred to Appendix B to avoid excess notational complexity. An aspect of the difference between η_1^{rc} and η_2^{rc} that is worth mentioning but is perhaps buried in the notation is that η_1^{rc} depends on β only through β_0, whereas η_2^{rc} depends on both β_1 and β_0. This is simply a consequence from the fact that $\widehat{\tau}_1^{dr,rc}$ does not rely on outcome regressions for the treated units, but $\widehat{\tau}_2^{dr,rc}$ does.

Consider the following claims:

$$\exists\left(\beta_{0,1}^{*,rc},\beta_{0,0}^{*,rc}\right) \in \Theta_{0,1}^{reg}\times\Theta_{0,0}^{reg}:\mathbb{P}\left(\mu_{0,1}^{rc}\left(X;\beta_{0,1}^{*,rc}\right)-\mu_{0,0}^{rc}\left(X;\beta_{0,0}^{*,rc}\right)=m_{0,1}^{rc}(X)-m_{0,0}^{rc}(X)\right)=1, \quad \text{(A.6)}$$

$$\forall(d,t) \in \{0,1\}^2\ \exists\left(\beta_{d,t}^{*,rc}\right)\in\Theta_{d,t}^{reg}:\mathbb{P}\left(\mu_{d,t}^{rc}\left(X;\beta_{d,t}^{*,rc}\right)=m_{d,t}^{rc}(X)\right)=1. \quad \text{(A.7)}$$

Theorem A.2 *Let $n=n_1+n_0$, where n_1 and n_0 are the sample sizes of the post-treatment and pre-treatment periods, respectively. Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix A hold, and that $n_1/n \xrightarrow{p} \lambda \in (0,1)$ as $n_0,n_1\to\infty$.*

(a) Provided that either (A.3) or (A.6) is true, as $n\to\infty$, for $j=1,2$,

$$\widehat{\tau}_j^{dr,rc} \xrightarrow{p} \tau.$$

Furthermore,

$$\begin{aligned}\sqrt{n}(\widehat{\tau}_j^{dr,rc}-\tau_j^{dr,rc}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_j^{rc}(W_i;\gamma^*,\beta^{*,rc})+o_p(1)\\ &\xrightarrow{d} N\left(0,V_j^{rc}\right),\end{aligned}$$

where $V_j^{rc}=\mathbb{E}[\eta_j^{rc}(W;\gamma^*,\beta^{*,rc})^2]$.

(b) Suppose that both (A.3) and (A.7) are true. Then, $\eta_2^{rc}(W;\gamma^*,\beta^{*,rc})=\eta^{e,rc}(Y,D,T,X)$ a.s., and V_2^{rc} is equal to the semiparametrically efficiency bound (2.14). On the other hand, V_1^{rc} does not attain the semiparametric efficiency bound when (A.3) and (A.7) are true.

In other words, Theorem A.2 states that both proposed estimators for the ATT, $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$, are doubly robust, $\sqrt{n}$-consistent and asymptotically normal. Similar to the panel data case, the exact form of the V_j^{rc}, $j=1,2$, depends on which working models are correctly specified, implying that the generic estimators $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$ are doubly robust in terms of consistency but in terms of inference.

Part (b) of Theorem A.2 indicates that $\widehat{\tau}_2^{dr,rc}$ is semiparametrically efficient when the working model for the propensity score, and all working models for the outcome regressions, for both treated and comparison units, are correctly specified. When compared to Theorem A.1(b), it is evident that such a requirement is stronger than when panel data are available.

B Appendix B: Influence function of the DR DID estimators with repeated cross-section

As it is evident from Theorem A.2, the influence functions of $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$ play a major role in study of the large sample properties of our proposed DR DID estimators. In this section, we state the precise definition of $\eta_j^{rc}(W;\gamma,\beta)$, $j=1,2$, introduced in (A.5).

We first focus on $\widehat{\tau}_1^{dr,rc}$. For generic γ and $\beta=\left(\beta_1',\beta_0'\right)'$, where, for $d=0,1$, $\beta_d=\left(\beta_{d,1}',\beta_{d,0}'\right)'$, let

$$\eta_1^{rc}(W;\gamma,\beta)=\eta_1^{rc,1}(W;\beta_0)-\eta_0^{rc,1}(W;\gamma,\beta_0)-\eta_{est}^{rc,1}(W;\gamma,\beta_0),$$

where

$$\eta_1^{rc,1}(W;\beta_0)=\eta_{1,1}^{rc,1}(W;\beta_{0,1})-\eta_{1,0}^{rc,1}(W;\beta_{0,0}), \tag{B.1}$$

$$\eta_0^{rc,1}(W;\gamma,\beta_0)=\eta_{0,1}^{rc,1}(W;\gamma,\beta_{0,1})-\eta_{0,0}^{rc,1}(W_i;\gamma,\beta_{0,0}), \tag{B.2}$$

$$\eta_{est}^{rc,1}(W;\gamma,\beta_0)=\eta_{est,reg}^{rc,1}(W;\gamma,\beta_0)+\eta_{est,ps}^{rc,1}(W;\gamma,\beta_0),$$

and, for $t=0,1$,

$$\eta_{1,t}^{rc,1}(W;\gamma,\beta)=w_{1,t}^{rc}(D,T)\cdot\left(Y-\mu_{0,t}^{rc}(X;\beta_{0,t})-\mathbb{E}[w_{1,t}^{rc}(D,T)\cdot\left(Y-\mu_{0,t}^{rc}(X;\beta_{0,t})\right)]\right),$$

$$\eta_{0,t}^{rc,1}(W;\gamma,\beta)=w_{0,t}^{rc}(D,T,X;\gamma)\cdot\left(Y-\mu_{0,t}^{rc}(X;\beta_{0,t})-\mathbb{E}[w_{0,t}^{rc}(D,T,X;\gamma)\cdot\left(Y-\mu_{0,t}^{rc}(X;\beta_{0,t})\right)]\right),$$

and the influence functions associated with the estimation effects of the nuisance parameters are

$$\eta_{est,reg}^{rc,1}(W;\gamma,\beta)=l_{reg}(W;\beta)'\cdot\mathbb{E}[(w_{1,1}^{rc}-w_{1,0}^{rc})-(w_{0,1}^{rc}(\gamma)-w_{0,0}^{rc}(\gamma))\cdot\dot{\mu}_{0,Y}^{rc}(T,X;\beta)],$$

and

$$\begin{aligned}&\eta_{est,ps}^{rc,1}(W;\gamma,\beta)\\&=l_{ps}(D,X;\gamma)'\cdot\mathbb{E}\left[\alpha_{ps,1}^{rc}(\gamma)\cdot\left(Y-\mu_{0,1}^{rc}(X;\beta_{0,1})-\mathbb{E}[w_{0,1}^{rc}(\gamma)\cdot\left(Y-\mu_{0,1}^{rc}(\beta_{0,1})\right)]\right)\dot{\pi}(X;\gamma)\right]\\&\quad-l_{ps}(D,X;\gamma)'\cdot\mathbb{E}\left[\alpha_{ps,0}^{rc}(\gamma)\cdot\left(Y-\mu_{0,0}^{rc}(\beta_{0,0})-\mathbb{E}[w_{0,0}^{rc}(\gamma)\cdot\left(Y-\mu_{0,0}^{rc}(\beta_{0,0})\right)]\right)\dot{\pi}(X;\gamma)\right],\end{aligned}$$

where, for $t=0,1$,

$$\alpha_{ps,t}^{rc}(\gamma)\equiv\alpha_{ps,t}^{rc}(D,T,X;\gamma)=\frac{(1-D)1\{T=t\}}{(1-\pi(X;\gamma))^2}\Big/\mathbb{E}\left[\frac{\pi(X;\gamma)(1-D)1\{T=t\}}{1-\pi(X;\gamma)}\right],$$

and $w_{1,t}^{rc}\equiv w_{1,t}^{rc}(D,T)$, $w_{0,t}^{rc}(\gamma)\equiv w_{0,t}^{rc}(D,T,X;\gamma)$.

The influence function of $\widehat{\tau}_2^{dr,rc}$ is given by

$$\eta_2^{rc}(W;\gamma,\beta)=\eta_1^{rc,2}(W;\beta)-\eta_0^{rc,2}(W;\gamma,\beta_0)-\eta_{est}^{rc,2}(W;\gamma,\beta_0),$$

where

$$\eta_1^{rc,2}(W;\beta)=\eta_{1,1}^{rc,2}(W;\beta)-\eta_{1,0}^{rc,2}(W;\beta), \tag{B.3}$$

$$\eta_0^{rc,2}(W;\gamma,\beta_0)=\eta_0^{rc,1}(W;\gamma,\beta_0), \tag{B.4}$$

$$\eta_{est}^{rc,1}(W;\gamma,\beta_0)=\eta_{est}^{rc,1}(W;\gamma,\beta_0),$$

and, for $d = 0, 1$, $\mu^{rc}_{d,\Delta}(X;\beta_{d,1},\beta_{d,0}) \equiv \mu^{rc}_{d,1}(X;\beta_{d,1}) - \mu^{rc}_{d,0}(X;\beta_{d,0})$, and

$$\begin{aligned}
\eta^{rc,2}_{1,1}(W;\beta^*) &= \frac{D}{\mathbb{E}[D]}\left(\mu^{rc}_{1,\Delta}(X;\beta_{1,1},\beta_{1,0}) - \mathbb{E}\left[\frac{D}{\mathbb{E}[D]}\mu^{rc}_{1,\Delta}(X;\beta_{1,1},\beta_{1,0})\right]\right) \\
&\quad + w^{rc}_{1,1}(D,T)\cdot\left(\left(Y - \mu^{rc}_{1,1}(X;\beta_{1,1})\right) - \mathbb{E}[w^{rc}_{1,1}\cdot\left(Y - \mu^{rc}_{1,1}(X;\beta_{1,1})\right)]\right), \\
\eta^{rc,2}_{1,0}(W;\beta) &= \frac{D}{\mathbb{E}[D]}\left(\mu^{rc}_{0,\Delta}\left(X;\beta^{rc}_{0,1},\beta^{rc}_{0,0}\right) - \mathbb{E}\left[\frac{D}{\mathbb{E}[D]}\mu^{rc}_{0,\Delta}(X;\beta_{0,1},\beta_{0,0})\right]\right) \\
&\quad + w^{rc}_{1,0}(D,T)\cdot\left(Y - \mu^{rc}_{1,0}(X;\beta_{1,0}) - \mathbb{E}[w^{rc}_{1,0}\cdot\left(Y - \mu^{rc}_{1,0}(X;\beta_{1,0})\right)]\right).
\end{aligned}$$

Note that estimating the OR coefficients associated with the treated group does not lead to any estimation effect.

Acknowledgements

We thank the editor, Serena Ng, the associate editor, two anonymous referees, Brantly Callaway, Alex Poirier, Vitor Possebom, Yuya Sasaki, Tymon Słoczyński, Qi Xu, and the audiences of the 2018 SEA conference, 2019 New York Econometrics Camp, and the 2019 IAAE conference for valuable comments. Part of this paper was written when Pedro H. C. Sant'Anna was visiting the Cowles Foundation at Yale University, whose hospitality is gratefully acknowledged.

References

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

Ai, C., and Chen, X. (2003), "Efficient estimation of models with conditional moment restrictions containin unknown functions," *Econometrica*, 71(6), 1795–1843.

Ai, C., and Chen, X. (2007), "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141(1), 5–43.

Ai, C., and Chen, X. (2012), "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170(2), 442–457.

Angrist, J. D., and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist ' s Companion*, Princeton, NJ: Princeton University Press.

Bang, H., and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61(4), 962–972.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017), "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85(1), 233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81(2), 608–650.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, New York: Springer-Verlag.

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.

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

Callaway, B., and Sant'Anna, P. H. C. (2018), "Difference-in-Differences with Multiple Time Periods," *arXiv preprint arXiv:1803.09015*, pp. 1–47.

Cattaneo, M. D. (2010), "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chen, X., and Christensen, T. M. (2018), "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression," *Quantitative Economics*, 9(1), 39–84.

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36(2), 808–843.

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, pp. 1–71.

Dehejia, R., and Wahba, S. (1999), "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

Dehejia, R., and Wahba, S. (2002), "Propensity score-matching methods for nonexperimental causal studies," *The Review of Economics and Statistics*, 84(1), 151–161.

Farrell, M. H. (2015), "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189(1), 1–23.

Firpo, S. (2007), "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75(1), 259–276.

Graham, B., Pinto, C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models with Missing Data," *The Review of Economic Studies*, 79(3), 1053–1079.

Graham, B., Pinto, C., and Egel, D. (2016), "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *Journal of Business and Economic Statistics*, 34(2), 288–301.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Hájek, J. (1971), "Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu," in *Foundations of Statistical Inference*, eds. V. P. Godambe, and D. A. Sprott, Toronto: Holt, Rinehart, and Winston.

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

Heckman, J. J., Ichimura, H., and 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.

Hong, S.-H. (2013), "Measuring the effect of Napster on recorded music sales: difference-in-differences estimates under compositional changes," *Journal of Applied Econometrics*, 28(2), 297–324.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

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

Kang, J. D. Y., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.," *Statistical Science*, 22(4), 569–573.

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76(4), 604–620.

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

Muris, C. (2019), "Efficient GMM Estimation with Incomplete Data," *The Review of Economics and Statistics*, pp. 1–41.

Newey, W. K. (1990), "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

Newey, W. K. (1994), "The asymptotic variance of semiparametric estimators," *Econometrica*, 62(6), 1349–1382.

Newey, W. K., and McFadden, D. (1994), "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Vol. 4, Amsterdam: North-Holland: Elsevier, chapter 36, pp. 2111–2245.

Robins, J. M., Rotnitzky, A., and 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.

Rothe, C., and Firpo, S. (2018), "Properties of Doubly Robust Estimators when Nuisance Functions are Estimated Nonparametrically," *Econometric Theory*, pp. 1–40.

Sant'Anna, P. H. C., Song, X., and Xu, Q. (2018), "Covariate Distribution Balance via Propensity Scores," *Working Paper*, pp. 1–34.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94(448), 1096–1120.

Seaman, S. R., and Vansteelandt, S. (2018), "Introduction to Double Robust Methods for Incomplete Data," *Statistical Science*, 33(2), 184–197.

Słoczyński, T. (2018), "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands," *Working Paper*, .

Słoczyński, T., and Wooldridge, J. M. (2018), "A General Double Robustness Result for Estimating Average Treatment Effects," *Econometric Theory*, 34(1), 112–133.

Smith, J. A., and Todd, P. E. (2005), "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125, 305–353.

Tan, Z. (2019), "Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data," *Annals of Statistics*, .

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

Vermeulen, K., and Vansteelandt, S. (2015), "Bias-Reduced Doubly Robust Estimation," *Journal of the American Statistical Association*, 110(511), 1024–1036.

Wooldridge, J. M. (2007), "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141(2), 1281–1301.

Zimmert, M. (2019), "Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding," *arXiv preprint arXiv: 11809.01643v4*, .