Doubly Robust Difference-in-Differences Estimators*

Pedro H. C. Sant'Anna[†] Jun B. Zhao[‡]

Vanderbilt University Vanderbilt University

September 13, 2019

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. 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.

*First complete version: November 29, 2018. We thank the editor, 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.

[†]Department of Economics, Vanderbilt University. E-mail: pedro.h.santanna@vanderbilt.edu. Part of this article was written when I was visiting the Cowles Foundation at Yale University, whose hospitality is gratefully acknowledged.

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

1 Introduction

Researchers and policy makers are often interested in evaluating the causal effect of a given program or treatment on an outcome of interest. Although randomized experiments are often viewed as the "gold standard" of causal inference, in many situations, researchers only have access to observational data. In such cases, a popular empirical strategy to credibly address the possibility of selection bias is to exploit natural experiments and use difference-in-differences (DID) methods to conduct policy evaluation. 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). This simple procedure can be used with either panel or repeated cross-section data, does not rely on strong functional form assumptions and accounts for time-invariant unobserved confounders. These features highlight the practical appeal of DID methods in many empirical settings.

It is worth mentioning that the causal interpretation of the canonical DID procedure relies on a potentially strong identifying assumption: it requires that, in the absence of the treatment, the average outcome for the treatment and comparison groups would have followed parallel paths over time. This is the so-called (unconditional) parallel trends assumption (PTA). 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.

When the credibility of the PTA is at risk, researchers usually incorporate pre-treatment covariates into the DID analysis and assume that the PTA is satisfied only after conditioning on these covariates, see e.g. Heckman et al. (1997), Heckman et al. (1998), Blundell et al. (2004), Abadie (2005), Bonhomme and Sauder (2011) and Callaway and Sant'Anna (2018) for semi and nonparametric DID procedures. In practice, however, researchers commonly incorporate covariates in DID models in a more straightforward manner, i.e., by adding them linearly in a two-way fixed effects (TWFE) linear regression model. However, this particular TWFE model specification restricts the data generating process beyond the conditional PTA: it implicitly assumes that treatment effects are homogeneous across different subpopulations, and rules out covariate-specific trends in both treated and comparison groups. As a consequence, when treatment effects are heterogeneous, this model specification does not recover the ATT and policy evaluation based on it may be misleading; see e.g. Wooldridge (2005), Chernozhukov et al. (2013) and Słoczyński (2018).

In order to allow for treatment effect heterogeneity and covariate-specific trends under DID designs, researchers can, of course, use alternative, more general estimation procedures than the TWFE specification. For instance, one can consider more flexible outcome regression specifications where interactions between covariates, group, and time indicators are included, see e.g. Meyer (1995), Heckman et al. (1997) and Heckman et al. (1998). Alternatively, one can model the propensity score (i.e., the probability of being in the treatment group given observed covariates) and form inverse probability weighted (IPW) estimators for the ATT, see e.g. Abadie (2005).

It is important to emphasize that the reliability of the outcome regression and the IPW DID approaches hinges on the correct specification of different, non-nested working models: the former depends on the researcher's ability to correctly specify models for the outcome of interested, whereas the latter depends on her ability to model the propensity score correctly. Motivated by this observation, in this paper, we propose that, instead of picking one of the two approaches, one can combine them to form DID estimators for the ATT that enjoy more attractive properties.

Towards this goal, a first contribution of this paper is to derive doubly robust (DR) moments for the ATT under DID settings when either panel or repeated cross-section data are available. When panel data are available, we 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. When only repeated cross-section data are available, we propose two 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. Similarly to the panel data case, we show that both DR DID estimators are consistent for the ATT when either (but not necessarily both) a propensity score working model or a putative model for the outcome evolution for the comparison group is correctly specified. This result is particularly interesting as the DR property does not depend on whether one uses potentially misspecified outcome regression models for the treated group. Thus, in terms of consistency, the price of modelling the evolution of the outcome among treated units seems to be rather low.

A second contribution of this paper is to derive the semiparametric efficiency bounds for the ATT under DID designs in both panel and repeated cross-section data settings. Such efficiency bounds provide a standard to which we can compare the efficiency of any regular semiparametric DID estimator. They also allow us to quantify the loss of efficiency associated with observing repeated cross-section instead of panel data.

Building on these results, we show that our proposed DR DID estimator for the panel data setting attains the semiparametric efficiency bound when both the working model for the propensity score and the working model for the outcome evolution for the comparison group are correctly specified. When only repeated cross-section data are available, we show that our proposed DR DID estimator that depends on working models for the propensity score and for the outcome of both treated and comparison groups in both pre and post treatment periods attains the semiparametric efficiency bound when all these nuisance working models are correctly specified. On the other hand, the "simpler" DR DID estimator that mimics the one with panel data and does not make use of regression models for the outcome of the treated group does not achieve the semiparametric efficiency bound. As we illustrate via Monte Carlo simulations, the loss of efficiency can be of first order importance.

Our proposed methodology accommodates linear and nonlinear working models for the nuisance functions and does not restrict treatment effect heterogeneity beyond the conditional PTA. 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, i.e., their asymptotic linear representation is doubly robust, implying that their associated asymptotic variances do not depend on which models are consistently estimated. 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 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 follows 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.

Finally, 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 builds on Graham et al. (2012), as their proposed propensity score estimator is one important component of our proposal.

Organization of the paper: In the next section, we describe this paper's framework, briefly give and 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, and derive their large sample properties. In Section 4, we discuss the important role that first-step estimators for the nuisance functions can play. We examine the finite sample properties of our proposed methodology by means of a Monte Carlo study in Section 5, and provide an empirical illustration in Section 6. Section 7 concludes. Mathematical proofs are gathered in the Supplemental Appendix[1].

Finally, all proposed policy evaluation tools discussed in this article can be implemented via open-source R package `DRDID`, which is freely available from GitHub (https://github.com/pedrohcgs/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[2]. We formalize these conditions in the following assumption. Let T be a dummy variable that takes value one if the observation belongs to

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

2 Assumption 1(b) below allows for different sampling schemes including the binomial sampling, where $n=n_{T=0}+n_{T=1}$ units are independently sampled from (Y_0,D,X) or (Y_1,D,X) with fixed probability $\lambda\in(0,1)$, and the "conditional" sampling, where $n_{T=1}$ and $n_{T=0}$ units are drawn from (Y_1,D,X) and (Y_1,D,X), respectively, and the deterministic sequence $n_{T=1}/(n_{T=0}+n_{T=1})\rightarrow\lambda\in(0,1)$. On the other hand, Assumption 1(b) rules out settings with non-stationarity / compositional changes in (D,X), see e.g. Hong (2013) for a discussion. A detailed description of this latter setup is beyond the scope of this paper and we leave it for another occasion.

the post-treatment sample and zero if it belongs to the pre-treatment sample.

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) *conditional on* $T = t$*,* $t \in \{0,1\}$*, the data are iid from the distribution of* (Y_t, D, X)*, and* $(D,X)|T = 0 \overset{d}{\sim} (D,X)|T = 1$*.*

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[3]

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

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/group 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).

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 exploits that, under the aforementioned assumptions,

$$\mathbb{E}[Y_1(0)|D = 1] \quad = \quad \mathbb{E}\left[\mathbb{E}[Y_0|D = 1, X] + \mathbb{E}[Y_1|D = 0, X] - \mathbb{E}[Y_0|D = 0, X]\;|D = 1\right]$$

3 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 $\sum_{t=0}^{1} \mathbb{P}(T = t) \cdot \mathbb{P}(Y_t \leq y, D = d, X \leq x|T = t)$.

$$= \mathbb{E}[Y_0|D=1] + \mathbb{E}\left[\mathbb{E}[Y_1|D=0,X] - \mathbb{E}[Y_0|D=0,X]|D=1\right].$$

This representation and (2.1) suggest that one can estimate the ATT using the following OR DID estimator[4]

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

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[5], and $\widehat{\mu}_{d,t}(x)$ is an estimator of the true, unknown $m_{d,t}(x) \equiv \mathbb{E}[Y_t|D=d,X=x]$[6], see e.g. Heckman et al. (1997).

The IPW approach proposed by Abadie (2005) exploits that, under Assumptions 1-3, the ATT can be expressed as

$$\tau = \frac{1}{\mathbb{E}[D]} \mathbb{E}\left[\frac{D-p(X)}{1-p(X)}(Y_1 - Y_0)\right] \tag{2.3}$$

when panel data are available, and as

$$\tau = \frac{1}{\mathbb{E}[D]} \mathbb{E}\left[\frac{D-p(X)}{1-p(X)} \frac{T-\lambda}{\lambda(1-\lambda)} Y\right] \tag{2.4}$$

when repeated cross-section data are available, where $p(X) \equiv \mathbb{P}(D=1|X)$ and $\lambda \equiv \mathbb{P}(T=1)$. 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,

$$\widehat{\tau}^{ipw,p} = \frac{1}{\mathbb{E}_n[D]} \mathbb{E}_n\left[\frac{D-\widehat{\pi}(X)}{1-\widehat{\pi}(X)}(Y_1 - Y_0)\right], \tag{2.5}$$

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. On the other hand, the consistency of the IPW estimator (2.5) relies on the propensity score estimator $\widehat{\pi}(\cdot)$ of $p(\cdot)$ being correctly specified. As a consequence, if the putative model for $p(\cdot)$ is misspecified, the DID IPW estimator (2.5) will, in general, be inconsistent for the ATT. Likewise, if the OR models for $m_{d,t}(\cdot)$

4 Alternatively, one can also use $\widehat{\tau}_2^{reg} = n_{treat}^{-1}\sum_{i|D_i=1}\left[\widehat{\mu}_{1,1}(X_i) - \left(\widehat{\mu}_{1,0}(X_i) + \widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)\right)\right]$ where n_{treat} is the total number of observations with $D=1$.

5 In the panel data case, we have $n_{d,0} = n_{d,1}$, for $d \in \{0,1\}$.

6 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.

are misspecified, the DID OR estimator (2.2) will be inconsistent[7].

Remark 1 It is not uncommon to see practitioners adopting the two-way fixed effects linear regression model

$$Y_{it} = \alpha_1 + \alpha_2 \, T_i + \alpha_3 \, D_i + \tau^{fe} \, (T_i \cdot D_i) + \theta' X_i + \varepsilon_{it}, \tag{2.6}$$

and interpreting estimates of τ^{fe} as estimates of the ATT, see e.g. chapter 5.2 in Angrist and Pischke (2009). Although (2.6) may be perceived as a "natural" specification, it implicitly imposes additional restrictions on the data generating process beyond Assumptions 1-3. More specifically, (2.6) implicitly imposes that $(i)\, \mathbb{E}\left[Y_1(1) - Y_1(0) | X, D = 1\right] = \tau^{fe}$ $a.s.$, 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[8]. 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 5; see also Słoczyński (2018) for related results.

2.2 Doubly robust difference-in-differences estimands

From the discussion in Section 2.1, one can conclude that, as the reliability of ATT estimators based on the OR and IPW approaches rely on different modelling assumptions, it is hard to compare these procedures in terms of robustness against misspecification. In this section, we argue that instead of choosing one of the two 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[9].

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

7 More specifically, the consistency of (2.2) relies on the researchers' ability to model the outcome evolution among the comparison units, $m_{0,1}(\cdot) - m_{0,0}(\cdot)$.

8 Note that under Assumptions 1-3, (2.6) 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.

9 In this section, we discuss DR DID estimands for the ATT, whereas in Section 3, we build on these results and propose DR DID plug-in estimators for the ATT.

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

$$\tau^{dr,p} = \mathbb{E}\left[\left(w^p_1(D) - w^p_0(D,X;\pi)\right)\left(\Delta Y - \mu^p_{0,\Delta}(X)\right)\right], \tag{2.7}$$

where, for a generic g,

$$w^p_1(D) = \frac{D}{\mathbb{E}[D]}, \quad \text{and} \quad w^p_0(D,X;g) = \frac{g(X)(1-D)}{1-g(X)} \Big/ \mathbb{E}\left[\frac{g(X)(1-D)}{1-g(X)}\right]. \tag{2.8}$$

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

$$\tau^{dr,rc}_1 = \mathbb{E}\left[\left(w^{rc}_1(D,T) - w^{rc}_0(D,T,X;\pi)\right)\left(Y - \mu^{rc}_{0,Y}(T,X)\right)\right], \tag{2.9}$$

and

$$\begin{aligned}\tau^{dr,rc}_2 = \tau^{dr,rc}_1 &+ \left(\mathbb{E}\left[\mu^{rc}_{1,1}(X) - \mu^{rc}_{0,1}(X)\middle| D=1\right] - \mathbb{E}\left[\mu^{rc}_{1,1}(X) - \mu^{rc}_{0,1}(X)\middle| D=1, T=1\right]\right) \\ &- \left(\mathbb{E}\left[\mu^{rc}_{1,0}(X) - \mu^{rc}_{0,0}(X)\middle| D=1\right] - \mathbb{E}\left[\mu^{rc}_{1,0}(X) - \mu^{rc}_{0,0}(X)\middle| D=1, T=0\right]\right),\end{aligned} \tag{2.10}$$

where, for a generic g,

$$w^{rc}_1(D,T) = w^{rc}_{1,1}(D,T) - w^{rc}_{1,0}(D,T), \quad \text{and} \quad w^{rc}_0(D,T,X;g) = w^{rc}_{0,1}(D,T,X;g) - w^{rc}_{0,0}(D,T,X;g), \tag{2.11}$$

and, for $t=0,1$,

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

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 $\mu^p_\Delta(X) = m^p_{0,1}(X) - m^p_{0,0}(X)$ a.s.;

(b) When repeated cross-section data are available, $\tau^{dr,rc}_1 = \tau^{dr,rc}_2 = \tau$ if either (but not necessarily both) $\pi(X) = p(X)$ a.s. or $\mu^{rc}_{0,\Delta}(X) = m^{rc}_{0,1}(X) - m^{rc}_{0,0}(X)$ a.s..

Theorem 1 is the first main result of this paper, and provides powerful identification results. It 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 indeed "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, $\mathbb{E}\left[g\left(X\right)|D=1\right]=\mathbb{E}\left[g\left(X\right)|D=1,T=t\right]$, $t=0,1$. Given that this holds for any generic function g, it must also hold for $\mu_{1,t}^{rc}\left(\cdot\right)-\mu_{0,t}^{rc}\left(\cdot\right)$, $t=0,1$, even when $\mu_{d,t}^{rc}\left(\cdot\right)$ are misspecified models of $m_{d,t}^{rc}\left(\cdot\right)$. 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. Given that attractive estimators are not only "robust" but also exploit the available information in an efficient manner, the results in this subsection complement those of the previous one.

Let $m_{0,\Delta}^{p}\left(x\right)\equiv m_{0,1}^{p}\left(x\right)-m_{0,0}^{p}\left(x\right)$, and, for $d=0,1$, $m_{d,\Delta}^{rc}\left(X\right)\equiv m_{d,1}^{rc}\left(X\right)-m_{d,0}^{rc}\left(X\right)$. Recall that $\lambda\equiv\mathbb{P}\left(T=1\right)$.

Theorem 2 *Let Assumptions 1-3 hold. Then:*

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

$$\eta^{e,p}\left(Y_1,Y_0,D,X\right)= w_1^p\left(D\right)\left(m_{1,\Delta}^{p}\left(X\right)-m_{0,\Delta}^{p}\left(X\right)-\tau\right)$$

$$+w_1^p(D)\left(\Delta Y-m_{1,\Delta}^p(X)\right)-w_0^p(D,X;p)\left(\Delta Y-m_{0,\Delta}^p(X)\right), \quad (2.12)$$

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

$$\mathbb{E}\left[\eta^{e,p}(Y_1,Y_0,D,X)^2\right]=\frac{1}{\mathbb{E}[D]^2}\mathbb{E}\left[D\left(m_{1,\Delta}^p(X)-m_{0,\Delta}^p(X)-\tau\right)^2\right.$$
$$\left.+D\left(\Delta Y-m_{1,\Delta}^p(X)\right)^2+\frac{(1-D)\,p(X)^2}{(1-p(X))^2}\left(\Delta Y-m_{0,\Delta}^p(X)\right)^2\right]. \quad (2.13)$$

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

$$\begin{aligned}\eta^{e,rc}(Y,D,T,X)=\;&\frac{D}{\mathbb{E}[D]}\left(m_{1,\Delta}^{rc}(X)-m_{0,\Delta}^{rc}(X)-\tau\right)\\&+\left(w_{1,1}^{rc}(D,T)\left(Y-m_{1,1}^{rc}(X)\right)-w_{1,0}^{rc}(D,T)\left(Y-m_{1,0}^{rc}(X)\right)\right)\\&-\left(w_{0,1}^{rc}(D,T,X;p)\left(Y-m_{0,1}^{rc}(X)\right)-w_{0,0}^{rc}(D,T,X;p)\left(Y-m_{0,0}^{rc}(X)\right)\right), \quad (2.14)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left[\eta^{e,rc}(Y,D,T,X)^2\right]=\;&\frac{1}{\mathbb{E}[D]^2}\mathbb{E}\left[D\left(m_{1,\Delta}^{rc}(X)-m_{0,\Delta}^{rc}(X)-\tau\right)^2\right.\\&+\frac{DT}{\lambda^2}\left(Y-m_{1,1}^{rc}(X)\right)^2+\frac{D(1-T)}{(1-\lambda)^2}\left(Y-m_{1,0}^{rc}(X)\right)^2\\&\left.+\frac{(1-D)\,p(X)^2\,T}{(1-p(X))^2\lambda^2}\left(Y-m_{0,1}^{rc}(X)\right)^2+\frac{(1-D)\,p(X)^2(1-T)}{(1-p(X))^2(1-\lambda)^2}\left(Y-m_{0,0}^{rc}(X)\right)^2\right]. \quad (2.15)\end{aligned}$$

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, $m_{1,\Delta}(X)-m_{0,\Delta}(X)$[10]. The second and third terms in (2.12) and (2.14) 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 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 (2.14) depends on different weights for each pair $(D,T)\in\{0,1\}^2$.

It is also worth mentioning that the efficient influence functions (2.12) and (2.14) 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

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

emphasizing that the efficient influence function for the ATT when panel data are available *does not*

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

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 Theorem 2 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[11].

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[D\left(Y_0 - m_{1,0}(X)\right)^2 + \frac{p(X)^2}{1-p(X)}\left(Y_0 - m_{0,0}(X)\right)^2\right] = \\
&\qquad \mathbb{E}\left[\frac{(1-D)\,p(X)^2}{(1-p(X))^2}\left(Y_1 - m_{1,1}(X)\right)^2 + \frac{p(X)^2}{1-p(X)}\left(Y_1 - m_{0,1}(X)\right)^2\right], \quad (2.16)
\end{aligned}$$

we can show that $\lambda = 0.5$ is "optimal". However, when (2.16) 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.

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

3 Estimation and inference

In this section, we build on the DR DID estimands in Theorem 1 and the semiparametric efficiency bounds in Theorem 2, and discuss estimation and inference procedures for the ATT in DID designs. Indeed, the moment equations (2.7), (2.9), and (2.10) suggest a simple two-step strategy to estimate the ATT based on the analogy principle. In the first step, you estimate 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, you plug 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[12].

In the case when panel data are available, our proposed DR DID estimator for the ATT is based on (2.7) and is given by

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

where

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

$\widehat{\gamma}$ 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, $\mu_{0,\Delta}^{p}(\cdot;\beta_0,\beta_1)=\mu_{0,1}^{p}(\cdot;\beta_1)-\mu_{0,0}^{p}(\cdot;\beta_0)$.

12 Let $g(x)$ be a generic notation for $p(x)$, $m_{d,t}^{l}(X)$, $m_{d,t}^{l}(X)$, $d,t=0,1$, $l=p,rc$. 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".

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 (2.9) 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 (2.11), 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 (2.10) 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 (2.11), but with $\pi(x;\widehat{\gamma})$ playing the role of g.

In what follows, we derive the asymptotic properties of $\widehat{\tau}^{dr,p}$, $\widehat{\tau}_1^{dr,rc}$ and $\widehat{\tau}_2^{dr,rc}$. To do so, we impose some relatively weak, high-level conditions on the parametric first step estimators, and also impose some weak integrability conditions such that we can appropriately apply law of large numbers and central limit theorems. Given that such assumptions are standard in the literature, we list them in the Appendix 7.

3.1 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

$$\eta^p(W;\gamma,\beta) = \eta_1^p(W;\beta_0) - \eta_0^p(W;\gamma,\beta_0) - \eta_{est}^p(W;\gamma,\beta_0), \tag{3.5}$$

where

$$\eta_1^p(W;\beta_0) \quad = \quad w_1^p\cdot\left[\left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right) - \mathbb{E}\left[w_1^p\cdot\left(\Delta Y - \mu_{0,\Delta}^p(\beta_0)\right)\right]\right],$$

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

and

$$\eta_{est}^p(W;\gamma,\beta_0) = l_{reg}(\beta_0)'\cdot\mathbb{E}\left[\left(w_1^p-w_0^p(\gamma)\right)\cdot\dot{\mu}_{0,\Delta}^p(\beta_0)\right] + l_{ps}(\gamma)'\cdot\mathbb{E}\left[\alpha_{ps}^p(\gamma)\left(\left(\Delta Y-\mu_{0,\Delta}^p(\beta_0)\right)-\mathbb{E}\left[w_0^p(\gamma)\cdot\left(\Delta Y-\mu_{0,\Delta}^p(\beta_0)\right)\right]\right)\cdot\dot{\pi}(\gamma)\right], \quad (3.6)$$

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) in the Appendix, $l_{ps}(\cdot)$ 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:

$$\exists\gamma^* \in \Theta^{ps}:\mathbb{P}(\pi(X;\gamma^*)=p(X))=1, \quad (3.7)$$

$$\exists\left(\beta_{0,1}^{*,p},\beta_{0,0}^{*,p}\right) \in \Theta_{0,1}^{reg}\times\Theta_{0,0}^{reg}:\mathbb{P}\left(\mu_{0,1}^p\left(X;\beta_{0,1}^{*,p}\right)-\mu_{0,0}^p\left(X;\beta_{0,0}^{*,p}\right)=m_{0,1}^p(X)-m_{0,0}^p(X)\right)=1. \quad (3.8)$$

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

Theorem 3 *Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix 7 hold.*

(a) Provided that either (3.7) or (3.8) is true, as $n\to\infty$,

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

Furthermore,

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

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

(b) When both (3.7) and (3.8) are true, $\eta^p\left(W;\gamma^*,\beta_0^{*,p}\right)=\eta^{e,p}(Y_1,Y_0,D,X)$ a.s. and V^p is equal to the semiparametrically efficiency bound (2.13).

Theorem 3 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 3 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 3 also suggests that one can use the analogy principle to estimate V^p and conduct asymptotically valid inference[13]. 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^{p}_{est} when estimating V^p. Failing to do so may lead to asymptotically invalid inference procedures.

3.2 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

$$\eta^{rc}_{j}\left(W;\gamma,\beta\right) = \eta^{rc,j}_{1}\left(W;\beta\right) - \eta^{rc,j}_{0}\left(W;\gamma,\beta\right) - \eta^{rc,j}_{est}\left(W;\gamma,\beta\right), \tag{3.9}$$

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

$$\begin{aligned}
\eta^{rc,j}_{1}\left(W;\beta\right) &= \eta^{rc,j}_{1,1}\left(W;\beta\right) - \eta^{rc,j}_{1,0}\left(W;\beta\right), \\
\eta^{rc,j}_{0}\left(W;\gamma,\beta\right) &= \eta^{rc,j}_{0,1}\left(W;\gamma,\beta\right) - \eta^{rc,j}_{0,0}\left(W;\gamma,\beta\right), \\
\eta^{rc,j}_{est}\left(W;\gamma,\beta\right) &= \eta^{rc,j}_{est,reg}\left(W;\gamma,\beta\right) + \eta^{rc,j}_{est,ps}\left(W;\gamma,\beta\right),
\end{aligned}$$

and the precise definitions of all these η^{rc} functions are deferred to Appendix 7 to avoid excess notational complexity. An aspect of the difference between η^{rc}_1 and η^{rc}_2 that is worth mentioning but is perhaps buried in the notation is that η^{rc}_1 depends on β only through β_0, whereas η^{rc}_2 depends on both β_1 and β_0. This is simply a consequence from the fact that $\widehat{\tau}^{dr,rc}_1$ does not rely on outcome regressions for the treated units, but $\widehat{\tau}^{dr,rc}_2$ does.

Consider the following claims:

$$\exists\left(\beta^{*,rc}_{0,1}, \beta^{*,rc}_{0,0}\right) \in \Theta^{reg}_{0,1} \times \Theta^{reg}_{0,0} : \mathbb{P}\left(\mu^{rc}_{0,1}\left(X;\beta^{*,rc}_{0,1}\right) - \mu^{rc}_{0,0}\left(X;\beta^{*,rc}_{0,0}\right) = m^{rc}_{0,1}\left(X\right) - m^{rc}_{0,0}\left(X\right)\right) = 1, \tag{3.10}$$

$$\forall\left(d,t\right) \in \{0,1\}^2\ \exists\left(\beta^{*,rc}_{d,t}\right) \in \Theta^{reg}_{d,t} : \mathbb{P}\left(\mu^{rc}_{d,t}\left(X;\beta^{*,rc}_{d,t}\right) = m^{rc}_{d,t}\left(X\right)\right) = 1. \tag{3.11}$$

Theorem 4 *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*

---

13 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.

7 hold, and that $n_1/n \xrightarrow{p} \lambda \in (0,1)$ as $n_0, n_1 \to \infty$.

(a) Provided that either (3.7) or (3.10) 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 (3.7) and (3.11) 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 (2.15). On the other hand, V_1^{rc} does not attain the semiparametric efficiency bound when (3.7) and (3.11) are true.

In other words, Theorem 4 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 4 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 3(b), it is evident that such a requirement is stronger than when panel data are available.

Finally, it is worth emphasizing that, although $\widehat{\tau}_2^{dr,rc}$ is locally semiparametrically efficient when all working models are correctly specified, in general, $\widehat{\tau}_1^{dr,rc}$ does not attain the semiparametric efficiency bound. In the next corollary, we state precisely the potential loss of efficiency associated with using $\widehat{\tau}_1^{dr,rc}$ instead of $\widehat{\tau}_2^{dr,rc}$, when all nuisance working models are correctly specified. Such a result is useful in the sense that it quantifies the efficiency loss of using an estimator analogue to $\widehat{\tau}^{dr,p}$ when only repeated cross-section data are available.

Corollary 2 *Let the assumptions of Theorem 4 hold, and assume that both (3.7) and (3.11) are true. Then,*

$$V_1^{rc} - V_2^{rc} = \mathbb{E}[D]^{-1} \cdot Var\left[\sqrt{\frac{1-\lambda}{\lambda}}\left(m_{1,1}^{rc}(X) - m_{0,1}^{rc}(X)\right) + \sqrt{\frac{\lambda}{1-\lambda}}\left(m_{1,0}^{rc}(X) - m_{0,0}^{rc}(X)\right) \middle| D = 1\right] \geq 0.$$

4 On the choice of first-step estimators

Until now, we have shown that our proposed DID estimators are doubly robust consistent, and when all the working models are correctly specified, $\widehat{\tau}^{dr,p}$ and $\widehat{\tau}_2^{dr,rc}$ are locally semiparametric efficient. It is also worth emphasizing that our proposed DR DID estimators are generic in the sense that they accommodate linear and nonlinear models for the nuisance functions, covariates can enter into these nuisance models in an asymmetric manner, and the nuisance parameters can be estimated using a variety of estimation methods. On one hand, this level of generality in terms of first-step estimators can be seen as an attractive feature of our proposed generic DR DID estimators. On the other hand, given that the generic DR DID estimators are not DR for inference, the choice of estimation method in the first-step may have an important impact on inference procedures, especially under model misspecifications; see e.g. Kang and Schafer (2007).

In this section, we aim to provide further guidance on the choice of first-step estimators in order to construct DR DID estimators that are also DR for inference, i.e., their asymptotic linear representation is also doubly robust. To do so, 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 Section 3, they are much weaker than those implicitly imposed in the TWFE specification (2.6), 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. For the sake of space, in what follows, we focus on the case where panel data are available. The analysis for the case where only repeated cross-section data are available is deferred to Appendix 7.

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

$$\pi(X,\gamma) = \Lambda\left(X'\gamma\right) \equiv \frac{\exp\left(X'\gamma\right)}{1+\exp\left(X'\gamma\right)}, \text{ and } \mu_{0,\Delta}^{p}\left(X;\beta_{0,1}^{p},\beta_{0,1}^{p}\right) = \mu_{0,\Delta}^{lin,p}\left(X;\beta_{0,\Delta}^{p}\right) \equiv X'\beta_{0,\Delta}^{p}. \quad (4.1)$$

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

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

where the first two-steps consist of computing

$$\widehat{\gamma}^{ipt} \quad = \quad \arg\max_{\gamma\in\Gamma} \mathbb{E}_n\left[DX'\gamma - (1-D)\exp\left(X'\gamma\right)\right],$$

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

while in the third and last step, one plugs the fitted values of the working models (4.1) 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, note that the results of Theorem 3 imply that the key to forming estimators for the ATT that are DR for inference is to obtain $\eta^{p}_{est}\left(W;\gamma^{*},\beta^{*}\right)=0$ a.s., where γ^{*} and β^{*} are the probability limits of the possibly misspecified first-step estimators used in constructing $\widehat{\tau}^{dr,p}$. By paying closer attention to $\eta^{p}_{est}\left(W;\gamma^{*},\beta^{*}\right)$ in (3.6), it is clear that if

$$\begin{aligned}
\mathbb{E}\left[\left(w^{p}_{1}-w^{p}_{0}\left(\gamma^{*}\right)\right)\cdot\dot{\mu}^{p}_{0,\Delta}\left(\beta^{*}\right)\right] &= 0,\\
\mathbb{E}\left[\frac{(1-D)}{(1-\pi(X;\gamma^{*}))^{2}}\left(\Delta Y-\mu^{p}_{0,\Delta}\left(\beta^{*}\right)\right)\cdot\dot{\pi}(\gamma^{*})\right] &= 0, \qquad (4.2)\\
\mathbb{E}\left[w^{p}_{0}\left(\gamma^{*}\right)\cdot\left(\Delta Y-\mu^{p}_{0,\Delta}\left(\beta^{*}\right)\right)\right] &= 0,
\end{aligned}$$

then η^{p}_{est} is asymptotically negligible. As the first component of X is a constant and we adopt the working models (4.1), it follows that (4.2) reduces to

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

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 $\eta^{p}_{est}\left(W;\gamma^{*,ipt},\beta^{*,wls,p}_{0,\Delta}\right)=0\ a.s.$, where $\gamma^{*,ipt}$ and $\beta^{*,wls,p}_{0,\Delta}$ are the pseudo-true parameters associated with $\widehat{\gamma}^{ipt}$ and $\widehat{\beta}^{wls,p}_{0,\Delta}$, respectively[14].

The next proposition formalizes this discussion. Define

$$\tau^{dr,p}_{imp}=\mathbb{E}\left[\left(w^{p}_{1}\left(D\right)-w^{p}_{0}\left(D,X;\gamma^{*,ipt}\right)\right)\left(\Delta Y-\mu^{lin,p}_{0,\Delta}\left(X;\beta^{*,wls,p}_{0,\Delta}\right)\right)\right], \qquad (4.3)$$

and let

$$\eta^{dr,p}_{imp}\left(W;\gamma^{*,ipt},\beta^{*,wls,p}_{0,\Delta},\tau^{dr,p}_{imp}\right)=\left(w^{p}_{1}\left(D\right)-w^{p}_{0}\left(D,X;\gamma^{*,ipt}\right)\right)\left(\Delta Y-\mu^{lin,p}_{0,\Delta}\left(X;\beta^{*,wls,p}_{0,\Delta}\right)\right)-w^{p}_{1}\left(D\right)\cdot\tau^{dr,p}_{imp}.$$

14 More formally, $\gamma^{*,ipt}=\arg\max_{\gamma\in\Gamma}\mathbb{E}\left[DX'\gamma-(1-D)\exp\left(X'\gamma\right)\right]$ and $\beta^{*,wls,p}_{0,\Delta}=\arg\min_{b\in\Theta}\mathbb{E}\left[\frac{\Lambda\left(X'\gamma^{*,ipt}\right)}{1-\Lambda\left(X'\gamma^{*,ipt}\right)}\left(\Delta Y-X'b\right)^2\middle|D=0\right]$.

Proposition 1 *Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix 7 hold, and that the working nuisance models (4.1) are adopted. Then,*

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

where $V_{imp}^{p}=\mathbb{E}\left[\eta_{imp}^{dr,p}\left(W;\gamma^{*,ipt},\beta_{0,\Delta}^{*,wls,p},\tau_{imp}^{dr,p}\right)^2\right]$.

Proposition 1 generalizes the cross-section results of Vermeulen and Vansteelandt (2015) to the DID framework. An important consequence of Proposition 1 is that one can treat the summands of $\widehat{\tau}_{imp}^{dr,p}$ as if they were independent and identically distributed, and, therefore, can estimate V_{imp}^{p} by

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

This simple but powerful result greatly simplifies inference procedures.

Finally, note that when one combines Proposition 1 with Theorem 3, we get the third main result of this paper, that is, by using the DR moment for the ATT and appropriately choosing the estimation method for the first-step nuisance parameters, we have that $\widehat{\tau}_{imp}^{dr,p}$ is not only DR consistent and locally semiparametrically efficient, but also DR for inference.

Remark 2 From the discussion above, it may be natural to directly use the moment conditions (4.2) to from (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 (4.2) 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 (4.1) 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.

5 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. In particular, 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 (4.3), respectively, to the OR DID estimator (2.2), the Horvitz and Thompson (1952) type IPW estimator (2.5), and the TWFE regression model (2.6). Given that the weights of the IPW estimator (2.5) 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

$$\widehat{\tau}^{ipw,p}_{std} = \mathbb{E}_n\left[\left(\widehat{w}_1^p\left(D\right) - \widehat{w}_0^p\left(D,X;\widehat{\gamma}\right)\right)\left(Y_1 - Y_0\right)\right], \tag{5.1}$$

where the weights $\widehat{w}_1^p\left(D\right)$ and $\widehat{w}_0^p\left(D,X;\widehat{\gamma}\right)$ 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 (C.2) and (C.3) in the Appendix 7, to the OR DID estimator (2.2), the plug-in IPW estimator based on (2.4), and the TWFE regression model (2.6). As in the case of panel data, we also consider the Hájek (1971) type IPW estimator for the ATT

$$\widehat{\tau}^{ipw,rc}_{std} = \mathbb{E}_n\left[\left(\widehat{w}_1^{rc}\left(D,T\right) - \widehat{w}_0^{rc}\left(D,T,X;\widehat{\gamma}\right)\right)Y\right], \tag{5.2}$$

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 4 and in Appendix 7, 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, root mean square error (RMSE), empirical 95% coverage probability, the average length of a 95% confidence interval, and the average of their standard errors. The confidence intervals are based on the normal approximation, with the asymptotic variances being estimated by their sample analogues. Finally, we emphasize that our measures of performance highlight not only the behavior of DID point estimates but also the accuracy of their associated inference procedures.

5.1 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):

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

where ε_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 $\left\{Y_{0,i}, Y_{1,i}, D_i, Z_i\right\}_{i=1}^{n}$, 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

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

	DGP1: OR correct, PS correct					DGP2: OR correct, PS incorrect				
	Bias	RMSE	Std. error	Coverage	CI length	Bias	RMSE	Std. error	Coverage	CI length
$\widehat{\tau}^{fe}$	-20.9518	21.1227	2.5271	0.0000	9.9061	-19.2859	19.4683	2.5754	0.0000	10.0955
$\widehat{\tau}^{reg}$	-0.0012	0.1005	0.1010	0.9500	0.3960	-0.0008	0.0997	0.1004	0.9492	0.3937
$\widehat{\tau}^{ipw,p}$	0.0257	2.7743	2.6636	0.9518	10.4412	2.0100	3.2982	2.5049	0.8376	9.8193
$\widehat{\tau}^{ipw,p}_{std}$	0.0075	1.1320	1.0992	0.9476	4.3090	-0.7942	1.2253	0.9241	0.8564	3.6226
$\widehat{\tau}^{dr,p}$	-0.0014	0.1059	0.1052	0.9473	0.4124	-0.0008	0.1036	0.1031	0.9469	0.4043
$\widehat{\tau}^{dr,p}_{imp}$	-0.0013	0.1057	0.1043	0.9451	0.4088	-0.0007	0.1042	0.1030	0.9445	0.4039
	DGP3: OR incorrect, PS correct					DGP4: OR incorrect, PS incorrect				
	Bias	RMSE	Std. error	Coverage	CI length	Bias	RMSE	Std. error	Coverage	CI length
$\widehat{\tau}^{fe}$	-13.1703	13.3638	3.5611	0.0035	13.9596	-16.3846	16.5383	3.6268	0.0000	14.2169
$\widehat{\tau}^{reg}$	-1.3843	1.8684	1.2286	0.8001	4.8159	-5.2045	5.3641	1.2890	0.0145	5.0531
$\widehat{\tau}^{ipw,p}$	0.0114	3.1982	3.0043	0.9468	11.7769	-1.0846	2.6557	2.3746	0.9487	9.3084
$\widehat{\tau}^{ipw,p}_{std}$	-0.0299	1.4270	1.3990	0.9447	5.4840	-3.9538	4.2154	1.4585	0.2282	5.7172
$\widehat{\tau}^{dr,p}$	-0.0513	1.2142	1.1768	0.9416	4.6132	-3.1878	3.4544	1.2946	0.3076	5.0749
$\widehat{\tau}^{dr,p}_{imp}$	-0.0709	1.0151	0.9842	0.9423	3.8581	-2.5291	2.7202	0.9837	0.2737	3.8561

Notes: Simulations based on 10,000 Monte Carlo experiments. $\widehat{\tau}^{fe}$ is the TWFE outcome regression estimator of τ^{fe} in (2.6), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,p}$ is the IPW DID estimator (2.5), $\widehat{\tau}^{ipw,p}_{std}$ is the standardized IPW DID estimator (5.1), $\widehat{\tau}^{dr,p}$ is our proposed DR DID estimator (3.1), and $\widehat{\tau}^{dr,p}_{imp}$ is our proposed DR DID estimator (4.3). 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 4. Finally, "Bias", "RMSE", "Std. error", "Coverage" and "CI length', stand for the average simulated bias, simulated root mean-squared errors, plug-in standard errors, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

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 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, standard error, 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. 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 and the standard error of $\widehat{\tau}^{ipw,rc}$ are more than two 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 $\widehat{\tau}^{ipw,p}$.

Finally, 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.

5.2 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. This binomial sampling scheme is consistent with Assumption 1(b).

Table 2 present the simulation results with $\lambda = 0.5$ and with $n \equiv n_1 + n_0 = 1{,}000$ [15]. Overall, the simulation exercise reveals that the RMSE, standard errors, 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

15 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					DGP2: OR correct, PS incorrect				
	Bias	RMSE	Std. error	Coverage	CI length	Bias	RMSE	Std. error	Coverage	CI length
$\widehat{\tau}^{fe}$	-20.7916	21.0985	3.5705	0.0002	13.9962	-19.1783	19.5289	3.6345	0.0005	14.2472
$\widehat{\tau}^{reg}$	0.0263	7.5878	7.5702	0.9510	29.6751	-0.0244	8.1906	8.1493	0.9481	31.9454
$\widehat{\tau}^{ipw,rc}$	-0.6619	55.9708	55.5516	0.9493	217.7621	1.8203	55.0496	54.9614	0.9491	215.4486
$\widehat{\tau}^{ipw,rc}_{std}$	-0.0502	9.6477	9.5815	0.9487	37.5596	-0.8119	9.8141	9.7018	0.9459	38.0310
$\widehat{\tau}^{dr,rc}_{1}$	0.0129	3.0414	3.0340	0.9504	11.8934	-0.0102	3.2814	3.2651	0.9486	12.7991
$\widehat{\tau}^{dr,rc}_{2}$	0.0041	0.2159	0.2102	0.9441	0.8239	-0.0002	0.2108	0.2054	0.9454	0.8051
$\widehat{\tau}^{dr,rc}_{1,imp}$	0.0136	3.0413	3.0337	0.9507	11.8921	-0.0095	3.2818	3.2650	0.9488	12.7989
$\widehat{\tau}^{dr,rc}_{2,imp}$	0.0047	0.2163	0.2049	0.9371	0.8032	0.0002	0.2127	0.2030	0.9403	0.7958
	DGP3: OR incorrect, PS correct					**DGP4: OR incorrect, PS incorrect**				
	Bias	RMSE	Std. error	Coverage	CI length	Bias	RMSE	Std. error	Coverage	CI length
$\widehat{\tau}^{fe}$	-13.1310	14.0577	5.0424	0.2598	19.7664	-16.3305	17.1263	5.1307	0.1138	20.1123
$\widehat{\tau}^{reg}$	-1.3763	8.1367	8.0046	0.9421	31.3782	-5.3378	9.9773	8.5196	0.9075	33.3969
$\widehat{\tau}^{ipw,rc}$	-0.9734	57.2618	56.9005	0.9465	223.0501	-1.3912	55.1777	55.6717	0.9518	218.2330
$\widehat{\tau}^{ipw,rc}_{std}$	0.0508	9.4283	9.3068	0.9431	36.4826	-4.1487	10.5195	9.6864	0.9304	37.9707
$\widehat{\tau}^{dr,rc}_{1}$	-0.0855	5.6917	5.6276	0.9453	22.0602	-3.3422	7.0709	6.1963	0.9157	24.2897
$\widehat{\tau}^{dr,rc}_{2}$	-0.0289	4.7419	4.6585	0.9416	18.2613	-3.2751	6.0158	4.8876	0.8863	19.1593
$\widehat{\tau}^{dr,rc}_{1,imp}$	-0.1191	4.8371	4.7970	0.9450	18.8042	-2.6888	5.5642	4.8416	0.9134	18.9790
$\widehat{\tau}^{dr,rc}_{2,imp}$	-0.0762	4.0623	3.9669	0.9436	15.5503	-2.6138	4.8453	3.9673	0.8923	15.5519

Notes: Simulations based on 10,000 Monte Carlo experiments. $\widehat{\tau}^{fe}$ is the TWFE outcome regression estimator of τ^{fe} in (2.6), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,rc}$ is the IPW DID estimator based on the sample analogue of (2.4), $\widehat{\tau}^{ipw,rc}_{std}$ is the standardized IPW DID estimator (5.2), 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 (C.2) and (C.3) in the Appendix 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 $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ where we use the estimation methods described in the Appendix 7. Finally, "Bias", "RMSE", "Std. error", "Coverage" and "CI length', stand for the average simulated bias, simulated root mean-squared errors, plug-in standard errors, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

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 are correctly specified. Finally, 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.

6 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 [16].

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 (2.6), 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 (2.5). We also consider the Hájek (1971) type IPW estimator $\widehat{\tau}^{ipw,p}_{std}$ as defined in (5.1). 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 4.

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.

16 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 (2.6), $\widehat{\tau}^{reg}$ is the OR-DID estimator (2.2), $\widehat{\tau}^{dr,p}$ is the IPW DID estimator (2.5), $\widehat{\tau}^{dr,p}_{std}$ is the standardized IPW DID estimator (5.1), $\widehat{\tau}^{dr,p}$ is our proposed DR DID estimator (3.1), and $\widehat{\tau}^{dr,p}_{imp}$ is our proposed DR DID estimator (4.3). 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 4. 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.

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

7 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 $\mathrm{CATT}(X_1) \equiv \mathbb{E}\left[Y(1) - Y(0) \,|X_1, D=1\right]$ 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. We leave the detailed analysis of these extensions to future work.

Appendix A: Regularity assumptions on first-step estimators

Let $g(x)$ be a generic notation for $\pi(x)$, $\mu_{d,t}^{p}(x)$ and $\mu_{d,t}^{rc}(x)$, $d,t=0,1$. Analogously and with some abuse of notation, let $g(x;\theta)$ be a generic notation for $\pi(x;\gamma)$, $\mu_{d,t}^{p}\left(x,\beta_{d,t}^{p}\right)$ and $\mu_{d,t}^{rc}\left(x,\beta_{d,t}^{rc}\right)$, $d,t=0,1$. 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

$$
\begin{aligned}
h^{p}\left(W;\kappa^{p}\right) &= \left(w_{1}^{p}(D)-w_{0}^{p}(D,X;\gamma)\right)\left(\Delta Y-\mu_{0,\Delta}^{p}(X;\beta_{0,0}^{p},\beta_{0,1}^{p})\right),\\
h^{rc\ 1}\left(W;\kappa^{rc\ 1}\right) &= \left(w_{1}^{rc}(D,T)-w_{0}^{rc}(D,T,X;\gamma)\right)\left(Y-\mu_{0,Y}^{rc}\left(T,X;\beta_{0,0}^{rc},\beta_{0,1}^{rc}\right)\right),\\
h^{rc\ 2}\left(W;\kappa^{rc\ 2}\right) &= \left(D/\mathbb{E}[D]\right)\cdot\left(\mu_{1,\Delta}^{rc}\left(X;\beta_{1,1}^{rc},\beta_{1,0}^{rc}\right)-\mu_{0,\Delta}^{rc}\left(X;\beta_{0,1}^{rc},\beta_{0,0}^{rc}\right)\right)\\
&\quad +w_{1,1}^{rc}(D,T)\left(Y-\mu_{1,1}^{rc}\left(X;\beta_{1,1}^{rc}\right)\right)-w_{1,0}^{rc}(D,T)\left(Y-\mu_{1,0}^{rc}\left(X;\beta_{1,0}^{rc}\right)\right)\\
&\quad -\left(w_{0,1}^{rc}(D,T,X;\gamma)\left(Y-\mu_{0,1}^{rc}\left(X;\beta_{0,1}^{rc}\right)\right)-w_{0,0}^{rc}(D,T,X;\gamma)\left(Y-\mu_{0,0}^{rc}\left(X;\beta_{0,0}^{rc}\right)\right)\right)
\end{aligned}
$$

where $\kappa^{p}=\left(\gamma',\beta_{0,0}^{p\prime},\beta_{0,1}^{p\prime}\right)'$, $\kappa^{rc\ 1}=\left(\gamma',\beta_{0,0}^{rc\prime},\beta_{0,1}^{rc\prime}\right)'$ and $\kappa^{rc\ 2}=\left(\gamma',\beta_{0,0}^{rc\prime},\beta_{0,1}^{rc\prime},\beta_{1,1}^{rc\prime},\beta_{1,0}^{rc\prime}\right)'$. In obvious notation, the vector of pseudo-true parameter[17] is given by $\kappa^{*,p}$, $\kappa_{0}^{*,rc\ 1}$, 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.

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

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 $\mathbb{E}\left[l_g(W;\theta^*)\right] = 0$, $\mathbb{E}\left[l_g(W;\theta^*)\, l_g(W;\theta^*)'\right]$ exists and is positive definite and $\lim_{\delta \to 0} \mathbb{E}\left[\sup_{\theta \in \Theta^*: \|\theta - \theta^*\| \le \delta} \left\| l_g(W;\theta) - l_g(W;\theta^*)\right\|^2\right] = 0$. In addition, (vi) for some $\varepsilon > 0$, $0 < \pi(X;\gamma) \le 1 - \varepsilon$ a.s., for all $\gamma \in int(\Theta^{ps})$, 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 $\mathbb{E}\left[\sup_{\kappa \in \Gamma^{*,p}} \left|\dot{h}^p(W;\kappa)\right|\right] < \infty$, 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 $\mathbb{E}\left[\sup_{\kappa \in \Gamma^{*,rc\ j}} \left|\dot{h}^{rc,j}(W;\kappa)\right|\right] < \infty$, 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 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).

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

As it is evident from Theorem 4, 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 (3.9).

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

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

where, for $t = 0,1$,

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

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

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

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

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

Appendix C: Improved DR DID estimators when repeated cross-section data are available

In Section 4, we have shown that, when panel data are available, by paying particular attention to the estimation method used to estimate the nuisance parameters, one can form further improved DR DID estimators for the ATT, i.e., estimators that are not only DR consistent and locally semiparametrically efficient, but also DR for inference. In this section, we show the analogous results to the case where one only has access to repeated cross-section data.

As in Section 4, 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. More precisely, we consider the case where

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

We propose two "improved" DR DID estimators, namely

$$
\widehat{\tau}_{1,imp}^{dr,rc}=\mathbb{E}_n\left[\left(\widehat{w}_1^{rc}(D,T)-\widehat{w}_0^{rc}\left(D,T,X;\widehat{\gamma}^{ipt}\right)\right)\left(Y-\mu_{0,Y}^{lin,rc}\left(X;\widehat{\beta}_{0,1}^{wls,rc},\widehat{\beta}_{0,0}^{wls,rc}\right)\right)\right], \tag{C.2}
$$

and

$$
\begin{aligned}
\widehat{\tau}_{2,imp}^{dr,rc}=\widehat{\tau}_{1,imp}^{dr,rc} & +\left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]}-\widehat{w}_{1,1}^{rc}(D,T)\right)\left(\mu_{1,1}^{rc}\left(X;\widehat{\beta}_{1,1}^{ols,rc}\right)-\mu_{0,1}^{rc}\left(X;\widehat{\beta}_{0,1}^{wls,rc}\right)\right)\right]\right) \\
& -\left(\mathbb{E}_n\left[\left(\frac{D}{\mathbb{E}_n[D]}-\widehat{w}_{1,0}^{rc}(D,T)\right)\left(\mu_{1,0}^{rc}\left(X;\widehat{\beta}_{1,0}^{ols,rc}\right)-\mu_{0,0}^{rc}\left(X;\widehat{\beta}_{0,0}^{wls,rc}\right)\right)\right]\right),
\end{aligned} \tag{C.3}
$$

where

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

Here, note that $\widehat{\tau}^{dr,rc}_{1,imp}$ does not rely on OR models for the treated group while $\widehat{\tau}^{dr,rc}_{2,imp}$ does. In addition, when one compares $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ with $\widehat{\tau}^{dr,p}_{imp}$, 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 $\beta^{*,rc}_{1,t}$, $t=0,1$, does not lead to any estimation effect, and therefore one can choose her favorite estimation method. Given this observation and the linear specification in (C.1), we find it natural to estimate $\beta^{*,rc}_{1,t}$, $t=0,1$, via OLS as this is the most widespread estimation procedure adopted by practitioners.

Let

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

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

$$\begin{aligned}
\eta^{dr,rc}_{1,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right) &= \eta^{rc,1}_1(W;\beta^{*,wls,rc}_{0,1},\beta^{*,wls,rc}_{0,0}) - \eta^{rc,1}_0(W;\gamma^{*,ipt},\beta^{*,wls,rc}_{0,1},\beta^{*,wls,rc}_{0,0}),\\
\eta^{dr,rc}_{2,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right) &= \eta^{rc,2}_1(W;\beta^{*,rc}_{imp}) - \eta^{rc,2}_0(W;\gamma^{*,ipt},\beta^{*,wls,rc}_{0,1},\beta^{*,wls,rc}_{0,0}),
\end{aligned}$$

where $\eta^{rc,1}_1$, $\eta^{rc,1}_0$, $\eta^{rc,2}_1$, and $\eta^{rc,2}_0$ are defined as in (B.1)-(B.4), respectively. Next proposition states that $\widehat{\tau}^{dr,rc}_{1,imp}$ and $\widehat{\tau}^{dr,rc}_{2,imp}$ are indeed doubly robust for inference.

Proposition C.1 *Suppose Assumptions 1-3 and Assumptions A.1-A.2 stated in Appendix 7 hold, and that the working nuisance models (C.1) are adopted. Then,*

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

where $V^{rc}_{1,imp} = \mathbb{E}\left[\eta^{dr,rc}_{1,imp}\left(W;\gamma^{*,ipt},\beta^{*,wls,rc}_{0,1},\beta^{*,wls,rc}_{0,0}\right)^2\right]$.

Analogously,

$$\begin{aligned}
\sqrt{n}(\widehat{\tau}^{dr,rc}_{2,imp} - \tau^{dr,rc}_{imp}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta^{dr,rc}_{2,imp}\left(W;\gamma^{*,ipt},\beta^{*,rc}_{imp}\right) + o_p(1)\\
&\xrightarrow{d} N\left(0,V^{rc}_{2,imp}\right),
\end{aligned}$$

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

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.

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 and an Application on the Minimum Wage and Employment,” *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.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013), “Average and Quantile Effects in Nonseparable Panel Models,” *Econometrica*, 81(2), 535–580.

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.

Meyer, B. D. (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics*, 13(2), 151–161.

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. (2005), "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Review of Economics and Statistics*, 87(2), 385–390.

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