Difference-in-Differences with Compositional Changes*

Pedro H. C. Sant'Anna[†] Qi Xu[‡]

November 11, 2022

Abstract

This paper proposes a doubly robust estimator for the average treatment effect on the treated (ATT) in a difference-in-differences setup. Identification of the ATT is studied without assuming that the covariate distribution stays constant over time. We derive semiparametric efficiency bound for our target policy parameter and show that our proposed estimator attains this semiparametric efficiency bound. A new uniform stochastic expansion of local polynomial logit estimator is provided, based on which, we show the proposed ATT estimator favorable theoretical properties under mild conditions on the first-step convergence speed. Finally, we develop a Hausman-type test for the compositional stationarity assumption. The finite sample performance of the estimator and the test is examined by means of a Monte Carlo experiment and an empirical application.

[*] We thank tXYZ and the seminar and conference participants at many institution for their valuable comments and suggestions.

[†] Microsoft and Vanderbilt University. E-mail: pedro.h.santanna@vanderbilt.edu.

[‡] Department of Economics, Vanderbilt University. E-mail: qi.xu.1@vanderbilt.edu.

1 Introduction

Difference-in-differences (DID) designs has been used widely for identifying and estimating causal effects, such as the average treatment effect on the treated (ATT). Identification with this research design typically relies on assumptions stipulating that, conditional on a set of covariates, the treated and comparison populations would have evolved “in parallel” had there been no treatment. When the set of covariate is rich enough, and in particular, if it accommodates time-varying covariates, this parallel trend assumption is more likely to hold. That being said, it seems a prevalent practice in the DID literature to restrict the time variation of the covariates, in one form or another.

When panel data of representative random samples are available, researchers often limit themselves to covariates that do not vary over time, or are set at their pre-treatment level. Such a restriction can sometimes be justified when non-random sample attrition is not a material concern and if the covariates are not affected by the treatment itself (Caetano, Callaway, Payne and Rodrigues, 2022). On the other hand, when we only observe different samples of a population at different time periods (i.e. repeated cross sections), it is commonly assumed that there is no compositional change among the covariates. This requirement is less plausible than that imposed in the panel setting and it rules out many interesting applications.

For example, Hong (2013) studies the effect of Napster on recorded music sales. He uses data from the 1996–2002 Interview surveys of the Consumer Expenditure Survey (CEX). Over this period, composition of internet users has changed substantially. The small number of early adopters tend to be younger, richer, more educated, and technically savvy, whereas later adopters exhibit higher level of diversity in terms of demographics. If one ignores such imbalances of group composition across time, the (negative) effect of Napster may be overestimated since the decrease in the average music expenditure is likely due to post-Napster group having more households with low reservation prices for recorded music.

For another instance, a large reduction in the average nominal tariff rate between South Africa and Mozambique occurred in 2008. This quasi-experimental variation is used by Sequeira (2016) to study the effect of tariff rate reduction on trade costs and corruption behavior with a DID design. The author collected multiple cross sections of shipping data between 2006 and 2014. Covariates used in the study includes firm-level characteristics, such as firm size and ownership structure, and the shipment information, like the value and tonnage of the imported products. The time frame is long enough for firms to adjust its import behavior in response to the tariff change. Researchers may wonder if the distribution of characteristics remains constant across the different surveys and if this has any implication for the causal inference. We revisit this study and address these questions in Section 6. Our findings suggest that even though covariates vary over time for the treated group, the ATT estimates are not significantly different from those estimated under the stationarity assumption.

In this paper, we studies identification and estimation of treatment effects when repeated cross section data are available and when the composition of covariates vary over time. Our proposal builds directly on Sant’Anna and Zhao (2020), extending their results in two directions. First, we derive an estimand for the ATT that is doubly robust (DR) even when composition of the covariates changes over time. That means, the estimand will recover the true ATT if either the propensity score (PS) or the outcome regression (OR) models is correctly specified, but it’s not necessary that both are. We then show that the two estimands proposed by Sant’Anna and Zhao (2020) are no longer DR in this general setup. As a matter of fact, even when one models both nuisance functions correctly, it is generally not able to identify ATT.

Based on the identification results, we derive the semiparametric efficiency bound for the proposed

ATT estimand. The bound serves as a benchmark for gauging how an estimator performs in terms of efficiency, against the ideal case where one exhausts all the information implied by the identifying assumptions. Our efficiency bound is then compared with the one derived by Sant'Anna and Zhao (2020). This helps us to quantity the efficiency loss from not incorporating the information that covariate distribution is stationary. As such, the extra layer of robustness does not come for free. We show that the efficiency cost can be sizable with simulation exercises.

Another major difference from Sant'Anna and Zhao (2020) lies in how we estimate the ATT. Here, our proposed DR DID estimator is based on a fully nonlinear procedure for the first-step nuisance functions. In particular, we use the local polynomial estimator for the OR models and the local multinomial logit regression to estimate the PS, the latter of which is fairly new in the DID literature. As a side contribution of this paper, we provide a new result on the uniform expansion of the local logit estimators, which accommodates both continuous and discrete variables. Using the asymptotic results on the first-step estimators, we show our estimator is always consistent for the true ATT, and it attains the semiparametric efficiency bound so long as some mild conditions on the first-step estimation errors are satisfied. It means that estimating the nuisance functions does not have an effect on the ATT estimator. It is important to emphasize that our methodology can also allow for parametric working models. However, in this case, the ATT estimator may not always be consistent, and the form of its asymptotic variance is considerably more complicated.

A natural question arises from the discussion so far: how do we assess whether there exists compositional change in covariates, and more importantly, whether the variation in time leads to bias in the estimation of treatment effects? We attempt to answer this query by means of a test that is based on the insight of Hausman (1978). As we have mentioned, our proposed estimator behaves differently in two aspects from those based on Sant'Anna and Zhao (2020), namely robustness and efficiency. These discrepancies can be utilized to detect deviations from the null hypothesis of no compositional change. It is possible, however, that the proposed test has trivial power against certain alternative hypotheses. Nonetheless, we show that this should not be a concern since changes in the covariate composition does not cause bias in the treatment effects in such cases.

Related literature: This article belongs to the extensive literature on semiparametric DID methods. We refer reader to Imbens and Wooldridge (2009) for an overview, and to Roth, Sant'Anna, Bilinski and Poe (2022) for a synthesis of recent advances in the econometrics of DID. Among the references mentioned within the surveys, we are most influenced by Heckman, Ichimura and Todd (1997), Abadie (2005), and Sant'Anna and Zhao (2020). The first develops a matching-based DID regression estimator, the second proposes the DID inverse probability weighted (IPW) estimators, and the last combines the first two procedures into a DR DID estimator. All three, however, rely on the stringent assumption on the evolution of covariate process, and therefore, rules out many empirically relevant examples.

Prior to our work, there has been a few other papers dealing with time-varying covariates in the DID framework. Hong (2013) develops a matching-based estimator that is tailored to the application we mentioned before. He bases the identification of the Napster effect on a "selection-on-observable"-type assumption, which is strictly stronger than the parallel trends condition invoked in this paper. Zimmert (2019) provides several efficiency results similar to ours. He then propose estimators when the covariates are very high dimensional. The estimators use "machine learning" tools for the first-step nuisance functions. However, we note that his estimators need not always attain the low dimensional efficiency bounds derived in his paper, which is in sharp contrast to our estimator as we will show in Section 3. Different from our repeated cross section setup, Caetano et al. (2022) studies the identification of causal effects when panel is available and when the treatment can affect the covariates. They are able to identify the average treatment effect for the treated group (AToT) under additional conditions restricting the counterfactual covariate process. Instead, we abstract away from "post-treatment" control problem and focus on recovering the ATT when covariates evolve "exogenously".

The method proposed in this article is closely related to the literature of doubly robust estimation. General results that based on "unconfoundedness" or "IV/LATE" type assumptions have been obtained by Robins, Rotnitzky and Zhao (1994), Scharfstein, Rotnitzky and Robins (1999), Bang and Robins (2005), Kang and Schafer (2007), Wooldridge (2007), Cattaneo (2010), Graham, Pinto and Egel (2012), Słoczyński (2018), Rothe and Firpo (2019), and the list goes on. As a trending development, the double robustness or local robustness has been used to accommodate machine learning of high-dimensional preliminary functions. See, e.g. Belloni, Chernozhukov and Hansen (2014), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2017), Tan (2019), and Chang (2020). Our nonparametric approach to the first-step estimation complements their works.

Finally, we contribute to semiparametric two-stage estimation that depend on nonparametrically estimated functions. See e.g. Newey (1994), Chen, Linton and Van Keilegom (2003), Escanciano, Jacho-Chávez and Lewbel (2016), Mammen, Rothe and Schienle (2016), among many others. Our results on local logit regression builds on Fan, Heckman and Wand (1995), Li and Ouyang (2005), and Kong, Linton and Xia (2010). Our new result on the uniform expansion of local logit estimator may be of independent interest beyond the scope of this article.

Organization of the paper: Section 2 introduces the identification framework of the DID parameter under compositional change. In Section 3, we present our semi-parametric DR DID estimator and then discuss its large sample properties. Next, we describe a test of covariate stationarity in Section 4. Results for two Monte Carlo simulations are provided in Section 5. Then, in Section 6, we illustrate our proposed methods with an empirical application. Section 7 concludes. Proofs, as well as additional results, are reported in the Supplemental Appendix.

2 Difference-in-Differences framework

In this section, we introduce the DID model and then discuss the identifying assumptions. We will focus on the canonical two-period and two-group model, although it is interesting to extend this idea to the case of more complicated adoption patterns, as we will discuss towards the end of this article. We have two time periods, $t = 0$, where no unit is exposed to the treatment, and time $t = 1$, where units in the group with $D = 1$ are exposed to treatment; here, D is a binary group indicator. We adopt the potential outcome notation where $Y_{it}(0)$ denotes the potential outcome for unit i at time t in the absence of treatment, and $Y_{it}(1)$ denotes the potential outcome for unit i at time t in the presence of treatment. Given that nobody is treated at time $t = 0$, we have that $Y_{i0} = Y_i(0)$ for all units i, but $Y_{i1} = D_i Y_{i1}(1) + (1 - D_i) Y_{i1}(0)$. We also assume that a k-dimensional vector of pre-treatment characteristics $X_i \in \mathbb{R}^k$ is available.

As mentioned before, we only have access to repeated cross-sectional data. To formalize this idea, let T_i be a dummy variable that takes value one if the observation i is observed only in the post-treatment period $t = 1$, and zero if observation i is only observed in the pre-treatment period $t = 0$. Define $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$, and let n_1 and n_0 be the sample sizes of the post-treatment and pre-treatment periods such that $n = n_1 + n_0$.

Assumption 1 *Assume that the pooled repeated cross-section data $\{Y_i, D_i, X_i, T_i\}_{i=1}^{n}$ consists of i.i.d.*

draws from the mixture distribution

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

where $(y, d, x, t) \in \mathcal{Y} \times \{0, 1\} \times \mathcal{X} \times \{0, 1\}$.

Assumption 1 covers the case where repeated cross-section data are available, and allows for different sampling schemes. For instance, it accommodates the binomial sampling scheme where an observation i is randomly drawn from either (Y_1, D, X) or (Y_0, D, X) with fixed probability. It also accommodates the "conditional" sampling scheme where n_1 observations are sampled from (Y_1, D, X), n_0 observations are sampled from (Y_0, D, X) and $\mathbb{P}\left(T = 1\right) = n_1/n$ (here, T is treated as fixed). Importantly, Assumption 1 accommodates settings with compositional changes in the distribution of (D, X) across the two sampling periods $t = 1$ and $t = 0$ – scenarios usually ruled out by extant works in the DID literature. To restrict the time variation of covariates it is often the case that conditions a la Assumption 1(b) in Sant'Anna and Zhao (2020) are included as part of the identifying assumptions. That is,

Assumption 2 (Stationarity) $(D, X) \perp\!\!\!\perp T$.

The condition says that the joint distribution of the treatment and the covariates is time-invariant. Similar requirements can also be found in Abadie (2005), Athey and Imbens (2006), Bertrand, Duflo and Mullainathan (2004), etc. There are various ways that Assumption 2 may be violated: (1) covariates (or their distribution) do not change over time, whereas the propensity to participate in the treatment is different in two cross sections; (2) covariates vary over time but their evolution path is unaffected by D; (3) the treatment has an effect on the covariates themselves. Here, we focus mainly on the first two scenario. This type of situation arises naturally when the covariate process is determined prior to treatment. However, we do not limit ourselves to such cases. When the covariates are functions of the treatment, they act as mediators. Conditioning on such covariates can lead to a "bad control" problem that may induce bias (Zeldow and Hatfield, 2021). We refer readers to Caetano et al. (2022) for a thorough analysis on the treatment-dependent covariates.

In the above setup, we are particular interested in the average treatment effect on the treated. That is,

$$ATT = \underbrace{\mathbb{E}\left[Y_{i1}\left(1\right)|D = 1, T = 1\right]}_{\text{observed in the data}} - \underbrace{\mathbb{E}\left[Y_{i1}\left(0\right)|D = 1, T = 1\right]}_{\text{counterfactual measure}}. \tag{2.1}$$

We need to distinguish the ATT from another widely-adopted policy measure – the AToT, defined by $\mathbb{E}[Y_1(1) - Y_1(0)|D = 1]$. It is the main target parameter for Abadie (2005), Blundell, Dias, Meghir and van Reenen (2004), Sant'Anna and Zhao (2020), and Caetano et al. (2022), to name a few. If there is no compositional change, the time at which one evaluates causal effects does not matter. As such, the ATT and the AToT are equal. When covariates varies over time, however, the latter will become a weighted average of the policy effects evaluated at different times, and is generally different from the ATT. We refer to the difference between the two parameters as the "post-treatment bias" in what follows.

The potential outcome $Y_1(1)$ is observed from data, and thus, the first term in (2.2) can be identified directly. However, the second term, $\mathbb{E}\left[Y_{i1}\left(0\right)|D = 1, T = 1\right]$ is not observed in the data. The DID literature overcomes this "missing value" problem via assumptions that allow us to recover the counterfactual mean. The key assumptions are listed as follows,

Assumption 3

$$\begin{aligned}(i)\ & \mathbb{E}[Y_1(0)|D=1,T=1,X]-\mathbb{E}[Y_0(0)|D=1,T=0,X]= \\ & \quad \mathbb{E}[Y_1(0)|D=0,T=1,X]-\mathbb{E}[Y_0(0)|D=0,T=0,X]\ a.s. \\ (ii)\ & \mathbb{E}[Y_0(0)|D=1,T=0,X]=\mathbb{E}[Y_0(1)|D=1,T=0,X]\ a.s.\end{aligned}$$

Assumption 3 (i) is commonly referred to as the conditional parallel trends (CPT) assumption in the DID literature. Intuitively, it says that the average outcome for the treated and untreated groups would have experienced parallel shift if the treatment has not occurred. Assumption 3(i) is the no anticipatory effect (NAE) assumption, which implies that the average outcome prior to the treatment does not change in anticipation of such an event.

In addition to the CPT and NAE, we will also impose a strong overlapping condition.

Assumption 4 *For some* $\varepsilon > 0$, $\mathbb{P}(D=1,T=1) > \varepsilon$ *and* $\mathbb{P}(D=d,T=t|X) \geqslant \varepsilon$ *for* $(d,t) = (0,1),(1,0)$*, and* $(0,0)$ *a.s.*

Assumption 4 guarantees that there are at least some units in the post-treatment treated group. In addition, there will be some units that belong to each of the three remaining treatment groups, for any given value of the covariates. This condition is standard in the DID literature, and it is crucial for using standard inference procedures.

Remark 1 *Assumption 3 implicitly requires that there be no causal relationship between the covariates and the treatment. To allow for such possibility, modification of CPT and additional assumptions on the covariate process are necessary. When panel data are available, Caetano et al. (2022) show that causal effects can be recovered by imposing either a rank similarity or an uncounfoundedness condition on the potential covariates. It is of interest to extend their analysis to our setting, and we leave it for future research.*

Combining Equation (2.1) and Assumption 3, it is immediately clear that we can identify ATT by

$$\tau = \mathbb{E}\left[\tau(X)|\, D=1, T=1\right] \tag{2.2}$$

where $\tau(x) = \sum_{d,t\in\{0,1\}}(-1)^{d+t}m_{d,t}(x)$, and $m_{d,t}(x) = E[Y|D=d,T=t,X=x]$, for $d,t=0,1$. Intuitively, Equation (2.2) states that the ATT can be recovered by (i) taking the path of the conditional mean of potential outcome for the treated group, adjusting it by the change in the conditional mean for the control group, given values of X; (ii) and then integrate the double differences over the covariate distribution for the treated population at time $t=1$. This approach requires modeling the process of potential outcomes. Therefore, it is commonly referred to as the outcome regression (OR) approach. The estimand in (2.2) resembles the standard DID regression-adjustment proposed by Heckman et al. (1997):

$$\tau_{sc} = \mathbb{E}\left[\tau(X)|\, D=1\right]. \tag{2.3}$$

The only difference between the two objects lies in the conditioning set. We focus on the treated population at time $t=1$, whereas Heckman et al. (1997) considers the average effect on the treated across both periods. When Assumption 2 fails to hold, the conditional expectation is not invariant to T. As a result, the two estimands may differ.

Alternatively, ATT can be recovered via the IPW approach proposed by Abadie (2005). Instead of modeling the conditional mean of outcomes, we express the ATT as functions of generalized propensity scores of the form: $p(d,t,x) = \mathbb{P}(D = d, T = t|X = x)$. Consider the following Horvitz and Thompson (1952)-type estimator,

$$\tau = \mathbb{E}\left[\left(\frac{DT}{\mathbb{E}[DT]} + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} (-1)^{d+t} \frac{I_{d,t}p(1,1,x)}{p(d,t,x)\,\mathbb{E}[DT]}\right) Y\right], \tag{2.4}$$

where $I_{d,t} = 1\{D = d, T = t\}$ and $(d,t) \in \mathcal{S}_{\setminus(1,1)} = \{(1,0),(0,1),(0,0)\}$.[1] The validity of IPW approach depends crucially on whether the generalized propensity score models are correctly specified. Likewise, we compare 2.4 with the standard IPW estimator by Abadie (2005):

$$\tau = \frac{1}{\mathbb{E}[D]}\mathbb{E}\left[\left(\frac{D - p(X)}{1 - p(X)} \frac{T - \mathbb{E}[T]}{\mathbb{E}[T](1 - \mathbb{E}[T])}\right) Y\right], \tag{2.5}$$

To facilitate comparison, we rewrite (2.5) as

$$\tau = \mathbb{E}\left[\sum_{t=0,1} (-1)^{t+1} \left(\frac{D1\{T = t\}}{\mathbb{E}[D1\{T = t\}]} - \frac{p(X)(1 - D)1\{T = t\}}{(1 - p(X))\,\mathbb{E}[D1\{T = t\}]}\right) Y\right]. \tag{2.6}$$

Note that for Abadie (2005), propensity score weights applies to the control population only, and the part of weights that depends on the propensity score is the same across the time periods. On the contrary, when Assumption 2 fails to hold, all treatment groups except the treated population at time $t = 1$ is weighted by generalized propensity scores, and the PS differs for each treatment group in general.

In practice, it is possible that PS recovers the true ATT while OR does not, or vice versa. It is therefore desirable to combine the two approaches. We show how this may be achieved in the following subsection.

2.1 Doubly robust Difference-in-Differences

when a DID estimand is concerned, DR refers to the property that the estimand identifies the true ATT as long as either one of the working models, the OR and the generalized PS models, is correctly specified, but not necessarily both. Consequently, a DR estimand is less demanding in terms of its requirement on nuisance models. Beyond this favorable property for identification, estimators based on DR moment equations typically fare better in terms of efficiency.

We first focus on the robustness to identification. Some additional notations are in order before introducing the DR estimand. Let $\pi : \{0,1\}^2 \times \mathcal{X} \mapsto [0,1]$ be a model for the generalized PS model, $p(\cdot,\cdot,\cdot)$ and let $\mu_{d,t} : \mathcal{X} \mapsto \mathbb{R}$ be a model for the true OR $m_{d,t}(\cdot)$, for $(d,t) \in \mathcal{S}_{\setminus(1,1)}$. Given a generic $g : \{0,1\}^2 \times \mathcal{X} \mapsto [0,1]$, we let

$$\begin{aligned} w_{1,1}(D,T) &= \frac{DT}{\mathbb{E}[DT]}, \\ w_{d,t}(D,T,X;g) &= \frac{I_{d,t}g(1,1,X)}{g(d,t,X)} \Big/ \mathbb{E}\left[\frac{I_{d,t}g(1,1,X)}{g(d,t,X)}\right], \text{ for } (d,t) \in \mathcal{S}_{\setminus(1,1)}. \end{aligned} \tag{2.7}$$

We consider the following estimand for τ,

$$\tau_{dr}(\pi,\mu) = \mathbb{E}[w_{1,1}(D,T)Y] + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} (-1)^{(d+t)} \{\mathbb{E}[w_{d,t}(D,T,X,\pi)(Y - \mu_{d,t}(X))]$$

[1] Proof of (2.4) uses similar reasoning as that of Theorem 2.1, and thus omitted for brevity.

$$+\mathbb{E}\left[w_{1,1}(D,T)\mu_{d,t}(X)\right]\}, \quad (2.8)$$

where for $d, t \in \{0,1\}$. Note first that (2.8) involves both OR and PS models. This sort of construction shares the structure of the so-called augmented-inverse-probability-weighting-type (AIPW) estimand in the DR literature (Robins et al., 1994).

This estimand is similar in structure to those proposed by Sant'Anna and Zhao (2020) for ATT when repeated cross section data are available:

$$\begin{aligned}
\tau_{sc,1}(\pi,\mu) &\equiv \mathbb{E}\left[\sum_{d,t=0,1}(-1)^{d+t}w_{d,t}^{sc}(D,T,X;\pi)(Y-(T\mu_{0,1}(X)+(1-T)\mu_{0,0}(X)))\right],\\
\tau_{sc,2}(\pi,\mu) &\equiv \tau_{sc,1}+\sum_{d=0,1}(-1)^{d+1}\left\{\mathbb{E}[\mu_{d,1}(X)|D=1]-\mathbb{E}[\mu_{d,1}(X)|D=1,T=1]\right\}\\
&\quad+\sum_{d=0,1}(-1)^{d}\left\{\mathbb{E}[\mu_{d,0}(X)|D=1]-\mathbb{E}[\mu_{d,0}(X)|D=1,T=0]\right\}. \qquad (2.9)
\end{aligned}$$

where

$$\begin{aligned}
w_{d,t}^{sc}(D,T,X;g) = 1\{d=1\}&\frac{D1\{T=t\}}{\mathbb{E}[D1\{T=t\}]}\\
&+1\{d=0\}\frac{g(X)(1-D)1\{T=t\}}{1-g(X)}\Big/\mathbb{E}\left[\frac{g(X)(1-D)1\{T=t\}}{1-g(X)}\right].
\end{aligned}$$

Sant'Anna and Zhao (2020) show that both estimand are DR, and furthermore, $\tau_{sc,2}$ attains the semiparametric efficiency bound, when covariate process is stationary. We investigate whether these properties are preserved when Assumption 2 fails to hold in the following two subsections.

Another feature worth mentioning is that the weights in (2.7) are of the Hájek (1971)-type. This guarantees that all the weights sum up to one. We prefer the stabilized weights as the resulting treatment effect estimators tend to have better finite sample properties; see e.g. Millimet and Tchernis (2009); Busso, Dinardo and McCrary (2014) for further details.

Theorem 2.1 *(Double Robustness) Suppose Assumptions 1, 3, and 4 hold. Then $\tau_{dr}(\pi,\mu)=\tau$ if either (but not necessarily both) $\pi(\cdot,\cdot,X)=p(\cdot,\cdot,X)$ a.s., or (but not necessarily both) $\mu_{d,t}(X)=m_{d,t}(X)$ a.s., for $(d,t)\in\mathcal{S}_{\backslash(1,1)}$.*

Theorem 2.1 shows that τ_{dr} is DR, meaning that it can recover the true ATT if at least one of the models is correctly specified. Relative to the requirements on the working models in Theorem 1(b) of Sant'Anna and Zhao (2020), those imposed by Theorem 2.1 are stronger when OR is misspecified. Our estimand requires that propensity scores be correctly specified for each of the four treatment groups, whereas Sant'Anna and Zhao (2020) only impose the constraint on the treated population. On the other hand, if PS is misspecified, τ can be identified only if each of the three conditional mean, $\{\mu_{d,t}(\cdot)\}_{(d,t)\in\mathcal{S}_{\backslash(1,1)}}$ is correctly specified. Instead, Sant'Anna and Zhao (2020) requires $\mu_{1,t}-\mu_{0,t}=m_{1,t}-m_{0,t}$ in such cases. This discrepancy stems from the breakdown of $\mathbb{E}[\cdot|D=1,T=1]=\mathbb{E}[\cdot|D=1]$, when we relax Assumption 2.

2.2 Semiparametric efficiency bound

We have shown how the DR estimand provide an extra layer of protection when identification is considered. In this subsection, we turn to the second motivation for using DR estimands, namely its efficiency property.

We derive the semiparametric efficiency bounds for τ. These bounds are the semiparametric version of the Cramér-Rao lower bound in parametric settings. Researcher can use them as benchmarks to measure how precise a given semiparametric DID estimator is.

Theorem 2.2 *(Semiparametric Efficiency Bound) (a) Suppose Assumptions 1, 3, and 4 hold. Then, the efficient influence function for the ATT, τ, is given by*

$$\eta_{dr}(W;p,m) = w_{1,1}(D,T)(\tau(Y,X)-\tau) + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} (-1)^{(d+t)} w_{d,t}(D,T,X;p)(Y-m_{d,t}(X)),$$

where $\tau(Y,X) = Y + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}}(-1)^{d+t}m_{d,t}(X)$. Furthermore, the semiparametric efficiency bound for the set of regular estimator of τ is

$$\mathbb{E}[\eta_{dr}(W;p,m)^2] = \frac{1}{p(1,1)^2}\,\mathbb{E}\left[DT(\tau(Y,X)-\tau)^2 + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} \frac{I_{d,t}p(1,1,X)^2}{p(d,t,X)^2}(Y-m_{d,t}(X))^2\right].$$

(b) Suppose that Assumptions 1-4 hold. Then, the efficient influence function for the ATT, τ, is given by

$$\eta_{sc}(W;p,m) \equiv \frac{D}{\mathbb{E}[D]}(\tau(X)-\tau) + \sum_{d,t\in\{0,1\}}(-1)^{(d+t)}w^{sc}_{d,t}(D,T,X;p)(Y-m_{d,t}(X)).$$

Moreover, the semiparametric efficiency bound for the set of regular estimator of τ is

$$\begin{aligned}\mathbb{E}[\eta_{sc}(W;p,m)^2] =& \frac{1}{\mathbb{E}[D]^2}\,\mathbb{E}\left[D(\tau(X)-\tau)^2\right]\\ &+\frac{1}{\mathbb{E}[D]^2}\,\mathbb{E}\Bigg[\frac{DT}{\mathbb{E}[T]^2}(Y-m_{1,1}(X))^2 + \frac{D(1-T)}{(1-\mathbb{E}[T])^2}(Y-m_{1,0}(X))^2\\ &+\frac{(1-D)Tp(X)^2}{(1-p(X))^2\,\mathbb{E}[T]^2}(Y-m_{0,1}(X))^2 + \frac{(1-D)(1-T)p(X)^2}{(1-p(X))^2(1-\mathbb{E}[T])^2}(Y-m_{0,0}(X))^2\Bigg].\end{aligned}$$

This theorem provides the efficient influence functions[2] and the semiparametric efficiency bounds for the ATT when Assumption 2 is violated and when it holds. Close inspection reveals that the bound under Assumption 2 is equal to that of Sant'Anna and Zhao (2020). This should not be unexpected since τ and τ_{sc} are equal when covariate distribution is invariant, and therefore, any estimator for the ATT that fully exploits the empirical content of the maintained assumptions should utilize the same amount of information as that for the τ_{sc}.

Next, we quantify the difference between the two bounds in Theorem 2.2, by means of which, we show how much efficiency is lost when stationarity condition holds but we use an estimator that ignores this information.

Corollary 1 *(Efficiency Loss under Stationarity) Suppose that Assumptions 1-4 hold. Then*

$$\rho_{sc} \equiv \mathbb{E}[\eta_{dr}(W;p,m)^2] - \mathbb{E}[\eta_{sc}(W;p,m)^2] = \frac{1-\mathbb{E}[T]}{\mathbb{E}[D]\,\mathbb{E}[T]}\,\mathbb{E}[(\tau(X)-\tau)^2|D=1], \tag{2.10}$$

It is evident from the corollary that the term involving $\tau(X)-\tau$ is the source of the efficiency loss. Under stationarity, the treated group in both pre- and post-treatment periods contributes to this term;

[2] Precise definition of the efficient influence function and other related technical details such as the path-wise differentiability and tangent space, can be found in Chapter 3 of Bickel, Klaassen, Ritov and Wellner (1998).

however, when the stationarity is violated, only the treated group from the post-treatment period affects this term, which inflates the variance.

Three factors determine the magnitude of ρ_{sc}. The efficiency loss is larger if either one of the following three quantities is larger: (a) the population ratio of the pre-treatment period vs the post-treatment period, (b) the population proportion of the comparison group vs the treated group, and (c) the expected variability of treatment effect heterogeneity on the treated. In the extreme case where the treatment effect on the treated is homogeneous, our ATT estimator would achieve the same level of efficiency as the one that imposes stationarity *a priori*.

2.3 Bias decomposition

When the covariate distribution is stationary, Sant'Anna and Zhao (2020) show that both $\tau_{sc,1}$ and $\tau_{sc,2}$ enjoy the DR property. One may wonder, however, whether any of the two estimands can still recover the unknown ATT if Assumption 2 fails to hold. We address this question in the following theorem.

Theorem 2.3 *(Bias Decomposition) Suppose Assumptions 1, 3, and 4 hold. Then*

$$
\begin{aligned}
bias_{sc,1}(\pi,\mu) &\equiv \tau_{sc,1}(\pi,\mu) - \tau \\
&= -\sum_{t=0,1}(-1)^t\,\mathbb{E}\left[\left(\frac{\pi(1,X)p(0,t,X)}{\pi(0,X)\,\mathbb{E}[\pi(1,X)p(0,t,X)/\pi(0,X)]} - \frac{p(1,1,X)}{p(1,1)}\right)\Delta_{0,t}(X)\right] \\
&+ \mathbb{E}\left[\left(\frac{p(1,0,X)}{p(1,0)} - \frac{p(1,1,X)}{p(1,1)}\right)\Delta_{1,0}(X)\right] \\
&+ \mathbb{E}[\Delta_{1,0}(X) - \Delta_{0,0}(X)|D=1,T=1] - \mathbb{E}[\Delta_{1,0}(X) - \Delta_{0,0}(X)|D=1,T=0] \\
&+ \mathbb{E}[m_{1,0}(X) - m_{0,0}(X)|D=1,T=1] - \mathbb{E}[m_{1,0}(X) - m_{0,0}(X)|D=1,T=0], \\
bias_{sc,2}(\pi,\mu) &\equiv \tau_{sc,2}(\pi,\mu) - \tau \\
&= -\sum_{t=0,1}(-1)^t\,\mathbb{E}\left[\left(\frac{\pi(1,X)p(0,t,X)}{\pi(0,X)\,\mathbb{E}[\pi(1,X)p(0,t,X)/\pi(0,X)]} - \frac{p(1,1,X)}{p(1,1)}\right)\Delta_{0,t}(X)\right] \\
&+ \mathbb{E}\left[\left(\frac{p(1,0,X)}{p(1,0)} - \frac{p(1,1,X)}{p(1,1)}\right)\Delta_{1,0}(X)\right] \\
&+ \sum_{d,t=0,1}(-1)^{d+t}\left\{\mathbb{E}[\Delta_{d,t}(X)|D=1] - \mathbb{E}[\Delta_{d,t}(X)|D=1,T=1]\right\}, \\
&+ \mathbb{E}[\tau(X)|D=1] - \mathbb{E}[\tau(X)|D=1,T=1],
\end{aligned}
$$

where $\pi(d,x)$ is the candidate function for $p(d,x) = \mathbb{P}(D=d|X=x)$, $\tau_\mu(x) = \sum_{d,t\in\{0,1\}}(-1)^{d+t}\mu_{d,t}(x)$, and $\Delta_{d,t}(x) = \mu_{d,t}(x) - m_{d,t}(x)$. If, in addition, the outcome regression model is correctly specified, then

$$
\begin{aligned}
bias_{sc,1}(\pi,m) &= \mathbb{E}[m_{1,0}(X) - m_{0,0}(X)|D=1,T=1] - \mathbb{E}[m_{1,0}(X) - m_{0,0}(X)|D=1,T=0] \\
bias_{sc,2}(\pi,m) &= \mathbb{E}[\tau(X)|D=1] - \mathbb{E}[\tau(X)|D=1,T=1].
\end{aligned}
$$

Theorem 2.3 provides bias decomposition for $\tau_{sc,1}$ and $\tau_{sc,2}$, respectively, when the joint distribution of X and D varies over time. Here, we first consider the case where the OR functions are correctly specified, in which case, the first three lines in each decomposition vanish. Hence, the bias is always the post-treatment bias, which is invariant to the specification of the propensity score model. Next, if OR model is subject to misspecification, on top of the its effect is captured by the first term in each bias expression. Summing up, the DR property of $\tau_{sc,j}$, $j = 1, 2$, breaks down when the covariate process is non-stationary.

Remark 2 *There is no ordering between $bias_{sc,1}$ and $bias_{sc,2}$ in general. If we assume in addition that*

$$\mathbb{E}[m_{1,t}(X) - m_{0,t}(X)|D = 1, T = 0] \leqslant \mathbb{E}[m_{1,t}(X) - m_{0,t}(X)|D = 1, T = 1],$$

for $t = 0, 1$, then $bias_{sc,2}(\pi, m) \leqslant bias_{sc,1}(\pi, m)$.

3 Estimation and inference

By definition, DR estimators are only consistent if at least one of the working nuisance functions is correctly specified. Misspecification of both models can lead to substantial bias. On top of that, the plenitude of covariates from which empirical researchers may choose from render functional form choices even more taxing. Semiparametric or fully nonparametric approaches relax restrictive functional form assumptions and alleviate concerns about misspecification error. Two-step estimators with nonparametric first step has been studied extensively in the treatment effect literature. See, e.g. Cattaneo (2010), Lee (2018), and Rothe and Firpo (2019). Sieve and kernel-type[3] regressions are the most popular choices, and among the latter class, local polynomial smoothing is arguably the most popular due to its adaptive control of boundary bias (Fan and Gijbels, 1996).

In the following subsection, we first introduce the local polynomial estimators for the PS and OR functions, respectively. Next, we proceed to describe an estimator for the ATT and then establish its asymptotic behavior. Lastly, we provide a data-driven bandwidth selection method in Subsection 3.3.

3.1 Local polynomial estimation of nuisance functions

We first focus on estimator for the PS functions. Conditional probability functions are bounded in the unit interval. When linear probability models are employed, such bounds may not be respected. Local logit regression, being a nonparametric generalization of the parametric logit regression, enforces such bounds by design. Through extensive Monte Carlo simulations, Frölich (2006) shows that local logit estimator consistently outperform local least squares, Klein–Spady and Nadaraya–Watson regressions, and is often more precise than parametric logit in finite samples. We therefore favor this estimator over other nonparametric methods.

Suppose there exist functions, $\{g_{d,t}(\cdot)\}_{(d,t)\in\mathcal{S}_{\backslash(1,1)}}$, such that

$$p(d, t, x) = \frac{\exp(g_{d,t}(x))}{1 + \sum_{(d',t')\in\mathcal{S}_{\backslash(1,1)}} \exp(g_{d',t'}(x))},$$

for $(d, t) \in \mathcal{S}_{\backslash(1,1)}$, and $p(1, 1, x) = \left(1 + \sum_{(d',t')\in\mathcal{S}_{\backslash(1,1)}} \exp(g_{d',t'}(x))\right)^{-1}$. That is, we assume the generalized PS can be represented by a multinomial transformation of unknown functions $\{g_{d,t}(\cdot)\}_{d,t\in\{0,1\}}$. Instead of imposing specific functional forms, the local logit estimator proceed to approximate the unknown functions locally by polynomials.

[3] Kernel-type estimators include the Nadaraya–Watson estimator, local polynomial method, local likelihood regression, etc. To distinguish likelihood-based from least squares-based local polynomial methods, we refer to the latter as the local least squares regression.

Towards this end, we first differentiate between continuous and discrete covariates. We assume that $x = (x_c, x_d)$, where x_c is a υ_c-vector of continuous covariates, and x_d is the subvector of discrete variables. In addition, we distinguish between ordered and unordered discrete variables. That is, $x_d = (x_u, x_o)$, where x_u is a υ_u-vector of unordered covariates and x_o is a υ_o-vector of ordered covariates. Following the tradition in local polynomial estimation, we use the notations

$$\mathbf{k} = (k_1, ..., k_v), \quad |\mathbf{k}| = \sum_{\ell=1}^{v} k_\ell, \quad \mathbf{k}! = \prod_{\ell=1}^{v} k_\ell !, \quad x^{\mathbf{k}} = \prod_{\ell=1}^{v} x_\ell^{k_\ell},$$

$$\sum_{0 \leqslant |\mathbf{k}| \leqslant p} f(\mathbf{k}) = \sum_{\ell=0}^{p} \sum_{k_1=0}^{\ell} \cdots \sum_{\substack{k_v=0 \\ k_1+\ldots+k_v=\ell}}^{\ell} f(k_1, ..., k_v),$$

to shorthand for common operators on vectors. Define $n_k = \begin{pmatrix} k+\ell-1 \\ \ell-1 \end{pmatrix}$ as the number of distinct ℓ-tuples $\mathbf{k}$ with $|\mathbf{k}| = k$. We arrange these n_k ℓ-tuples as a sequence in a lexicographic order, where the highest priority is given to the last position. Let $\pi_k(\cdot)$ denote the mapping from the rank in the sequence to the corresponding ℓ-tuple. For each $k = 0, ..., p$, we arrange $(\mathbf{x} - \mathbf{x_0})^{\mathbf{k}}$ and $g_s^{(\mathbf{k})}(\mathbf{x_0})$ according to the above order. Then, for a generic function, $g : \mathcal{X} \mapsto \mathbb{R}$, and a point, $x^* \in \mathcal{X}$, $g(\cdot)$ is approximated in a neighborhood of x^* by

$$g(x) \approx \sum_{1 \leqslant |\mathbf{k}| \leqslant p} \frac{1}{\mathbf{k}!} g^{(\mathbf{k})}(x^*)(x_c - x_c^*)^{\mathbf{k}} \equiv \underline{\mathbf{X}}(x_c^*)' \theta_g(x^*),$$

where $\underline{\mathbf{X}}_p(x_c) = (\underline{\mathbf{X}}^{(0)\prime}(x_c), ..., \underline{\mathbf{X}}^{(p)\prime}(x_c))'$ is a $N_p(= \sum_{k=0}^{p} n_k) \times 1$ vector collecting the sorted $(X_c - x_c)^{\mathbf{k}}$, with the l'th entry of $\underline{\mathbf{X}}^{(k)}(x_c)$, $\underline{\mathbf{X}}^{(k,l)}(x_c)$, equal to $(X_c - x_c)^{\pi_k(l)}$. Likewise, $\theta_g(x) = (\theta_{g,0}(x), \ldots, \theta_{g,p}(x))$ is defined as the vector of lexicographically-ordered $\frac{1}{\mathbf{k}!} g_s^{(\mathbf{k})}(x)$.

Local approximation is achieved through kernel smoothing. For continuous variables, we let the kernel function be denoted by $K^j(\mathbf{u})$, $j = ps, or$. It is a nonnegative, symmetric function supported on $[-1, 1]^{\otimes \upsilon_c}$. Suppose $h > 0$ is a generic bandwidth parameter. We denote the scaled kernel function by $K_h(\mathbf{u}) = K(\mathbf{u}/h)/h^d$. To accommodate discrete variables, we employ the kernel function proposed by Li and Racine (2007). For a generic smoothing parameter $\lambda = (\lambda_u, \lambda_o) \in [0, 1]^{\otimes \upsilon_d}$, we let

$$L_\lambda(x_d, z_d) = \prod_{s=1}^{\upsilon_u} \lambda_u^{1\{x_{u,s} - z_{u,s}\}} \prod_{s=1}^{\upsilon_o} \lambda_o^{|x_{o,s} - z_{o,s}|}. \tag{3.1}$$

When $\lambda = 0$, the estimator reduces to that based on the frequency estimator. Let $e_{\ell,k}$ denote a vector with its k'th element equal to one and all other elements being zero. For an observed X_j and $(d, t) \in \mathcal{S}_{\backslash(1,1)}$, the local logit polynomial estimator solves

$$\hat{\gamma}(X_j) = (\hat{\gamma}_{1,0}'(X_j), \hat{\gamma}_{0,1}'(X_j), \hat{\gamma}_{0,0}'(X_j))' \equiv \arg\max_{\gamma} \frac{1}{n-1} \sum_{i \neq j}^{n} \ell(W_i, X_j; \gamma) \widetilde{K}_{ps}(X_i; X_j, h, \lambda), \tag{3.2}$$

where $\widetilde{K}_{ps}(X_i; X_j, h, \lambda) = K_h^{ps}(\underline{\mathbf{X}}_{p,i}^{(1)}(X_{c,j})) L_\lambda(X_d, X_{d,j})$ and $\ell(w, x; \gamma)$ is the local logistic likelihood,

$$\ell(w, x; \gamma) = \sum_{(d',t')\in\mathcal{S}_{\backslash(1,1)}} I_{d,t}\underline{\mathbf{X}}_p(x_c)'\gamma_{d,t} - \log\left(1 + \sum_{(d',t')\in\mathcal{S}_{\backslash(1,1)}} \exp\left(\underline{\mathbf{X}}_p(x_c)'\gamma_{d',t'}\right)\right).$$

Note that we have used a "leave-one-out" version of the local regression estimator to construct $\hat{\gamma}$. That is, $\gamma(X_j)$ are estimated using every observations except j. This technique, standard in the literature (Powell and Stoker, 1996; Powell, Stock and Stoker, 1989; Rothe and Firpo, 2019), serves to avoid a "leave-in" bias that is of first-order importance. Given $\hat{\gamma}$, the generalized PS can thus be approximated as

$$\hat{p}(d, t, x) = \frac{\exp(e'_{N_p,1}\hat{\gamma}_{d,t}(x))}{1 + \sum_{(d',t')\in\mathcal{S}_{\backslash(1,1)}} \exp(e'_{N_p,1}\hat{\gamma}_{d',t'}(x))}, \tag{3.3}$$

for $(d, t) \in \mathcal{S}_{\backslash(1,1)}$, and $\hat{p}(1, 1, x) = 1 - \sum_{(d,t)\in\mathcal{S}_{\backslash(1,1)}} \hat{p}(d, t, x)$.

Now, we turn our attention to the OR models, for which, we use leave-one-out q-th order local polynomial least squares estimators. We first get the local polynomial coefficients by

$$\hat{\beta}_{d,t}(X_j) = \arg\min_{\beta} \frac{1}{n-1}\sum_{i\neq j}\left(Y_i - \underline{\mathbf{X}}_{q,i}(X_{c,j})'\beta\right)^2 I_{d,t,i}\widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}), \tag{3.4}$$

where $\widetilde{K}_{or}(X_i; X_j, b_{d,t}, \vartheta_{d,t}) = K^{or}_{b_{d,t}}(\underline{\mathbf{X}}^{(1)}_{q,i}(X_{c,j}))L_{\vartheta_{d,t}}(X_d, X_{d,j})$. Then, we estimate the OR functions by

$$\hat{m}_{d,t}(X_j) = e'_{N_q,1}\hat{\beta}_{d,t}(X_j) \tag{3.5}$$

for $(d, t) \in \mathcal{S}_{\backslash(1,1)}$.

We analyze asymptotic behaviors of these local polynomial estimators in Appendix B. We provide results on the uniform convergence rate for the approximation error. In particular, we establish a new stochastic expansion for the local multinomial logit regression that is similar in structure to the one for the local least squares regression.

Remark 3 *The choice of polynomial order depends on factors such as computational tractability and the trade-off between bias and variance properties. We follow the recommendation by Fan et al. (1995) and use odd-degree polynomial fits as they simplify the analysis for the boundary bias. We allow the orders of local polynomials to differ for estimators of the PS and the OR, and in the latter case, for different treatment groups. This is desirable as the propensity score and conditional mean functions may exhibit different level of smoothness.*

3.2 ATT estimator

Given the first-step estimators described in the last subsection, we estimate the ATT by

$$\hat{\tau}_{dr} = \sum_{d,t\in\{0,1\}} (-1)^{d+t}\,\mathbb{E}_n\left[\psi_{d,t}(W, \hat{w}, \hat{m})\right], \tag{3.6}$$

where

$$\psi_{d,t}(W, \hat{w}, \hat{m}) = \begin{cases} \hat{w}_{1,1}(D, T)Y, & \text{if } d = t = 1, \\ \hat{w}_{d,t}(D, T, X, \hat{p})(Y - \hat{m}_{d,t}(X)) + \hat{w}_{1,1}(D, T)\hat{m}_{d,t}(X), & \text{otherwise,} \end{cases}$$

and

$$\hat{w}_{1,1}(D,T) = \frac{DT}{\mathbb{E}_n[DT]},$$
$$\hat{w}_{d,t}(D,T,X,\hat{p}) = \frac{I_{d,t}\hat{p}(1,1,X)}{\hat{p}(d,t,X)} / \mathbb{E}_n\left[\frac{I_{d,t}\hat{p}(1,1,X)}{\hat{p}(d,t,X)}\right].$$

The estimator is the empirical analogue of (2.8), with unknown nuisance functions replaced by (3.3) and (3.5). Before deriving the large sample properties of $\hat{\tau}_{dr}$, we make the following set of assumptions.

Assumption 5 *1. (i) $\mathcal{X} = \mathcal{X}_c \otimes \mathcal{X}_d$, where $\mathcal{X}_c$ is a compact subset of $\mathbb{R}^{\upsilon_c}$ and $\mathcal{X}_d$ is finite; (ii) The marginal probability density of X_c, $f_{X_c}(\cdot)$, is continuously differentiable and bounded away from zero on $\mathcal{X}_c$;*

2. *For all $x \in \mathcal{X}$, and $d, t \in \{0,1\}$, $p(d,t,x)$ is $(p+1)$-times continuously differentiable in x_c, with uniformly bounded derivatives; (ii) $m_{d,t}(x)$ is $(q+1)$-times continuously differentiable in x_c, with uniformly bounded derivatives;*

3. *$\mathbb{E}[|Y|^{\zeta}|X,D,T] < \infty$ a.s. for some constant $\zeta > 2$.*

4. *For $j = ps, or$, (i) $K^j : [-1,1]^{\otimes \upsilon_c} \to \mathbb{R}_+$; (ii) $K^j(\cdot)$ is symmetric and twice continuously differentiable; (iii) $\|\mathbf{u}\|^{4p_j} K^j(\mathbf{u}) \in L_1(\mathcal{U}^{\otimes \upsilon_c})$; (iii) With $\mathbf{Q}_p(x_c)$ defined in (B.29), $\inf_{x_c \in \mathcal{X}_c}\{\lambda_{min}\mathbf{Q}_p(x_c)\}$ is strictly positive.*

5. *(i) $h = o(1)$; (ii) $\log n/\left(nh^{\upsilon_c+2p}\right) = o(1)$ and $\lambda/h^p = o(1)$; (iii) $h^{p+1} = o\left(n^{-1/4}\right)$ and $\log n/\left(nh^{\upsilon_c}\right) = o\left(n^{-1/2}\right)$. For $(d,t) \in \mathcal{S}_{\setminus(1,1)}$, (iv) $b_{d,t} = o(1)$; (v) $(\log n\{(\log n)(\log\log n)^{1+\delta}\})^{2/\zeta}/n^{1-2/\zeta}b_{d,t}^{\upsilon_c} \to 0$, for some $0 < \delta < 1$; (vi) $b_{d,t}^{q+1} = o\left(n^{-1/4}\right)$ and $\log n/\left(nb_{d,t}^{\upsilon_c}\right) = o\left(n^{-1/2}\right)$; (vii) $\lambda, \vartheta_{d,t} = o(n^{-1/4})$.*

A few remarks of the assumptions are in order. Assumption 5.1 indicates that our local polynomial estimator can accommodate categorical data. The key is to ensure that the uniform rate conditions in Lemma B.2 hold in the presence of discrete variables. See, e.g. Li and Ouyang (2005) for such results in the case of a kernel-type estimator. Assumption 5.2 describes the standard smoothness condition for the nuisance functions. Assumption 5.3 is a regularity condition that controls the conditional moments of Y. Assumption 5.4 collects the regularity conditions on the kernel functions. Note that we allow the kernel to be different for the propensity score and the conditional mean models. In practice, we may use the product kernel, i.e. $K_h(\mathbf{u}) = \prod_{i=1}^{\upsilon_c}\tilde{K}(u_i/h)/h$, where $\tilde{K}(\cdot)$ can be a commonly used univariate kernels, such as uniform, triangular, biweight, triweight, and Epanechnikov kernels to list a few. The Gaussian kernel, however, is ruled out by this restriction. Assumption 5 collects the rate condition on the bandwidths. Assumptions 5.5 (ii, v) are imposed to ensure linear expansions of the local polynomial estimators hold uniformly over $\mathcal{X}$. 5.5 (v) is due to Masry (1996). When Y has finite moments of any order, such as when it is a discrete variable, this condition is implied by Assumptions 5.5 (vi). Assumptions 5.5 (iii, vi, vii) specifies rate conditions on the bias and stochastic part of the first step estimation error. The usual $o_p(n^{-1/4})$ rate of convergence for the error applies here. Without the DR property, it typically requires more stringent rate conditions on the bias part, which can only be fulfilled with higher-order kernel functions; see, e.g. Newey (1994); Lee (2018).

Remark 4 *The result of Rothe and Firpo (2019) can be applied to weaken the rate conditions on the nuisance functions. They provide higher order expansion of semiparametric two-step DR estimators, through which, they show that, as long as the bias and the stochastic parts of the first-step error are* *of order $o_p(n^{-1/6})$, respectively, and their product is of order $o_p(n^{-1/2})$, the resulting DR estimator is root-n consistent. A full treatment along this line is left for future research.*

Theorem 3.1 *(Asymptotic Normality Doubly Robust Estimator) Under Assumptions 1, 3, 4, and 5, we have*

$$\sqrt{n}(\hat{\tau}_{dr} - \tau) = \frac{1}{n}\sum_{i=1}^{n} \eta_{dr}(W_i) + o_p(1) \xrightarrow{d} \mathcal{N}\left(0,\, \Omega_{dr}\right), \tag{3.7}$$

where $\Omega_{dr} = \mathbb{E}[\eta_{dr}(W)^2]$.

Theorem 3.1 states that $\hat{\tau}_{dr}$ is root-n consistent, and asymptotically normal. It also shows that the estimation error of the nuisance functions does not affect the asymptotic distribution of $\hat{\tau}_{dr}$. Moreover, the asymptotic variance of $\hat{\tau}_{dr}$ is equal to the semiparametric efficiency bound

The theorem can be used to calculate confidence intervals for the ATT. Towards this end, we need an estimator of the asymptotic variance, Ω_{dr}. One can construct such an estimator using either empirical analogues of the influence function or by bootstrapping. We focus on the first approach here, and a weighted bootstrap procedure is provided in Appendix C that accommodates clustered inference.

$$\hat{\eta}_{dr}(W;\hat{p},\hat{m}) = \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} (-1)^{d+t}\hat{w}_{d,t}(D,T,X,\hat{p})(Y-\hat{m}_{d,t}(X)) + \hat{w}_{1,1}(D,T,X,\hat{p})(\hat{\tau}(Y,X)-\hat{\tau}_{dr}), \tag{3.8}$$

and $\hat{\Omega}_{dr} = \mathbb{E}_n[\hat{\eta}_{dr}(W;\hat{p},\hat{m})^2]$. Under mild regularity conditions, consistency of $\hat{\Omega}_{dr}$ follows, the proof of which is contained in that of Theorem 4.1.

3.3 Bandwidth selection

In this subsection we consider practical bandwidth determination for the first-step local polynomial estimators. It is well-known that smoothing parameters plays a crucial role in the trade-off between reducing bias and variance. Although the robustness check with multiple bandwidths is useful but a reliable data-driven rule for selection is oftentimes preferred. The multitude of bandwidth selection methods in the extant literature mainly falls into two categories, namely the plug-in methods and the cross-validation based approaches.

The former involves plugging in estimates of unknown quantities to criteria that determine the asymptotically optimal bandwidth. It demonstrates good finite-sample properties for low dimensional regressors (Fan and Gijbels, 1996). On contrary, cross-validation methods, which target minimizing the out-of-sample prediction error, tend to fare better when the covariates have moderate dimension and/or contain discrete variables (Hall, Racine and Li, 2004). On top of it, bandwidth selected by cross validation methods has a better chance of being consistent than the plug-in methods, when the true nuisance models belong to the parametric world (Frölich, 2006; Eguchi, Yoon Kim and Park, 2003). Due to these reasons, we focus on the cross validation method.

Let

$$C_n^{ls}(h,\lambda,\{b_{d,t},\vartheta_{d,t}\}_{(d,t)\in\mathcal{S}_{\setminus(1,1)}}) \\ = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{d,t\in\{0,1\}}(I_{d,t,i}-\hat{p}(d,t,X_i))^2 + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} I_{d,t,i}(Y_i-\hat{m}_{d,t}(X_i))^2\right\}, \tag{3.9}$$

$$C_n^{ml}(h, \lambda, \{b_{d,t}, \vartheta_{d,t}\}_{(d,t)\in\mathcal{S}_{\setminus(1,1)}})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{-\sum_{d,t\in\{0,1\}} I_{d,t,i}\log(\hat{p}(d,t,X_i)) + \sum_{(d,t)\in\mathcal{S}_{\setminus(1,1)}} I_{d,t,i}(Y_i - \hat{m}_{d,t}(X_i))^2\right\}. \tag{3.10}$$

The least-squares criterion, C_n^{ls}, which is based on the sum of the least squares distance between the observed and leave-one-out fitted values for both PS and OR estimators, is standard in the kernel estimation literature. The second criterion, C_n^{ml}, replaces the least squares sum for the PS estimator with that of observed likelihood. This idea of using a likelihood based criterion in the context of local logistic estimation can be traced back to Staniswalis (1989).

The cross-validated bandwidths, $(\hat{h}^j, \hat{\lambda}^j, \{\hat{b}_{d,t}, \hat{\vartheta}_{d,t}\}_{d,t\in\{0,1\}})$, minimizes, C_n^j for $j = ls, ml$. In Appendix B.3, we study the mean integrated squared error (MISE) properties of the first-step estimators and derive the convergence rates of the optimal bandwidths. When local linear estimation is considered (i.e. $p = q = 1$), optimal bandwidths ensure that the rate conditions in Assumption 5.5 are satisfied if $\upsilon_c < 4$. To the contrary, the result does not impose any limit on the number of discrete variable.

Remark 5 *When coupled with the local logit estimation, cross-validation can be computationally intensive. This is partly due to the fact that, unlike local least squares estimator, local logit regression does not admit a close-form solution. For each evaluation of the criterion function, it entails solving n minimization problem. This can be time-consuming when data sets are somewhat large. To tackle this problem, we provide a plug-in method for frequency-based local polynomial estimators – Algorithm 1 in Appendix B.4. The algorithm exploits analytical expressions for the MISE, and therefore, avoids the computation burden of the cross-validation method. We recommend using the procedure when υ_d is small and the size of data is large.*

4 Test of covariate stationarity

Two estimators have featured prominently into our discussion so far. The one proposed in this paper, $\hat{\tau}_{dr}$, is based on DR estimand (2.8). The other, developed by Sant'Anna and Zhao (2020), is based on (2.9). One may wonder how do we choose between these two in practice. To address such a query, we propose a test in this section, which is based on the insight of Durbin (1954), Wu (1973, 1974), and Hausman (1978).

A typical Hausman test requires the existence of two estimators, A and B, of the same parameter, such that, under the null hypothesis, both A and B are consistent, but only A is efficient, while under the alternative, only B is consistent. Specializing to our case, A can be any ATT estimator that is consistent and achieves the semiparametric efficiency bound when the covariate process is stationarity, and B can be any estimator that is consistent irrespective of the stationarity condition. Our proposed estimator $\hat{\tau}_{dr}$ is a good candidate for the latter. And for A, we modify the estimator of Sant'Anna and Zhao (2020), by having the nuisance functions estimated nonparametrically.

Towards this end, we need nonparametric estimators for time-invariant PS, $p(\cdot)$ and OR functions, $\{m_{d,t}(\cdot)\}_{d,t\in\{0,1\}}$. For the former, we may reuse the local polynomial estimators from Section 3.1 by letting

$$\hat{p}(X) = \hat{p}(1,1,X) + \hat{p}(1,0,X),$$

where $\hat{p}(1,t,X)$ is given by (3.3). As opposed to (2.8), all four conditional mean functions,

$\{m_{d,t}(\cdot)\}_{d,t\in\{0,1\}}$ feature into (2.9). We can estimate them in a similar fashion as (3.5). With these quantities in hand, we then define the modified Sant'Anna and Zhao (2020) estimator by

$$\hat{\tau}_{sc} \equiv \mathbb{E}_n\left[\hat{w}_1^{sc}(D)\hat{\tau}(X) + \sum_{d,t\in\{0,1\}}(-1)^{(d+t)}\hat{w}_{d,t}^{sc}(D,T,X;\hat{p})(Y-\hat{m}_{d,t}(X))\right], \tag{4.1}$$

where $\hat{\tau}(X) = \sum_{d,t\in\{0,1\}}(-1)^{d+t}\hat{m}_{d,t}(X)$, and

$$\begin{aligned}\hat{w}_1^{sc}(D) &= \frac{D}{\mathbb{E}_n[D]},\\ \hat{w}_{d,t}^{sc}(D,T,X;\hat{p}) &= 1\{d=1\}\frac{I_{d,t}}{\mathbb{E}_n[I_{d,t}]} + 1\{d=0\}\frac{I_{d,t}\hat{p}(X)}{1-\hat{p}(X)}\Big/\mathbb{E}_n\left[\frac{I_{d,t}\hat{p}(X)}{1-\hat{p}(X)}\right].\end{aligned}$$

Now let us define our null and alternative hypotheses formally. We consider the following null hypothesis, $\mathbf{H}_0 : (D,X) \perp\!\!\!\perp T$, and the fixed alternative hypothesis, $\mathbf{H}_1 : (D,X)\not\perp\!\!\!\perp T$. $\mathbf{H}_1$ is the simple negation of $\mathbf{H}_0$. From Theorem 2.3, we know that $|bias_{sc,2}| \geqslant 0$ under $\mathbf{H}_1$. The inequality is not strict, and therefore, our test has trivial power along directions where $\tau_{sc,2}$ is not biased. Nonetheless, such a compromise is harmless as test power is practically irrelevant when the bias is absent, in which case, we may freely choose between both estimators.

We also investigate the behavior of the test under the sequence of Pitman local alternatives,

$$\mathbf{H}_{1n} : \{\mathbb{P}_n\left(D=d,T=t|X\right) = \mathbb{P}\left(D=d|X\right)\mathbb{P}\left(T=t\right) + n^{-1/2}\delta_{d,t}(X)\ a.s.\}_{d,t\in\{0,1\}},$$

for $\{\delta_{d,t}(\cdot)\}_{d,t\in\{0,1\}} \in \mathcal{H}_\delta$, with

$$\begin{aligned}\mathcal{H}_\delta \equiv \Bigg\{\{\delta_{d,t}(\cdot)\}_{d,t\in\{0,1\}} : (i) \sum_{d,t=0,1}\delta_{d,t}(\cdot) = 0, \text{ and}\\ (ii)\ \left(\mathbb{P}\left(D=d|X\right)\mathbb{P}\left(T=t\right) + c_0\delta_{d,t}(X)\right) \in [1-\epsilon,1) \text{ a.s. }\Bigg\},\end{aligned} \tag{4.2}$$

where c_0 is a constant that lies in $(0,1)$.

To construct the statistic, we also need estimators for the influence functions of $\hat{\tau}_{dr}$ and $\hat{\tau}_{sc}$. For the former, we let $\hat{\eta}_{dr}$ be given by (3.8), and for the latter, we use,

$$\hat{\eta}_{sc}(W;\hat{p},\hat{m}) \equiv \frac{D}{\mathbb{E}_n[D]}(\hat{\tau}(X)-\hat{\tau}_{sc}) + \sum_{d,t\in\{0,1\}}(-1)^{(d+t)}\hat{w}_{d,t}^{sc}(D,T,X;\hat{p})(Y-\hat{m}_{d,t}(X)). \tag{4.3}$$

Given these first-step estimators, our test statistic is defined as

$$\mathcal{T}_n = n\hat{V}_n^{-1}\left(\hat{\tau}_{dr}-\hat{\tau}_{sc}\right)^2, \tag{4.4}$$

where $\hat{V}_n \equiv \mathbb{E}_n\left[\left(\hat{\eta}_{dr}(W;\hat{p},\hat{m}) - \hat{\eta}_{sc}(W;\hat{p},\hat{m})\right)^2\right]$. Different from the original Hausman test where the denominator is the difference between the variances of the two estimators, we estimate the denominator by averaging the squared difference between the two asymptotic linear expansions. Our choice ensures that the denominator is always positive, which may not be the case when the original denominator is used.

In the following theorem, we characterize the asymptotic behavior of this statistic. Let $c_{1-\alpha}^*$ denote the $(1-\alpha)$'th quantile of the chi-squared distribution with one degree of freedom (χ_1^2).

Theorem 4.1 *Suppose Assumptions 1, 3, 4, and 5 hold. In addition, assume that $\tau(x)$ is not a degenerate*

function of x, then
(a) under the null space $\mathbf{H}_0$, $\hat{V}_n \xrightarrow{p} \rho_{sc} > 0$, and

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{T}_n \geqslant c^*_{1-\alpha}\right) = \alpha; \tag{4.5}$$

(b) under the alternative space $\mathbf{H}_1$,

$$\left|\frac{1}{n}\mathcal{T}_n - \rho_{sc}^{-1}(\tau_{sc} - \tau)\right| \xrightarrow{p} 0; \tag{4.6}$$

(c) under the local alternative $\mathbf{H}_{1n}$,

$$\mathcal{T}_n \xrightarrow{d} \chi^2_1(\mu_\delta), \tag{4.7}$$

where $\mu_\delta = \rho_{sc}^{-1}\left(\frac{\mathbb{E}[\delta_{1,1}(X)(\tau(X) - \tau_{sc})]}{\mathbb{E}[D]\,\mathbb{E}[T]}\right)^2$.

The theorem states that the test controls size under the null, and is consistent against fixed alternatives such that post-treatment bias is nonzero. In terms of the local power, our test will not detect local alternatives with deviations orthogonal to $(\tau(X) - \tau_{sc})$. Contrarily, local power is maximized when the deviation $\delta_{1,1}(X)$ is proportional to $(\tau(X) - \tau_{sc})$. It is also interesting to note that the test is only sensitive to local perturbations in $p(1, 1, \cdot)$, which makes sense as it is associated the sub-population that τ is based on.

Remark 6 *It is important to emphasize that our test should be seen as a "model validation" instead of a "model selection" procedure. For researchers who are concerned with the validity of Assumption 2, it may be tempting to perform a two-stage test, where in first stage, a Hausman specification test, such as the one discussed above, is used to "pretest" for covariate stationarity. And then in the second stage, the usual t-test is conducted based on $\hat{\tau}_{dr}$ or $\hat{\tau}_{sc}$, depending on the outcome of this test. As Guggenberger (2010a) and Guggenberger (2010b) have shown, standard inference procedure can lead to substantial size distortion. Although our context differs, similar analyses will apply. More discussion on this topic can be found in Leeb and Pötscher (2005), Sant'Anna and Song (2019), etc.*

5 Monte Carlo simulation study

To study the finite sample properties of the proposed estimators, we conduct two Monte Carlo experiments in this section. In the first experiment, Assumption 2 is violated, whereas in the second, the joint distribution of covariates and the treatment is independent of the timing of the treatment. For each design, we compare our estimator $\hat{\tau}_{dr}$, the one based on Sant'Anna and Zhao (2020) $\hat{\tau}_{sc}$, and canonical two-way fixed effect (TWFE) regression estimator.[4]

We provide simulation results for our DR DID estimator based on local linear $(p, q = 1)$ kernel estimators of the PS and OR functions. As described in Section 3.1, the PS is estimated using the local likelihood method with the logistic link function, whereas the OR is estimated using the standard local least squares estimator. We use the product Epanechnikov kernel for the continuous covariates, and for the

[4] By canonical TWFE estimator, we mean the coefficient, τ_{fe}, that is associated with the term, $T \cdot D$, in the following regression model,

$$Y = \alpha_1 + \alpha_2 T + \alpha_3 D + \tau_{fe}(T \cdot D) + \theta' X + \epsilon$$

discrete variables, we use the kernel given in (3.1). We compare the performance of our estimator $\hat{\tau}_{dr}$ and that of $\hat{\tau}_{sc}$ when the two different criteria of bandwidth selection: the log-likelihood and the least squares distance. In addition, we show how results differ depending on whether stabilizing weights are used. For TWFE regression, we consider two specifications: 1) a linear specification, where all the covariates enter linearly, and 2) a fully saturated specification, where, in addition to the linear terms, quadratic terms of the continuous variables and all the interactive terms of the covariates are also included.

We consider sample size $n = 1000$. For each design, we perform $5,000$ Monte Carlo simulations. We compare the various DID estimators for the ATT in terms of average bias, median bias, root mean square error (RMSE), empirical 95% coverage probability, the average length of a 95% confidence interval, and the average of their plug-in estimator for the asymptotic variance. We use normal approximation to compute the confidence intervals, with the asymptotic variances being estimated by their sample analogues. Additionally, we calculate the semiparametric efficiency bound under each design, so as to assess the potential loss of efficiency/accuracy associated with using inefficient DID estimators for the ATT. Finally, under each design, we conduct a Hausman-type test as described in Section 4.

5.1 Simulation 1: non-stationary covariate distribution

We first discuss the case where the stationarity condition fails to hold. Let $\mathbf{X} = (X_1, X_2, ..., X_6)$, where X_1 and X_2 are drawn from $\mathrm{Uniform}\,[-1,\ 1]$, X_3 and X_4 are binary variables, following $\mathrm{Bernoulli}\,(0.5)$, and the remaining two, X_5 and X_6, are distributed as $\mathrm{Binomial}\,(3,\ 0.5)$. The six variables are mutually independent.

Define

$$\begin{aligned}
f^{ps}_{1,0}(X) =& 0.4\sum_{s=1}^{2}(X_s - X_s^2) + 0.2\sum_{k=3}^{6} X_k + 0.1\left(\sum_{j\in\{3,5\}}(-1)^{j+1}X_jX_{j+1}\right.\\
&\left. + \sum_{l=1}^{2}\sum_{l'=3}^{6}(-1)^{l+1}X_lX_{l'} + \sum_{\ell=3}^{4}\sum_{\ell'=5}^{6}(-1)^{\ell+\ell'}X_\ell X_{\ell'}\right),\\
f^{ps}_{0,1}(X) =& 0.4(2X_1 + X_2 + X_1^2 - X_2^2 + X_1X_2) + 0.2\sum_{k=3}^{6}(-1)^{k+1}X_k + 0.1\left(\sum_{l=3}^{6}X_2X_l + \sum_{\ell=3}^{4}X_\ell X_6\right),\\
f^{ps}_{0,0}(X) =& 0.4(X_1 + 2X_2 - X_1^2 + X_2^2 - X_1X_2) + 0.2\sum_{k=3}^{6}(-1)^{k}X_k + 0.1\left(\sum_{l=3}^{6}X_1X_l + \sum_{\ell=3}^{4}X_\ell X_5\right),
\end{aligned}$$

and for the OR models,

$$\begin{aligned}
f^{or}_{base}(X) &= f^{or}_{het}(X) = 27.4X_1 + 27.4X_2 + 13.7X_1^2 + 13.7X_2^2 + 13.7X_1X_2,\\
f^{or}_{att}(X) &= 27.4X_1 + 13.7X_2 + 6.85\sum_{k=3}^{6}X_k - 15.
\end{aligned}$$

We consider the following data generating process

$$p^{s1}(d,t,X) = \begin{cases} \dfrac{\exp(f^{ps}_{d,t}(X))}{1 + \sum_{(d,t)\in\mathcal{S}_{\backslash(1,1)}}\exp(f^{ps}_{d,t}(X))}, & \text{if } (d,t)\in\mathcal{S}_{\backslash(1,1)}\\ \dfrac{1}{1 + \sum_{(d,t)\in\mathcal{S}_{\backslash(1,1)}}\exp(f^{ps}_{d,t}(X))}, & \text{otherwise.}\end{cases}$$

Let $U \sim \mathrm{Uniform}\,[0,\ 1]$. The treatment groups are assigned by

$$(D,T) = \begin{cases} (1,0), & \text{if } U \leqslant p^{s1}(1,0,X), \\ (0,1), & \text{if } p^{s1}(1,0,X) < U \leqslant p^{s1}(1,0,X) + p^{s1}(0,1,X), \\ (0,0), & \text{if } p^{s1}(1,0,X) + p^{s1}(0,1,X) < U \leqslant 1 - p^{s1}(1,1,X), \\ (1,1), & \text{if } 1 - p^{s1}(1,1,X) < U. \end{cases}$$

Next, building on Kang and Schafer (2007), we consider the following potential outcomes

$$Y_0(j) = 210 + f^{or}_{base}(X) + \epsilon_{het} + \epsilon_{j,0}, \text{ for } j = 0, 1, \tag{5.1}$$

$$Y_1(0) = 210 + 2f^{or}_{base}(X) + \epsilon_{het} + \epsilon_{0,1}, \tag{5.2}$$

$$Y_1(1) = 210 + 2f^{or}_{base}(X) + f^{or}_{att}(X) + \epsilon_{het} + \epsilon_{1,1}, \tag{5.3}$$

where $\epsilon_{het} \sim N(D \cdot f^{or}_{het}, 1)$ and $\epsilon_{d,t}$, $d, t \in \{0, 1\}$ are independent standard normal random variables.

Under this design, the covariate distribution does not exhibit time variation. However, the PS function different in two cross sections. The mean absolute difference between $p^{s1}(1,1,X)$ and $p^{s1}(1,0,X)$ and that between $p^{s1}(0,1,X)$ and $p^{s1}(0,0,X)$ are both around 0.125, with the maximum difference being as large as 0.63. Hence, we expect all but $\hat{\tau}_{dr}$ to yield biased estimates. In addition, the test of stationarity would reject the null hypothesis with high probability. These conjectures are confirmed by results reported in Table 1 and Table 2.

DGP 1: Non-stationary Distribution.

True value of ATT: 4.31. Semiparametric Efficiency Bound: 1753.6

Two-way Fixed Effect Estimators

	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{fe}$	Linear	-10.413	-10.463	10.918	10407.642	0.122	12.624
$\hat{\tau}_{fe}$	Saturated	-11.160	-11.151	11.563	8787.421	0.047	11.606

Doubly Robust Estimators with Stabilized Weights

	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	0.016	0.044	1.372	1863.380	0.951	5.333
$\hat{\tau}_{dr}$	LS	0.014	0.037	1.387	1908.760	0.951	5.357
$\hat{\tau}_{sc}$	ML	4.327	4.343	4.454	1030.981	0.017	3.967
$\hat{\tau}_{sc}$	LS	4.328	4.343	4.454	1031.856	0.017	3.968

Doubly Robust Estimators with Unstabilized Weights

	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	0.139	0.160	1.382	1877.111	0.952	5.350
$\hat{\tau}_{dr}$	LS	0.144	0.164	1.413	1977.300	0.952	5.387
$\hat{\tau}_{sc}$	ML	4.626	4.631	4.750	1036.923	0.010	3.978
$\hat{\tau}_{sc}$	LS	4.627	4.631	4.751	1038.073	0.010	3.980

Note: Simulations based on 5,000 Monte Carlo experiments. $\hat{\tau}_{fe}$ the TWFE regression estimator, $\hat{\tau}_{dr}$ is our proposed DR DID estimator (3.6), and $\hat{\tau}_{sc}$ is the DR DID estimator (4.1) by Sant'Anna and Zhao (2020). For TWFE regression, we use a linear specification, "Linear" and a saturated specification, "Saturated". For DR DID estimators, the PS and the OR models are estimated nonparametrically, using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, "ML", and the least squares criterion, "LS", respectively. Lastly, "Spec.", "CV Crit.", "Avg. Bias", "Med. Bias", "RMSE", "Asy. Var.", "Cover.", and "CIL", stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

Table 1: Monte Carlo study of the performance of DR DID estimators under DGP 1. Sample size: $n = 1,000$.

First, results in Table 1 suggests that both $\hat{\tau}_{fe}$ and $\hat{\tau}_{sc}$ are severely biased under this DGP, whereas $\hat{\tau}_{dr}$ exhibits negligible bias on average. In addition, among the three set of estimators considered, only our proposed estimator achieves nominal coverage rate. This result is robust to the bandwidth selection method and whether the inverse propensity score weights are stabilized. Notice that the performance of the TWFE does not improve with a fully-saturated specification, indicating that adding nonlinear terms into a TWFE regression does not help with identifying heterogeneous treatment effects in general.

This finding sheds light to the importance of an estimator being robust to the violation of the covariate stationarity assumption.

Note also that $\hat{\tau}_{dr}$ based on stabilized weights improves upon the unstabilized counterparts, in terms of the Monte Carlo bias, RMSE, average length of confidence intervals. This also holds true for $\hat{\tau}_{sc}$. Bias is visibly lower and coverage rate higher when stabilized weights are employed. Such a finding highlights the practical importance of using weights that are normalized to sum up to one.

With regards to efficiency, we find that our proposed estimator is consistently close to the semiparametric efficiency bound. This result corroborates the findings of Theorem 3.1. Compared to the DR DID estimators, the TWFE estimators have much larger asymptotic variances on average, causing the empirical coverage rate to be higher than that of $\hat{\tau}_{sc}$.

Test results are reported in Table 2. Under DGP 1, $\mathbf{H}_0$ does not hold, and therefore, the empirical rejection frequencies reported in Table 2 amount to empirical power of the stationarity test. Our proposed test achieves good empirical power with this DGP under all the scenarios we consider. Interestingly, tests based on unstabilized weights exhibits relative better power property.

Test of Stationarity under DGP 1: Non-stationary Distribution

Stb. Wgt.	CV Crit.	Avg. Test Stats.	Emp. Pow. (0.10)	Emp. Pow. (0.05)	Emp. Pow. (0.01)
Yes	ML	21.263	0.993	0.991	0.969
Yes	LS	21.214	0.992	0.989	0.968
No	ML	22.849	0.995	0.992	0.976
No	LS	22.721	0.993	0.991	0.975

Note: Simulations based on 5,000 Monte Carlo experiments. Test statistic is calculated based on (4.4). “Stb. Wgt.”, “Avg. Test Stats.”, and “Emp. Pow. (α)” stand for the stabilized weights, average test statistics, and empirical power of the test with a nominal size α, respectively.

Table 2: Monte Carlo result of the stationarity test under DGP 1. Sample size: $n = 1,000$.

5.2 Simulation 2: stationary covariate distribution

We now modify the first design slightly by taking the average of propensity scores over time, while holding all other aspects of the DGP fixed in the mean time. Specifically, we let

$$p^{s2}(d,t,X) = \mathbb{P}^{s1}(T=t)\,(p^{s1}(d,1,X) + p^{s1}(d,0,X)),$$

where $\mathbb{P}^{s1}(T=t) = \mathrm{E}[p^{s1}(1,t,X) + p^{s1}(0,t,X)]$. The treatment groups are then assigned based on the realization of a standard uniform random variable on the unit interval partitioned by $\{p^{s2}(d,t,X)\}_{d,t\in\{0,1\}}$. Furthermore, the potential outcomes are determined by (5.1)-(5.3). Contrary to the first DGP, both the covariate distribution and the propensity score function are stationary. Therefore, we expect that both $\hat{\tau}_{dr}$ and $\hat{\tau}_{sc}$ are consistent for the true ATT. Moreover, the empirical rejection rate of the stationarity test would converge to the nominal size. The Monte Carlo under this DGP results are summarized in Tables 3 and 4.

DGP 2: Stationary Distribution

True value of ATT: 9.13. Semiparametric Efficiency Bound: 796.8

Two-way Fixed Effect Estimators

	Spec.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{fe}$	Linear	-10.516	-10.557	10.993	9952.373	0.099	12.353
$\hat{\tau}_{fe}$	Saturated	-10.473	-10.496	10.861	7943.739	0.047	11.039

Doubly Robust Estimators with Stablized Weights

	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	-0.065	-0.056	1.304	1746.305	0.954	5.173
$\hat{\tau}_{dr}$	LS	-0.068	-0.056	1.309	1758.715	0.954	5.181
$\hat{\tau}_{sc}$	ML	-0.046	-0.058	0.977	958.952	0.950	3.835
$\hat{\tau}_{sc}$	LS	-0.046	-0.058	0.978	959.581	0.950	3.836

Doubly Robust Estimators with Unstablized Weights

	CV Crit.	Avg. Bias	Med. Bias	RMSE	Asy. Var.	Cover.	CIL
$\hat{\tau}_{dr}$	ML	-0.046	-0.042	1.305	1747.280	0.954	5.174
$\hat{\tau}_{dr}$	LS	-0.047	-0.041	1.314	1770.619	0.953	5.186
$\hat{\tau}_{sc}$	ML	-0.052	-0.071	0.986	957.708	0.947	3.833
$\hat{\tau}_{sc}$	LS	-0.051	-0.070	0.987	958.420	0.947	3.834

Note: Simulations based on 5,000 Monte Carlo experiments. $\hat{\tau}_{fe}$ the TWFE regression estimator, $\hat{\tau}_{dr}$ is our proposed DR DID estimator (3.6), and $\hat{\tau}_{sc}$ is the DR DID estimator (4.1) by Sant'Anna and Zhao (2020). For TWFE regression, we use a linear specification, "Linear" and a saturated specification, "Saturated". For DR DID estimators, the PS and the OR models are estimated nonparametrically, using a local linear least squares and a local linear logistic regression, respectively. Bandwidth for the PS function is selected with the log-likelihood criterion, "ML", and the least squares criterion, "LS", respectively. Lastly, "Spec.", "CV Crit.", "Avg. Bias", "Med. Bias", "RMSE", "Asy. Var.", "Cover.", and "CIL", stand for the specification, cross-validation criterion, average simulated bias, median simulated bias, simulated root mean-squared errors, average of the plug-in estimator for the asymptotic variance, 95% coverage probability, and 95% confidence interval length, respectively. See the main text for further details.

Table 3: Monte Carlo study of the performance of DR DID estimators under DGP 2. Sample size: $n = 1,000$.

Test of Stationarity under DGP 2: Stationary Distribution					
Stb. Wgt.	CV Crit.	Avg. Test Stats.	Emp. Size (0.10)	Emp. Size (0.05)	Emp. Size (0.01)
Yes	ML	0.994	0.094	0.046	0.009
Yes	LS	0.995	0.094	0.046	0.010
No	ML	1.009	0.098	0.050	0.008
No	LS	1.010	0.097	0.050	0.008

Note: Simulations based on 5,000 Monte Carlo experiments. Test statistic is calculated based on (4.4). "Stb. Wgt.", "Avg. Test Stats.", and "Emp. Size (α)" stand for the stabilized weights, average test statistics, and empirical size of the test with a nominal size α, respectively.

Table 4: Monte Carlo result of the stationarity test under DGP 2. Sample size: $n = 1,000$.

Contrary to the findings of Table 1, both $\hat{\tau}_{dr}$ and $\hat{\tau}_{sc}$ show little bias, and their confidence intervals achieve nominal coverage. Their performance is consistently good across different bandwidth selection methods and the weighting choices. The TWFE estimators, on the other hand, are still severely biased. Additionally, they have almost negligible coverage even with much wider confidence intervals than DR DID estimators.

In terms of efficiency, $\hat{\tau}_{sc}$ is reasonably close to the semiparametric efficiency bound, as predicted by Theorem 4.1. Asymptotic variance of $\hat{\tau}_{dr}$ is 2.2 times of the semiparametric efficiency bound on average, which is significantly lower than that of the TWFE estimators.

Under DGP 2, Assumption 2 and thus $\mathbf{H}_0$ hold, empirical rejection frequency is equivalent empirical size of the tests. Results in Table 4 demonstrate that our stationarity test control size in finite sample, across difference bandwidth and weighting choices consistently.

6 Empirical illustration: the effect of tariff reduction on corruption

In this section, we revisit a study from Sequeira (2016) on the effect of import tariff liberalization on corruption patterns. Before the phaseout of high tariffs between South Africa and Mozambique, bribery payment was pervasive. It serves as a means to evade tariff taxes. According to Sequeira and Djankov (2014), bribery payments can be found in nearly 80% of all shipments records in a random sample of tracked shipments prior to a tariff rate reduction in 2008.

This tariff change is the result of a long-standing trade agreement between South Africa and Mozambique. The agreement, Southern African Development Community (SADC) Trade Protocol, was signed in 1996. It set the road map for import tariff reductions between 2001 and 2015. The largest reduction took place in 2008, with the average nominal rate dropping by 5%. The effect of such tariff liberalization scheme is considerable, as both the probability and the amount of bribe payments decreased significantly after the phaseout.

Theoretically speaking, lower tariff rates may either disencourage bribery as expected profits from tax evasion are now reduced (Allingham and Sandmo, 1972; Mishra, Subramanian and Topalova, 2008; Sequeira and Djankov, 2014), or it may incentivize such behavior through an income effect, since reduced tariff rates make agents wealthier, thus increasing their ability to pay higher bribes (Feinstein, 1991; Slemrod and Yitzhaki, 2002). The empirical findings of Sequeira (2016) helps to shed light on this discussion.

To formally study the causal relationship between tariff rate reduction and changes in bribery, Sequeira

(2016) exploits a quasi-experimental variation induced by trade protocol: not all products were subject to the change in tariff rate during the period under analysis. As a result, the products that did not experience a change in the tariff serve as a credible control group. It is therefore possible to use the DID design to analyze how tariff rate changes may affect bribe patterns along the trade routes.

Sequeira (2016) collected data measuring the bribe payment along the trade routes between the two countries, from 2007 to 2013. This data set has a repeated cross section structure. Sequeira (2016) mainly considers the following two TWFE regressions:

$$\begin{aligned}
\text{(Linear)} \qquad & y_{it} = \gamma_1 TariffChangeCategory_i \times Post + \mu Post \\
& \gamma_2 TariffChangeCategory_i + \beta_2 BeselineTariff_i + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \\
\text{(Interactive)} \qquad & y_{it} = \gamma_1 TariffChangeCategory_i \times Post + \mu Post \\
& \gamma_2 TariffChangeCategory_i + \beta_2 BeselineTariff_i + \Gamma_i + \Gamma_i \times Post \\
& + p_i + w_t + \delta_i + \epsilon_{it},
\end{aligned}$$

where y_{it} is one of the measurements of bribery payments for shipment i in period t. $TariffChangeCategory$ is the treatment indicator, which takes value 1 if the product shipped experienced a tariff reduction in 2008, and 0 otherwise. $Post$ is the indicator for the post-treatment period. It equals 1 for the years following 2008. $BaselineTariff$ stands for the tariff rates before 2008. A vector of controls, Γ_i, industry, year, and clearing agent fixed effects, p, ω, δ, are also included in the regressions. The interactive specification differs from the linear one by an interaction of $Post$ and the covariates, Γ_i.

The parameter of interest is γ_1 in both specifications. It identifies the ATT of tariff rate reduction on the corruption under stringent conditions of the unknown treatment effects. As is well-known in the literature, identification by TWFE regression is likely to fail when the treatment effect is heterogeneous, the functional form is misspecified, and/or when there exist compositional changes. Our proposed DR DID estimator, τ_{dr}, the one based on Sant'Anna and Zhao (2020), $\hat{\tau}_{sc}$, and various other estimators are potentially better suited for the task of identifying and consistently estimating the ATT in the present context. . The debiased machine learning difference-in-differences (DML DID) estimator proposed by Chang (2020) is a notable example. The estimator removes the first-order bias of the semiparametric IPW estimator by Abadie (2005), and allows for machine learning methods to be used in the first-step estimation. Using this estimator, he found much larger effects than those reported by Sequeira (2016). In what follows, we estimate the ATT using our proposed DR DID estimator and compare the results with those produced by the two estimators discussed above.

Toward this end, we first estimate the PS and OR functions based on local linear logistic regression and local linear OLS, respectively. Following Sequeira (2016), we employ four different outcome measures: 1. a binary variable denoting if a bribe is paid, the logarithmic form, $log(x + 1)$, of the amount of bribe payment, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the value of the shipment, respectively. The covariates include: a dummy variable indicating if the shipper is a large firm, dummy variables for perishable products, agricultural goods, and shipments pre-inspected at origin, shipments from South Africa, and the arrival day of week.

For each estimator except DML DID, we report both the unclustered standard errors based on asymptotic approximation and the cluster-robust standard errors based on the bootstrap procedure in Algorithm 2. Likewise, we conduct two sets of tests – the one using unclustered influence functions based on (4.4) and the other that accounts for clustering using a bootstrap procedure given in Algorithm

3.

Outcome	Prob(bribe)	Log(bribe)	Log(bribe/shpt.val.)	Log(bribe/shpt.tonn.)
TWFE - Linear				
ATT	-0.429	-3.748	-0.011	-1.914
	(0.083)	(0.724)	(0.003)	(0.341)
	[0.135]	[1.065]	[0.003]	[0.504]
TWFE - Interactive				
ATT	-0.296	-2.928	-0.01	-1.597
	(0.082)	(0.746)	(0.004)	(0.402)
	[0.125]	[0.949]	[0.004]	[0.45]
DML DID				
ATT	-0.702	-6.43	-0.028	-4.647
	(0.246)	(2.154)	(0.009)	(1.010)
DR DID - Stationary				
ATT	-0.258	-2.35	-0.005	-3.484
	(0.068)	(0.651)	(0.004)	(2.22)
	[0.099]	[0.792]	[0.005]	[2.981]
DR DID - Nonstationary				
ATT	-0.298	-2.748	-0.003	-1.869
	(0.085)	(0.829)	(0.003)	(0.598)
	[0.11]	[0.917]	[0.004]	[0.616]

Notes: Same data used by Sequeira (2016). The results represent the estimated ATT of tariff rate reduction on bribery payment behavior. Columns 1 through 4 denote estimates for dependent variables representing whether a bribe is paid, the logarithmic form, $log(x + 1)$, of the amount of bribe paid, the logarithmic form of the amount of bribe paid as a share of the value of the shipment, and as a share of the value of the shipment, respectively. We compare five different DID estimators for the ATT: 1. the two-way fixed effect estimator based on specifications in Column (1) of Tables 8-11 in Sequeira (2016); 2. the two-way fixed effect estimator based on Column (2) from Tables 8-11 in Sequeira (2016); 3. DML DID estimator with first-step LASSO estimation based on (3.2) in Chang (2020); 4. DR DID estimator based on (4.1), and 5. DR DID estimator based on (3.6). The same set of covariates are used for the last three estimators. See the main text for further details of the covariates. Continuous variables are re-scaled between 0 and 1, and then added in with binary variables. For DR DID estimators, the PS and the OR models are estimated nonparametrically, using a local linear OLS and a local linear logistic regression, respectively. Bandwidth for the local linear logistic regression is selected with the log-likelihood criterion. Numbers in the parentheses are unclustered standard errors based on asymptotic approximation. Numbers in brackets refer to standard errors clustered at the level of four-digit HS code. Cluster-robust standard errors are calculated following Algorithm 2 with 9999 bootstrap draws.

Table 5: Difference-in-differences estimation results for Sequeira (2016)

Outcome	Prob(bribe)	Log(bribe)	Log(bribe/shpt.val.)	Log(bribe/shpt.tonn.)
Panel A: Cramér Test				
p value ($\mathbf{H}^{cr}_{0,1}$)	0.000		0.000	0.000
$\mathbf{H}^{cr}_{0,1}$ rejected?	Yes		Yes	Yes
p value ($\mathbf{H}^{cr}_{0,0}$)	0.258		0.251	0.228
$\mathbf{H}^{cr}_{0,0}$ rejected?	No		No	No
Panel B: Hausman Test				
p value	0.202	0.217	0.147	0.449
$\mathbf{H}_0$ rejected?	No	No	No	No
p value (cls.)	0.245	0.241	0.395	0.59
$\mathbf{H}_0$ rejected? (cls.)	No	No	No	No

Notes: Test statistic in Panel A is calculated based on (6.1). Rows 2 and 4 present results for tests with a 5% level of significance. The p values in Rows 1 and 3 are calculated using the bootstrap procedure implemented in the R package `cremer` with 9999 bootstrap draws. Test statistic in Panel B is calculated based on (4.4). Rows 2 and 4 present results for tests with a 5% level of significance. The p values in Row 3 are calculated following the bootstrap procedure in Algorithm 3 with 9999 bootstrap draws.

Table 6: Stationarity test results for Sequeira (2016)

Table 5 contains results based on various estimators. We first observe that the point estimates are negative for all measures of bribery payment, consistent with the finding of Sequeira (2016). We find that a 10 % tariff reduction reduces the probability of paying a bribe and the amount by 25.8% . Clustering slightly inflates the standard errors but does not change inference results qualitatively.

Next, we notice that the DML DID estimates are approximately twice as large as the TWFE and DR DID estimates, across the outcome measures adopted. For instance, when the outcome variable is the log of bribe payment, the effect found by DML DID is 119%, 173%, and 134% larger in magnitude than those by TWFE (interactive), $\hat{\tau}_{sc}$ and $\hat{\tau}_{dr}$, respectively. The finding is consistent with that reported by Table 2 of Chang (2020). The high standard errors indicate that the ATT's are not precisely estimated by the DML DID estimator. We conjecture that the relatively small sample size and the use of sample splitting may have caused this issue.

On the contrary, results based on the two DR DID methods are generally close to the TWFE estimates with the interactive specification. There are two exceptions. For the first, both of the DR DID estimators fail to find significant negative effect of tariff rate reduction on the log of ratio between bribery payment and shipment values, in contradiction with the TWFE estimates. For the second, when the log of ratio between bribery payment and tonnage is considered, DR DID proposed by Sant'Anna and Zhao (2020) reports a large yet insignificant ATT estimate, while our proposed estimator produces results similar to TWFE estimates. This difference may be driven by the presence of compositional changes. To formally answer this question, we resort to the test results that are summarized in Table 6.

First, we conduct a test on the equality of covariate distributions before and after the treatment. For this task, we use the two-sample Cramér test developed by Baringhaus and Franz (2004). Specifically, we test the hypotheses $\mathbf{H}^{cr}_{0,d}$: $\mathbb{P}(X \leqslant \cdot | D = d, T = 1) = \mathbb{P}(X \leqslant \cdot | D = d, T = 0)$ against the general alternative $\mathbf{H}^{cr}_{1,d}$: $\mathbb{P}(X \leqslant \cdot | D = d, T = 1) \neq \mathbb{P}(X \leqslant \cdot | D = d, T = 0)$, for $d = 0, 1$, based on the following statistic

$$
\begin{aligned}
\mathcal{T}_j^{cr} = & \frac{n_{d,1} \cdot n_{d,0}}{n_{d,1} + n_{d,0}} \left\{ \frac{1}{n_{d,1} n_{d,0}} \sum_{i=1}^{n_{d,1}} \sum_{d=1}^{n_{d,0}} \left\| X_i^{(d,1)} - X_j^{(d,0)} \right\| - \frac{1}{2n_{d,1}^2} \sum_{i=1}^{n_{d,1}} \sum_{j=1}^{n_{d,1}} \left\| X_i^{(d,1)} - X_j^{(d,1)} \right\| \right. \\
& \left. - \frac{1}{2n_{d,0}^2} \sum_{i=1}^{n_{d,0}} \sum_{j=1}^{n_{d,0}} \left\| X_i^{(d,0)} - X_j^{(d,0)} \right\| \right\}, \qquad (6.1)
\end{aligned}
$$

where $n_{d,t} = \mathbb{E}_n[1\{D = d, T = t\}]$, and $X^{(d,t)}$ is the subset of X such that $1\{D = d, T = t\} = 1$. For each outcome measure, we report the bootstrapped p-values for the corresponding specification of X in Panel A of Table 6. The p values suggest that, conditional on being treated, covariate distributions before and after the treatment are found to be significantly different at 5% level. These results imply that there may be compositional change on the covariates. However, we are not able to reject the null hypothesis for the control group. From these mixed results alone, we cannot infer whether the covariate distribution is not stationary across time and to what extent this change may affect the estimates of ATT. To tackle this question, it is imperative that the test incorporate information from the outcome variable. Our proposed test shares this feature.

The p values from Panel B of Table 6 show that there is no evidence of the post-treatment bias, regardless of the outcome measure used for the test. Clustering at the HS code level does not change the results. This finding suggests that even though there is strong evidence that, for the treated group, baseline covariates are not identically distributed before and after the treatment, this difference, however, does not seem to have direct implications for the estimation of ATT. The result also corroborates the assertion by Sequeira (2016) that her findings are robust to the inclusion of an interactive term between $Post$ and the covariates.

In sum, our results support the conclusion of Sequeira (2016) that tariff liberalization decreases corruption, and contrary to the findings of Chang (2020), our DR DID estimates suggest the size of effects is approximately the same as that of the original paper.

7 Concluding remarks

In this paper, we developed a doubly robust estimator for the ATT in difference-in-differences framework where covariates can vary over time. We establish large sample properties for the proposed estimator when the nuisance functions are estimated nonparametrically. In particular, we provide novel results on uniform convergence rate of local logit estimator with mixed data. We provide extensive discussions comparing our proposed DR estimator with those developed by Sant'Anna and Zhao (2020). A Hausman-type test is proposed to assess the validity of the ATT estimators under consideration. The finite sample performance of the tests is examined by means of two Monte Carlo experiments. All the finite sample findings are in line with the asymptotic results. Lastly, we illustrated the attractiveness through an empirical application concerning the effect of tariff liberalization on the corruption.

Our results can be extended in multiple directions. We have limited discussion to the canonical two-period models. When the number of time period is greater than two and when the treatment adoption is staggered, Callaway and Sant'Anna (2021) show that a family of group-time average treatment effects and their aggregates are identified under a general no-compositional-change assumption. Once we relax the condition, treatment effect depends not only on when the treatment is first received and the time that has passed since being treated, but also on the time when the treatment effect is evaluated. As such, careful interpretation of the resulting causal parameters will be of first order importance.

When estimating the nuisance functions, fully nonparametric procedures may not be feasible for moderately large dimensional covariates because of the "curse of dimensionality". In these cases, researchers often adopt semi-parametric methods, such as the single-index models. These type of estimators reduce the dimensionality by imposing a low-dimensional index structure, and in the meanwhile, have the link function estimated nonparametrically. Thus, flexibility in the functional specification is partially retained. Generalizing results from Liu, Ma and Wang (2018) and Sun, Yan and Li (2021) to our DR DID formulation seems prospective for future research.

References

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

Allingham, M. G., and Sandmo, A. (1972), "Income tax evasion: A theoretical analysis," *Journal of public economics*, 1(3-4), 323–338.

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

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

Baringhaus, L., and Franz, C. (2004), "On a new multivariate two-sample test," *Journal of multivariate analysis*, 88(1), 190–206.

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.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, 119(1), 249–275.

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

Blundell, R., Dias, M. C., Meghir, C., and van Reenen, J. (2004), "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, 2(4), 569–606.

Busso, M., Dinardo, J., and McCrary, J. (2014), "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *The Review of Economics and Statistics*, 96(5), 885–895.

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

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

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

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

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.

Durbin, J. (1954), "Errors in variables," *Revue de l'institut International de Statistique*, pp. 23–32.

Eguchi, S., Yoon Kim, T., and Park, B. U. (2003), "Local likelihood method: a bridge over parametric and nonparametric regression," *Nonparametric Statistics*, 15(6), 665–683.

Escanciano, J. C., Jacho-Chávez, D., and Lewbel, A. (2016), "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7(2), 561–589.

Fan, J., and Gijbels, I. (1996), *Local polynomial modelling and its applications* Routledge.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, 90(429), 141–150.

Feinstein, J. S. (1991), "An econometric analysis of income tax evasion and its detection," *The RAND Journal of Economics*, pp. 14–35.

Frölich, M. (2006), "Non-parametric regression for binary dependent variables," *The Econometrics Journal*, 9(3), 511–540.

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.

Guggenberger, P. (2010*a*), "The impact of a Hausman pretest on the asymptotic size of a hypothesis test," *Econometric Theory*, 26(2), 369–382.

Guggenberger, P. (2010*b*), "The impact of a Hausman pretest on the size of a hypothesis test: The panel data case," *Journal of Econometrics*, 156(2), 337–343.

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.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, 99(468), 1015–1026.

Hausman, J. A. (1978), "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, pp. 1251–1271.

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.

Kong, E., Linton, O., and Xia, Y. (2010), "Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model," *Econometric Theory*, 26(5), 1529–1564.

Lee, Y. Y. (2018), "Efficient propensity score regression estimators of multivalued treatment effects for the treated," *Journal of Econometrics*, 204(2), 207–222.

Leeb, H., and Pötscher, B. M. (2005), "Model selection and inference: Facts and fiction," *Econometric Theory*, 21(1), 21–59.

Li, Q., and Ouyang, D. (2005), "Uniform convergence rate of kernel estimation with mixed categorical and continuous data," *Economics Letters*, 86(2), 291–296.

Li, Q., and Racine, J. S. (2007), *Nonparametric econometrics: theory and practice* Princeton University Press.

Liu, J., Ma, Y., and Wang, L. (2018), "An alternative robust estimator of average treatment effect in causal inference," *Biometrics*, 74(3), 910–923.

Mammen, E., Rothe, C., and Schienle, M. (2016), "Semiparametric estimation with generated covariates," *Econometric Theory*, 32(5), 1140–1177.

Masry, E. (1996), "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Millimet, D. L., and Tchernis, R. (2009), "On the specification of propensity scores, with applications to the analysis of trade policies," *Journal of Business & Economic Statistics*, 27(3), 397–415.

Mishra, P., Subramanian, A., and Topalova, P. (2008), "Tariffs, enforcement, and customs evasion: Evidence from India," *Journal of public Economics*, 92(10-11), 1907–1925.

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

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

Powell, J. L., and Stoker, T. M. (1996), "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics*, 75(2), 291–316.

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.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2022), "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *arXiv preprint arXiv:2201.01194*, .

Rothe, C., and Firpo, S. (2019), "Properties of doubly robust estimators when nuisance functions are estimated nonparametrically," *Econometric Theory*, 35(5), 1048–1087.

Sant'Anna, P. H., and Song, X. (2019), "Specification tests for the propensity score," *Journal of Econometrics*, 210(2), 379–404.

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

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.

Sequeira, S. (2016), "Corruption, trade costs, and gains from tariff liberalization: Evidence from Southern Africa," *American Economic Review*, 106(10), 3029–63.

Sequeira, S., and Djankov, S. (2014), "Corruption and firm behavior: Evidence from African ports," *Journal of International Economics*, 94(2), 277–294.

Slemrod, J., and Yitzhaki, S. (2002), "Tax avoidance, evasion, and administration," in *Handbook of public economics*, Vol. 3 Elsevier, pp. 1423–1470.

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

Staniswalis, J. G. (1989), “The kernel estimate of a regression function in likelihood-based models,” *Journal of the American Statistical Association*, 84(405), 276–283.

Sun, Y., Yan, K. X., and Li, Q. (2021), “Estimation of average treatment effect based on a semiparametric propensity score,” *Econometric Reviews*, 40(9), 852–866.

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

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

Wu, D.-M. (1973), “Alternative tests of independence between stochastic regressors and disturbances,” *Econometrica: journal of the Econometric Society*, pp. 733–750.

Wu, D.-M. (1974), “Alternative tests of independence between stochastic regressors and disturbances: Finite sample results,” *Econometrica: Journal of the Econometric Society*, pp. 529–546.

Zeldow, B., and Hatfield, L. A. (2021), “Confounding and regression adjustment in difference-in-differences studies,” *Health services research*, 56(5), 932–941.

Zimmert, M. (2019), “Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding,” *arXiv preprint arXiv: 11809.01643v4*, .
URL: *http://arxiv.org/abs/1809.01643*