Alternative Semi-Parametric Approaches and Machine Learning Algorithms in The Difference-In-Difference Design: a Monte-Carlo Evaluation*

Tommaso Manfè†1 and Luca Nunziata‡2,3

1University of Chicago, Booth School of Business
2University of Padua
3IZA

December 22, 2022

Abstract

We discuss the potentially severe bias in the Difference-in-Difference (DiD) design of the commonly-used regression specification known as Two-Way-Fixed-Effects (TWFE) when researchers must invoke the conditional trend assumption and we evaluate alternative corrections building on Zeldow and Hatfield (2019). We propose a semi-parametric estimator robust when covariates are distributed heterogeneously among both the treatment groups and time periods and compare it with those suggested by the literature through Monte Carlo simulations, with a specific focus on the semi-parametric doubly robust DiD of Sant'Anna and Zhao (2020). The estimators are also modified to allow for machine-learning first-stage estimates, following the literature of Chernozhukov et al. (2018). Results show that semi-parametric estimators outperform regression, even if TWFE corrections provide substantial benefits. Following Sequeira (2016), we estimate the effect of tariff reduction on bribing behavior by analyzing trades between South Africa and Mozambique during the period 2006–2014. Contrarily to the replication in Chang (2020), our findings show that the effect is close and even lower in magnitude than the one in the original paper. Still, the contribution reinforces the evidence that indeed tariff reductions tend to weaken bribing behavior.

Keywords: Difference-in-Difference, Monte-Carlo simulations, Semi-parametric, Machine-learning.

JEL Codes: C10, C13, C14, C18, C23.

*We thank Pedro H.C. Sant'Anna and Scott Cunningham for their precious help and advise. The usual disclaimer applies. Available coding material can be found at: https://github.com/tommaso-manfe.

†tommaso.manfe@chicagobooth.edu

‡luca.nunziata@unipd.it

1 Introduction

Difference-in-Difference (DiD) is a widespread research design that estimates the causal effects of a policy treatment that affects a specific group of subjects, called the treated group, while leaving unaffected another typically comparable group, referred to as the control group. The rationale of this empirical strategy is that if treated and control groups are subject to the same time trend, the control group can be used to estimate the counterfactual potential outcome in absence of treatment for the treatment group. This is achieved by adding the mean change of the outcome variable for the non-treated over time to the mean level of the outcome variable for the treated before treatment to obtain the mean outcome the treated would have experienced in absence of treatment.

The standard DiD regression, the Two-Way-Fixed-Effects (TWFE) estimator, can deliver the causal effect of interest when some assumptions are satisfied. In particular, it is assumed that the differences over time in the expected potential outcomes in absence of treatment are independent of whether an individual belongs to either the treated or the control group. This assumption, known as the unconditional parallel trend, can be stated either as parallel counterfactual trends in absence of treatment in the post-intervention period t_1 or as parallel pre-trends in the pre-intervention period t_0 and common shocks in t_1.

Broadly speaking, the parallel counterfactual trend assumption is implausible if selection into treatment depends on individual characteristics that correlate with the outcome variable dynamics. As a consequence, a weaker assumption consists of assuming the parallel trend hypothesis holds after conditioning of individual observable characteristics, the so-called "conditional trend assumption". However, including covariates in the traditional TWFE does come with severe threats. In fact, Zeldow and Hatfield (2019) shows that, when the effect of the covariate on the outcome varies in time, the traditional TWFE works only in the case the covariate is time-invariant and the mean of its distribution is the same among treated and controls. To address potential imbalances between the two groups, the authors propose a TWFE specification that includes the interaction between the covariate and the time dummy. Building on that, we show that the correction that adds as well the interaction between the covariate and the treatment group dummy has a lower bias in real-world scenarios. However, (Sant'Anna and Zhao, 2020) shows that TWFE is still biased when dealing with heterogeneous treatment effects and does capture the potential non-linearities in the data generating process.

To overcome some limitations of the TWFE, several alternatives have been suggested by the literature. In particular, Heckman et al. (1997), to allow for covariate-specific trends, proposes an Outcome Regression (OR) strategy based on assuming a model for the conditional expected value of the outcome for the control group and using it to predict the outcome for the treated units based on their empirical distribution of the covariates. Alternatively, Abadie (2005) avoids directly modeling the outcome evolution but focuses on the treatment model, i.e. the conditional probability of being in the treatment group given a set of covariates. The suggested Inverse Probability Weighting (IPW) estimator adjusts for confounding factors by estimating a propensity score used to balance baseline individual characteristics in the treated and control groups. Recently Sant'Anna and Zhao (2020) combined the OR and the IPW approaches into a doubly robust estimand for the ATT (named DRDiD). The double robustness property means that if either the propensity score model or the outcome regression models are misspecified (but not both), then resulting will generally be consistent.

We propose a set of new alternative semi-parametric estimators that can be applied to a DiD setting. In particular, we propose an alternative version of the DRDiD estimator by Sant'Anna and Zhao (2020) that employs machine learning algorithms for the first-stage estimates, following Chernozhukov et al. (2018). The rationale is that using machine learning allows avoiding assuming a specific parametric form for both the treatment and outcome models. This is beneficial because, since the researchers do not know a prori the right parametric form of the phenomenon under study, they thus risk to misspecify it.

All these methodologies assume time-invariant controls, but in real practice, it is likely to observe the distribution of the covariates to change in time. As a consequence, building on Blundell et al. (2004) and Blundell and Dias (2009), we additionally propose an alternative estimator that aims at achieving balance in the distribution of the relevant observable characteristics among the four cells defined by eligibility and time. Under this approach, which we call triple inverse probability weighting regression adjusted (3IPWRA), the propensity score is computed as the probability of belonging to the treated group in the post-treatment period. More precisely, the initial sample is split according to the four groups defined by the interaction of time and treatment, and the propensity score is separately computed in the three sub-samples obtained by merging the treated group in the post-treatment period with each one of the three remaining sub-populations one at a time. In this way, a specific propensity score for each of the four groups is defined. Differently than Blundell and Dias (2009), the propensity score is then not used for matching but for inverse probability weighting. In fact, 3IPWRA uses the estimated propensity score to calculate the Horvitz-Thompson inverse probability weights for the ATT that are finally employed in the standard TWFE specification with covariates and their interactions with the time and treatment group dummy. In this way, the weighting scheme aims at balancing the distribution of the covariates among the four different groups. We propose also alternative versions of the estimator that employ machine learning methods to estimate the propensity score. This method just mitigates the bias of the TWFE in case of heterogeneous effects, but it offers an estimator that balances the changes in the distribution of the covariate over time and has proficient performances in real-world scenarios.

To corroborate our findings, we conduct a series of Monte Carlo simulations in order to investigate the finite sample properties of all semi-parametric estimators presented above. The simulations are designed for overcoming the typical problems encountered when estimating the causal effect of interest in a DiD setting, such as covariate imbalances between treated and controls groups and changes in their distribution over time. The different methodologies are overall tested across three different experimental settings in a repeated cross-sections scenario in order to provide practical guidance for practitioners.

Our analysis confirms that the commonly-used standard TWFE may be severely biased under recurrent settings. Typically, TWFE tends to be outperformed by other semi-parametric methods, in particular by DRDiD and 3IPWRA, which show limited biased with respect to the IPW and OR approaches.

When both the propensity score and outcome models are correctly specified, the class of DRDID estimators proposed by Sant'Anna and Zhao (2020) tend to have a lower bias than alternative estimators, surprisingly even in the case the distribution of the covariates varies over time, a condition that violates its assumptions. However, in the more realistic scenario when the propensity score and outcome models are misspecified, the class of 3IPWRA estimators outperforms DRDiD in presence of compositional changes. When analyzing the machine learning versions of such classes of estimators, the lasso version of the DRDID estimator has a consistently lower bias than its original version over all three experiments when the outcome and the propensity score models are both misspecified, while evidence for the LASSO 3IPWRA and random forest versions is more mixed. Overall, the simulations still provide evidence of the potential benefits of machine learning methods over traditional techniques.

Based on our Monte Carlo simulations' results, we propose the strategy in empirical DiD studies of the combination of the different versions of the DRDiD and the 3IPWRA, which proved the best performing estimators in our simulation. Such a comparison may be particularly indicative since the two classes of estimators rely on different sets of assumptions. While the first is not built to handle time-varying covariates, the second just partially mitigates the bias with respect to TWFE in case of heterogeneous effects. In addition, 3IPWRA requires particular care in order not to include bad controls (as in TWFE corrections). Therefore, in case of convergence of DRDiD and 3IPWRA results, this may constitute a strong indication in favor of the validity of the hypothesis under study.

These alternative estimation methods are finally applied by reproducing the analysis in Sequeira (2016) who investigate the effect of tariff reduction on corruption behaviors by using bribe payment data on the cargo shipments transiting from South Africa into the ports in Mozambique. By adopting our recommended empirical strategy, we can give strong evidence against the results of the replication in Chang (2020), since our findings show that the effect is close and even lower in magnitude than the traditional TWFE estimation present in the original paper.

The paper is organized as follows: Section 2 presents the baseline features of the DiD, it analyzes its common regression counterpart, also referred to as Two-Way-Fixed-Effects (TWFE), and finally introduces alternative semi-parametric estimators; Section 4 implements a Monte Carlo simulations under different scenarios to test the performance of the various estimators; Section 5 provides an empirical application of the results of the simulations by analyzing the effect of tariff reduction on bribing behavior between South

Africa and Mozambique during the period 2006–2014, as in Sequeira (2016); Section 6 concludes with the most relevant findings.

2 The model

2.1 Notation and Causal Effect

Following Lechner et al. (2011), define the treatment variable D, where d $\in \{0,1\}$,[1] as the binary indicator for whether the individual i belongs to the treated group, where the i subscript is dropped for ease of notation. Starting from the simplest scenario of only two time periods, define T, where t $\in \{0,1\}$, as the binary indicator that takes value zero in the pre-treatment period and one in the post-treatment period. Since the treatment is assumed to take place in between the two periods, every member of the population is untreated in the pre-treatment period. DiD estimates the mean effect of switching D from zero to one on the outcome variable of interest.

We define the potential levels of the outcome variable by using indexes that refer to the potential states of the treatment, so that Y_t^d denotes the outcome that would be realized for a specific value of d in period t. However, for each group and at each period only one of the potential outcomes is observed. The realized outcome is denoted by Y_t (i.e., not indexed by d), and the observable covariates by X. Initially, we assume the covariates X do not vary over time but later on we analyze the implications of relaxing such assumption. The object we are interested in estimating is the average effect on the treated (ATT), which is defined as follows:

$$\begin{aligned} ATT_t &= E(Y_t^1 - Y_t^0 | D = 1) \qquad (1) \\ &= E[E(Y_t^1 - Y_t^0 | X = x, D = 1) | D = 1] \\ &= E_{X|D=1}\delta_t(x) \end{aligned}$$

where $\delta_t(x)$, denoted as $E(Y_t^1 - Y_t^0 | X = x, D = 1)$, represents the causal effect in the respective subpopulations where X takes value x.[2]

[1] Capital letters denote random variables while small letters denote specific realizations or values of such variables.

[2] While usually another parameter on interest is the average treatment effect on the entire population (ATE), computing such a parameter requires additional assumptions that are unlikely to hold in this context and therefore the DiD setting usually focuses on the estimation of the ATT.

2.2 Assumptions

A typical database suitable for DiD analysis should contain information on both periods, i.e. before and after treatment. Our discussion will mainly focus on repeated cross-sections, but it can be easily extended to panel data with some minor adjustments. The identification of a causal effect under DiD relies on a set of assumptions, that we briefly discuss here.

2.2.1 Stable unit treatment values assumption

The first hypothesis, the so-called Stable Unit Treatment Value assumption (SUTVA) as in Rubin (1977), requires that the potential outcomes are not affected by the particular treatment assignment to the other units. As a consequence, only one of either the treated or the untreated potential outcome is observable for every member of the population at a specific point in time and the observed outcome is therefore defined as:

$$Y_t = dY_t^1 + (1-d)Y_t^0 \tag{2}$$

If SUTVA is violated, we observe neither of the two potential outcomes, invalidating the identification of the causal effect.

2.2.2 Exogeneity of the covariates

The second assumption, as standard practice for the identification of causal effects, is the exogeneity of the covariates that rules out reverse causality. Applying the outcome notation to the explanatory variable X^d, the assumption implies:

$$X^1 = X^0 = X, \qquad \forall x \in \chi \tag{3}$$

where χ denotes the subspace of X used in the analysis. Intuitively, this hypothesis excludes that the components of X are influenced by the treatment status. It is worth noting that individuals are forward-looking agents and may alter their behavior according to expectations about the future evolution of some variable. As a result, measuring variables before the treatment does not automatically ensure exogeneity. Indeed, if such anticipatory behavior affects the outcome variable as well, then the assumption of exogeneity may be violated. Variables that are constant over time are exogenous by construction since treatment is time-varying.

2.2.3 No effect on pre-treated

The third assumption is that in the pre-treatment period the treatment has no effect on the pre-treatment population (NEPT):

$$\delta_0(x) = 0, \qquad \forall x \in \chi \tag{4}$$

Note that NEPT also rules out the possibility of an anticipation effect of a future treatment on the pre-treatment outcome for the treated population.

2.2.4 Common trend

The key assumption for the identification of causal effects in the DiD design requires that the differences over time in the expected potential outcomes in absence of treatment are independent of whether an individual belongs to either the treated or the control group. i.e.:

$$\begin{aligned} E(Y_1^0|D=1) - E(Y_0^0|D=1) &= E(Y_1^0|D=0) - E(Y_0^0|D=0) \\ &= E(Y_1^0) - E(Y_0^0) \end{aligned}$$

This is also known as the unconditional parallel trend (UCP) assumption. The hypothesis is essential because the trend of potential nontreatment outcomes for units belonging to the treated group is not known, whereas the path of potential nontreatment outcomes is instead observable in the case of controls.

The parallel trends assumption is typically more plausible after conditioning on a set of observed covariates X, i.e.:

$$\begin{aligned}E(Y_1^0|X = x, D = 1) - E(Y_0^0|X = x, D = 1)\\ = E(Y_1^0|X = x, D = 0) - E(Y_0^0|X = x, D = 0)\\ = E(Y_1^0|X = x) - E(Y_0^0|X = x)\end{aligned} \tag{5}$$

The conditional parallel trend assumption (CPT) implies that if the treated group had not been subject to the treatment, it would have evolved, conditional on X, following the same trend observed in the control group. Therefore, the inclusion of the covariates X as controls is aimed at capturing all variables that may cause different time trends.

2.2.5 Common support

The conditional parallel trend assumption implies that it is necessary that observations with characteristics x exist for all four sub-samples determined by the treatment status and time dummy. This is guaranteed by the so-called common support (CS) assumption:

$$P[D|X] < 1 - \epsilon \quad and \quad P[D] > 0 \tag{6}$$

for some $\epsilon > 0$. In other words, the conditional probability of belonging to the treatment group given X is uniformly bounded away from one, imposing that for every value of the covariates X there is at least a small chance that the unit is not treated, and in addition the proportion of treated units is bounded away from zero, meaning that at least a small fraction of the population is treated. The common support assumption, in contrast to the previous ones, refers to observable quantities and is therefore testable. In the case common support is not verified for all values of X, researchers usually restrict the definition of average treatment effect on the treated units where $X(\chi)$ is observable in all four sub-populations.

2.3 DiD Regression: Two-Way-Fixed Effect

DiD regression, usually referred to as Two-Way-Fixed Effect (TWFE), assumes an additive linear structure for potential outcomes. The implicit assumption is that the conditional expectation function (CEF) is such that:

$$
\begin{aligned}
&E(Y_0^0|D=0) = \alpha \\
&E(Y_1^0|D=0) = \alpha + \gamma \\
&E(Y_0^1|D=1) = \alpha + \beta \\
&E(Y_1^1|D=1) = \alpha + \gamma + \beta + \delta
\end{aligned}
$$

where α represents the expected value of the outcome for the control sub-population at the pre-treatment period, γ is the constant time effect between t=0 and t=1, β represents the treatment-group fixed effect, namely the differential in the potential outcome between treated and controls in both periods t=0 and t=1, and δ represents the effect of the treatment. Under these assumptions, DiD identifies the ATT:

$$
\begin{aligned}
ATT &= E(Y_1|D=1) - E(Y_0|D=1) - E(Y_1|D=0) + E(Y_0|D=0) \\
&= (\alpha + \gamma + \beta + \delta) - (\alpha + \beta) - (\alpha + \gamma) + (\alpha) \\
&= \delta
\end{aligned}
$$

and the causal effect might also be estimated by means of a regression.

Typically, the TWFE, in case of no covariates, takes the following form:

$$Y_i = \alpha + \gamma T_i + \beta D_i + \delta(T_i \cdot D_i) + \epsilon_i \tag{7}$$

where i stands for individual i, $T \in \{0,1\}$ is a time dummy that takes value 0 in the pre-treatment period and 1 in the post, $D \in \{0,1\}$ is the treatment group dummy that has value 1 in case the unit belongs to the treated pool, and their interaction term captures the effect of the treatment. Note that in this simple setting, because the model is saturated, the conditional expectation of potential outcome coincides with the regression equation. As a consequence, in case the unconditional parallel trend hypothesis is verified, regression is unbiased without imposing several additional assumptions.

2.4 Limits and Extensions of the TWFE Model

2.4.1 TWFE With Covariates

In realistic settings, the common trend assumption is likely to hold only after conditioning for a set of covariates. Therefore, in the presence of X-specific trends, the TWFE specification needs to account for the presence of covariates.

Assuming that the CEF of the potential outcome depends linearly on a time-invariant set of covariates $X = (X_1, X_2, ..., X_p)'$ with coefficients $\theta = (\theta_1, \theta_2, ..., \theta_p)'$, then:

$$\begin{aligned}
E(Y_0^0|X, D = 0) &= \alpha + X'\theta \\
E(Y_1^0|X, D = 0) &= \alpha + \gamma + X'\theta \\
E(Y_0^1|X, D = 1) &= \alpha + \beta + X'\theta \\
E(Y_1^1|X, D = 1) &= \alpha + \gamma + \beta + \delta + X'\theta
\end{aligned}$$

We are interested in computing:

$$\delta^{DiD} = E(Y_1^1|X, D = 1) - E(Y_1^0|X, D = 1)$$

Similarly as before, the common trend assumption implies:

$$E(Y_1^0 - Y_0^0|X, D = 1) = E(Y_1^0 - Y_0^0|X, D = 0)$$

Rearranging:

$$\begin{aligned}
E(Y_1^0|X, D = 1) &= E(Y_1^0|X, D = 0) - E(Y_0^0|X, D = 0) + E(Y_0^1|X, D = 1) \\
&= \alpha + \gamma + \beta + X'\theta
\end{aligned}$$

Therefore:

$$\begin{aligned}\delta^{DiD} &= E(Y_1^1|X, D=1) - E(Y_1^0|X, D=1)\\ &= (\alpha+\gamma+\beta+\delta+X'\theta) - (\alpha+\gamma+\beta+X'\theta)\\ &= \delta\end{aligned}$$

In this case, the regression equation:

$$Y_i = \alpha + \gamma T_i + \beta D_i + \delta(T_i \cdot D_i) + X_i'\theta + \epsilon_i \tag{8}$$

still coincides with the CEF of the potential outcomes. However, it is important to note that the inclusion of the covariate holds only when three extremely restrictive additional assumptions are verified: homogeneous treatment effects in X, a restriction on how the covariates are allowed to vary over time (here we implicitly assumed X to be time-invariant, but next sections will also allow for time-varying covariates) and the additive linear form of how the covariates affect the outcome.

2.4.2 Dealing with X-specific trends

Under many plausible scenarios, the naive inclusion of covariates in the TWFE model may be a source of bias. To show this, define X_t^d as the mean value of X for treatment d at time t. Consider for simplicity just one covariate, then:

$$\begin{aligned}E(Y_0^0|X, D=0) &= \alpha_0 + \theta_0 X_0^0\\ E(Y_1^0|X, D=0) &= \alpha + \gamma + \theta_1 X_1^0\\ E(Y_0^0|X, D=1) &= \alpha + \beta + \theta_0 X_0^1\\ E(Y_1^0|X, D=1) &= \alpha + \gamma + \beta + \theta_1 X_1^1\end{aligned}$$

Then, assuming that the conditional parallel trend assumption holds, we have:

$$\begin{aligned}
E(Y_1^0 - Y_0^0|X, D = 1) &= E(Y_1^0 - Y_0^0|X, D = 0) \\
\alpha + \gamma + \beta + \theta_1 X_1^1 - (\alpha + \beta + \theta_0 X_0^1) &= \alpha + \gamma + \theta_1 X_1^0 - (\alpha + \theta_0 X_0^0) \\
(\theta_1 X_1^1 - \theta_0 X_0^1) - (\theta_1 X_1^0 - \theta_0 X_0^0) &= 0 \\
\theta_1 (X_1^1 - X_1^0) - \theta_0 (X_0^1 - X_0^0) &= 0 \qquad (9)
\end{aligned}$$

where in the third passage the right-hand side is subtracted from the left-hand side and the last line rearranges the terms.

This computed quantity may be different from zero under certain circumstances. For instance, assume X is time-invariant. Then, we can write $X_1^1 = X_0^1 \equiv X^1$ and $X_1^0 = X_0^0 \equiv X^0$ and Equation (9) becomes:

$$\theta_1 (X_1^1 - X_1^0) - \theta_0 (X_0^1 - X_0^0) = (\theta_1 - \theta_0) \cdot (X^1 - X^0) \qquad (10)$$

This implies that for a covariate that does not vary over time, TWFE identifies the ATT if either: (1) the means of the covariates are the same across groups or (2) the effects of the covariates on the outcome variable are equal in the pre and post-treatment periods (Zeldow and Hatfield, 2019). Therefore, whenever there are X-specific trends denoted as $\tau(X) = \gamma + \phi X$, this implies that $\theta_1 = \theta_0 + \phi X^1$ and for homogeneous treatment effects γ the $ATT = E(Y_1^1 - Y_1^0|D = 1)$ can be re-written as:

$$\begin{aligned}
ATT &= [\alpha + (\gamma + \phi X^1) + \beta + \delta + \theta_0 X^1 - (\alpha + \beta + \theta_0 X^1)] - [\alpha + (\gamma + \phi X^0) + \theta_0 X_0 - (\alpha + \theta_0 X^0)] \\
&= \delta + \phi (X^1 - X^0)
\end{aligned}$$

Thus, when $\phi \neq 0$, TWFE identifies the ATT only if $X_1 = X_0$, namely if the covariates X has the same distribution over the treated and the untreated individuals, which is unlikely to hold in non-randomized settings.

Instead, when we allow for time-varying covariates, by replacing Equation (9) and

$\tau(X)$ in the ATT, the following result is obtained:

$$
\begin{aligned}
ATT &= (\alpha + (\gamma + \phi X_1^1) + \beta + \delta + \theta_0 X_1^1 - (\alpha + \beta + \theta_0 X_0^1)) - (\alpha + (\gamma + \phi X_1^0) + \theta_0 X_1^0 - (\alpha + \theta_0 X_0^0)) \\
&= \delta + (\phi + \theta_0)(X_1^1 - X_1^0) - \theta_0 (X_0^1 - X_0^0)
\end{aligned}
$$

Consequently, when allowing for time-varying covariates, two conditions must be satisfied to guarantee that $ATT = \delta$: the relationship between the covariates and the outcome is constant (i.e. $\phi = 0$), and the difference in the mean of the covariates between treated and controls is the same in pre and post-treatment periods (i.e. $X_1^1 - X_1^0 = X_0^1 - X_0^0$) (Zeldow and Hatfield, 2019). Therefore, a time-varying covariate is a confounder if its relationship with the outcome is time-varying or the covariate evolves differently between the treated and control groups.

However, the standard TWFE specification can be improved by allowing some corrections. For example, the interaction terms between covariates and time can be included so that the model can be written as:

$$
Y_i = \alpha + \gamma T_i + \beta D_i + \delta (T_i \cdot D_i) + X_i' \theta + (T_i \cdot X_i')\omega + \epsilon_i \tag{11}
$$

Zeldow and Hatfield (2019) argue that this model version eliminates the bias in presence of homogeneous treatment effects in X, especially when dealing with time-invariant covariates X. If covariates X do not vary over time, they are exogenous to the treatment and therefore there is no risk of conditioning on covariates affected by the treatment. However, as shown in our Monte Carlo simulations in Section 4, the correction only partially works in case of time-varying covariates.

We therefore consider another possible correction that includes the interaction between the covariates and the treatment group dummy:

$$
Y_i = \alpha + \gamma T_i + \beta D_i + \delta (T_i \cdot D_i) + X_i' \theta + (T_i \cdot X_i')\omega + (D_i \cdot X_i')\mu + \epsilon_i \tag{12}
$$

This specification, by controlling for both the time and treatment group heterogeneity of the covariates, removes the trend also when dealing with time-varying covariates under homogeneous treatment effects. However, when covariates are allowed to vary over time, the correction is subject to the risk of conditioning on covariates affected by the treatment, namely bad controls. The performance of the TWFE and its corrections are therefore tested in our Monte Carlo simulations in Section 4.

2.4.3 Heterogeneous effects

In most realistic settings, the effect of the treatment is likely to vary for different values of the covariates X. However, TWFE and its correction implicitly assume homogeneous treatment effects in X and therefore, when this additional restriction is not satisfied, the estimated causal parameter may differ from the true ATT (Meyer, 1995; Abadie, 2005; Sant'Anna and Zhao, 2020; Roth et al., 2022). For instance, let the treatment effect be heterogeneous in X, as in Cunningham (2021), namely redefining the potential outcomes for the treated in the post period as $E(Y_1^1|X, D = 1) = \alpha + \gamma + \beta + (\delta + \rho X_1^1) + \theta X_1^1$. Then, even assuming time-invariant covariates and $\theta_1 = \theta_0$ we have:

$$\begin{aligned} ATT &= \delta + [(\theta + \rho)X^1 - \theta X^1] - \theta(X^0 - X^0) \\ &= \delta + \rho X_1 \end{aligned}$$

Therefore, whenever $\rho \neq 0$ and thus the treatment is heterogeneous in X, the regression estimate does not identify the true ATT, even when covariates are restricted to be time-invariant.

2.4.4 Non-additive linear form of the CEF for the covariates

Since in most settings it is not possible to use a fully saturated model in X, TWFE assumes a CEF that is a linear function of X, so that the regression equation might differ from the true CEF. Indeed, the linear specification for the control variables implies that the assumption of common trends is conditional on the linear index $X'\theta$ which is more restrictive than assuming common trends conditional on X. For example, if the vector X affects the potential outcome introducing non-linearities, then the potential outcome is:

$$
\begin{aligned}
E(Y_t^d|X) &= f(\alpha + \gamma T + \beta D + \delta TD + \theta X) \\
&\neq \alpha + \gamma T + \beta D + \delta TD + \theta X
\end{aligned}
$$

and yields biased estimates since it assumes a misspecified model that does not capture non-linearities.

In order to overcome the issues discussed above, a number of alternative semi-parametric estimators have been suggested by the literature as extensions to the DiD approach. These will be covered in the next section.

3 Alternative Semi-Parametric Estimators

3.1 Main Estimators Proposed by the Literature

As summarized in previous sections, all the TWFE limitations above relate to imposing unrealistic restrictions on the CEF. Instead, semi-parametric estimators allow for weaker assumptions on the CEF. Among these estimators, a recent novelty is the doubly robust Difference-in-Difference (DRDiD) (Sant'Anna and Zhao, 2020) is particularly flexible by owning the property of double-robustness: the researcher needs just to correctly specify either the propensity score model or the outcome regression models to retrieve the causal effect. Since DRDiD controls for pre-treatment levels of the covariates, we propose a method based on Blundell and Dias (2009) which deals with time-varying covariates as well, even if this comes at the cost of mitigating, and not fully solving, the heterogeneity in treatment effects present in TWFE and dealing with potential bad controls. The next sections introduce this family of semi-parametric estimators.

3.1.1 Outcome Regression

The outcome regression (OR) approach relies on the researchers' ability to correctly specify a model for the evolution of the outcome of interest. Intuitively, since the ATT under conditional parallel trends requires the computation of four conditional expectations of the outcome variable, OR computes the quantities that are not directly observable by specifying a functional model based on the covariates X. More precisely, the two conditional expectations of the observed outcome in the pre- and post- treatments periods for the treated are directly computed by means of sample averages. The remaining two conditional expectations rely on observations from the controls but refer to the treated sub-population, and must be predicted: an outcome model is estimated among untreated units given their covariate values, and then fitted values are predicted using the empirical distribution of X_i among treated units. Usually, the outcome model is estimated using regression, hence the alternative name of regression adjustment (RA), but other more flexible non-parametric methods can be employed as well, such as nearest neighbor matching, which associates treated with untreated units with close covariate values.

More formally, following Heckman et al. (1997), starting from the definition of the ATT under conditional parallel trends and using the law of iterated expectations, we obtain:

$$\begin{aligned} ATT &= E[E(Y_1 - Y_0|X, D = 1) - E(Y_1 - Y_0|X, D = 0)|D = 1] \\ &= E(Y_1 - Y_0|D = 1) - E[E(Y_1 - Y_0|X, D = 0)|D = 1] \end{aligned} \tag{13}$$

where the first term in Equation (13) can be computed by taking sample averages, while the second expected value must be estimated. One way of estimating it is by fitting a regression on the controls group data and taking predictions based on the empirical distribution of X_i among treated units. More formally:

$$\delta^{OR} = \bar{Y}_{1,1} - \bar{Y}_{1,0} - \left[\frac{1}{n_{treat}} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i)) \right] \tag{14}$$

where $\bar{Y}_{d,t} = \sum_{i|D_i=1} Y_{it}/n_{d,t}$ is the sample average outcome among treated units in treatment group d at time t, and $\hat{\mu}_{d,t}(X)$ is an estimator of the true, unknown $m_{d,t}(x) \equiv E[Y_t|D = d, X = x]$, which is usually estimated by running a regression in the observed control sub-population defined by d and t and obtaining fitted values based on the empirical distribution of X_i among the treated individuals. Intuitively, when using a linear specification for $\hat{\mu}_{d,t}(X)$, the model would be close to the version of TWFE with covariates as in Equation (12) that includes also the interactions between X_i with both treatment group and time dummies. The two models differ because the outcome regression approach re-weights based on the distribution of X_i among units with $D_i = 1$ (Roth et al., 2022). The condition for the consistency of the ATT of the outcome regression is the correct specification of $\hat{\mu}_{d,t}(X)$.

3.1.2 Inverse Probability Weighting

The Inverse Probability Weighting (IPW) approach proposed by Abadie (2005) avoids directly modeling the outcome evolution. Its focus is on the treatment model, namely the conditional probability of being in the treatment group given covariates, $p(X) \equiv P(D = 1|X)$. The idea of the IPW estimator is to adjust for confounding factors by using the propensity score to balance baseline individual characteristics in the treated and untreated groups. When estimating the ATE, the procedure corresponds to weighting each individual in the sample by its inverse probability of receiving the actual treatment. When estimating the ATT, the estimand is multiplied by the propensity score, since it refers to the treated sub-population. Therefore, in the case of panel data and under the standard assumptions expressed in section 2.2, the ATT can be expressed as:

$$\delta^{IPW} = \frac{1}{E(D)} \cdot E\left[\frac{D - p(X)}{1 - p(X)} \cdot (Y_1 - Y_0)\right] \tag{15}$$

Intuitively, IPW produces a weighting scheme that weights-down the distribution of $Y_1 - Y_0$ for the untreated individuals that have covariate values which are over-represented among controls (namely, with low $\frac{p(X)}{1-p(X)}$), and weighting-up $Y_1 - Y_0$ for the individuals with covariate values under-represented among controls (that is, with high $\frac{p(X)}{1-p(X)}$). Consequently, the adjustment balances the distribution of covariates between treated and untreated groups. The IPW is then estimated by using the sample analog:

$$\delta^{IPW} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n}(D_j)} \cdot \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \cdot (Y_{i1} - Y_{i0})\right] \tag{16}$$

where $\hat{\pi}(X)$ is an estimator of the true, unknown propensity score $p(x) = P(D = 1|X)$.

In the case of repeated cross-sections, when different samples are observed over time and thus it is not possible to directly observe the first difference in the observed outcome for each individual, the weighting scheme includes a specific balancing term that represents the proportion of individuals observed at $t = 1$. The ATT is then estimated by:

$$\delta^{IPW} = \frac{1}{E(D) \cdot \lambda} \cdot E\left[\frac{D - p(X)}{1 - p(X)} \cdot \frac{T - \lambda}{1 - \lambda} \cdot Y\right] \tag{17}$$

where λ represents the proportion of individuals at $t = 1$. Thus, the sample analog corresponds to:

$$\delta^{IPW} = \frac{1}{\lambda \cdot \frac{1}{n}\sum_{j=1}^{n}(D_j)} \cdot \sum_{i=1}^{n}\left[\frac{D_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \cdot \frac{T_i - \lambda}{1 - \lambda} \cdot Y_i\right] \tag{18}$$

The unknown propensity score $p(x) = P(D = 1|X)$ is usually estimated by means of logistic regression or a linear probability model, even if non-parametric models can be employed as well. The IPW approach will generally be consistent when the propensity score model is correctly specified.

3.1.3 Doubly Robust Difference-in-Difference

Sant'Anna and Zhao (2020) combine the OR and the IPW approaches into a doubly robust estimand for the ATT. The double robustness property means that if either the propensity score model or the outcome regression models are misspecified (but not both), the resulting estimand still identifies the ATT. Intuitively, the doubly robust Difference-in-Difference (DRDiD) estimator they propose has the advantages of each of the two individual DiD methods and, at the same time, circumvents some of their weaknesses.

Following Sant'Anna and Zhao (2020), we start by introducing the panel data model whose interpretation is more intuitive. Denote $\Delta Y = Y_1 - Y_0$ as the first difference of observed outcomes and $\mu^p_{d,\Delta}(X) = \mu^p_{d,1}(X) - \mu^p_{d,0}(X)$ where $\mu^p_{d,t}(X)$ is a model for the true, unknown, conditional expectation $E[Y_t|D = d, X = x]$ with $d, t \in \{0, 1\}$. Then the DRDiD estimand is:

$$\delta^{dr,p} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{(1-D)p(X)}{1-p(X)}}{E\left[\frac{(1-D)p(X)}{1-p(X)}\right]}\right)\left(\Delta Y - E[Y_1 - Y_0|D=0, X=x]\right)\right] \tag{19}$$

The estimand is made of two components. The first parenthesis is the IPW element of the estimator, namely the weighting scheme. For the treated group, $d = 1$ and therefore $1 - d$ reduces to zero. The weight is therefore $\frac{1}{E(D)}$, where the denominator is there just to guarantee that the weights integrate up to 1. For controls, only $1 - d$ does not reduce to zero and so the numerator displays the typical IPW weights for the ATT in the form of $\frac{p(X)}{1-p(X)}$. Similarly, the denominator has the function to let the weights sum up to one. The second parenthesis contains the outcome regression part of the estimator. $E[Y_1 - Y_0|D = 0, X = x]$ is usually obtained by estimating a linear regression model for the control group and fitting $Y_1 - Y_0$ based on the empirical distribution of X_i among treated individuals. Similarly, the sample analog can be written as:

$$\delta^{dr,p} = \sum_{i=1}^{n}\left[\left(\frac{D_i}{\sum_{j=1}^{n} D_j} - \frac{\frac{(1-D_i)\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}{\sum_{j=1}^{n}\left[\frac{(1-D_j)\hat{\pi}(X_j)}{1-\hat{\pi}(X_j)}\right]}\right)\left(\Delta Y - \hat{\mu}^{p}_{0,\Delta}(X)\right)\right] \tag{20}$$

where $\hat{\pi}(X)$ is the estimation of the true, unknown propensity score $p(X)$.

When dealing with repeated cross-section data, it is not possible to use the first difference of the observed outcomes for each individual because we do not observe the same sample over time. Hence, Sant'Anna and Zhao (2020) propose a modified version of the previous estimand in case of repeated cross-sections. In this case, denote $\mu^{rc}_{d,t}(X)$ as the arbitrary model for the true, unknown conditional expectation function $m^{rc}_{d,t}(x) \equiv E[Y|D = d, T = t, X = x]$, $d, t \in \{0, 1\}$ and for ease of notation, define for $d \in \{0, 1\}$, $\mu^{rc}_{d,Y}(T, X) \equiv T \cdot \mu^{rc}_{d,1}(X) + (1 - T) \cdot \mu^{rc}_{d,0}(X)$ which represents the outcome model for a given treatment group category. Then the DRDiD estimator is defined as:

$$\delta_1^{dr,rc} = E\left[[\omega_1^{rc}(D,T) - \omega_0^{rc}(D,T,X;p)][Y - \mu_{0,Y}^{rc}(T,X)]\right] \tag{21}$$

where:

$$\omega_1^{rc}(D,T) = \omega_{1,1}^{rc}(D,T) - \omega_{1,0}^{rc}(D,T) \tag{22}$$

$$\omega_0^{rc}(D,T,X;p) = \omega_{0,1}^{rc}(D,T,X;p) - \omega_{0,0}^{rc}(D,T,X;p) \tag{23}$$

and for $t \in 0,1$:

$$\omega_{1,t}^{rc}(D,T) = \frac{D \cdot 1\{T=t\}}{E[D \cdot 1\{T=t\}]} \tag{24}$$

$$\omega_{0,t}^{rc}(D,T,X;p) = \frac{(1-D)p(X) \cdot 1\{T=t\}}{1-p(X)} \Bigg/ E\left[\frac{(1-D)p(X) \cdot 1\{T=t\}}{1-p(X)}\right] \tag{25}$$

The relative sample analog is obtained by replacing $p(x)$ with $\hat{\pi}$ and the expectation with sample means. Again, the first term of $\delta_1^{dr,rc}$ represents the IPW weighting scheme based on the estimated propensity score, while the second term represents the outcome regression part of the estimand.

Sant'Anna and Zhao (2020) present a locally semi-parametrically efficient version of the above estimator, characterized by an asymptotic variance that achieves the semi-parametric efficiency bound when the working models for the nuisance functions, namely the propensity score and outcome regression functions, are correctly specified:

$$\begin{aligned}\delta_2^{dr,rc} = \delta_1^{dr,rc} &+ (E[\mu_{1,1}^{rc}(X) - \mu_{0,1}^{rc}(X)|D=1] - E[\mu_{1,1}^{rc}(X) - \mu_{0,1}^{rc}(X)|D=1,T=1]) \\ &- (E[\mu_{1,0}^{rc}(X) - \mu_{0,0}^{rc}(X)|D=1] - E[\mu_{1,0}^{rc}(X) - \mu_{0,0}^{rc}(X)|D=1,T=0])\end{aligned} \tag{26}$$

This is the most efficient version of the DRDiD estimand and the one we are going to evaluate in the continuation of this paper.

The outcome equation and the propensity score can be modeled either parametrically, for instance with a linear and a logistic regression respectively, or non-parametrically. The authors use the inverse probability tilting estimator (Graham et al., 2012) for the treatment model and weighted least-squares for the outcome model. Indeed, in the Monte Carlo simulations discussed in Sant'Anna and Zhao (2020), these two estimation methods for the nuisance functions outperform the traditional parametric models. For this reason, the locally-efficient version of the DRDiD is the one used in section 4 in our the Monte Carlo simulations. In addition, we present a modified version of the model, using lasso and random forest in the intermediate steps, whose performance will be tested as well. DRDiD will generally be consistent if either the propensity score or the outcome model is correctly specified.

In the next section we propose a set of novel semi-parametric extensions of the models discussed above, based on previous work by Sant'Anna and Zhao (2020) and Blundell et al. (2004) that will be evaluated through Monte Carlo simulations in section 4.

3.2 Novel Proposed Semi-Parametric Extensions of the TWFE model

3.2.1 Machine Learning DRDiD

We propose an alternative version of the DRDiD estimator in Sant'Anna and Zhao (2020) that employs machine learning algorithms for the first-stage estimates. Since machine learning methods optimize prediction, they are aimed to minimize the out of sample mean square error (MSE), finding a balance between bias and variance. This characteristic makes machine learning estimators not directly applicable to causal inference, where the aim is to obtain unbiased estimates of the causal parameter of interest. However, Chernozhukov et al. (2018) studied a rather flexible approach to employ the potential of machine learning in the field of causal inference. The idea is that in many econometric settings there are intermediate parts of the estimation process that focus on predicting values that are not readily available to the researchers. Chernozhukov et al. (2018) found that, when three main conditions are met, first-stage estimates can be obtained through machine learning predictors without creating bias in the final estimates of the causal parameter. We provide a brief discussion of these aspects in the Appendix, together with a description of the methods used to build the machine learning based DiD estimators proposed in this paper.

Indeed, the score function of the DRDiD satisfies the Neyman orthogonality condition defined in Equation (40), which is a key building block for the debiased machine learning literature (Chernozhukov et al., 2018). This enables the use of different machine learning methods in this context, such as lasso, ridge, random forests, deep neural nets, boosted trees and various hybrids and ensembles of these methods. In this paper we discuss and evaluate two DRDiD machine learning extensions of the DRDiD model. The first employs lasso in both the estimation of the propensity score and the outcome model, while the second utilizes random forest for both nuisance parameters. The advantage of machine learning over traditional estimation methods is that they do not assume parametrically the functional form of the model under study, avoiding the risk that the researcher misspecify it.

3.2.2 Triple Inverse Probability Weighting Regression Adjusted Estimator

Building on Blundell et al. (2004) and Blundell and Dias (2009), we propose an additional estimator for repeated cross-sections.

In the repeated cross-sections setting, treated and controls groups in the pre-treatment period are more likely to have structural differences with their respective group in the post-treatment period since we do not observe the same individuals over time. Indeed, Hong (2013) warns of the risks for identification under compositional changes in X over time. The authors show that in this scenario, matching on the standard definition of propensity score $P(D|X) = 1$ would lead to biased estimates since it is not equivalent to matching on the set of covariates X. As a consequence, Blundell and Dias (2009) suggests that, in the context of matching, one way of achieving balance in the distribution of the relevant observable characteristics among the four cells defined by eligibility and time is to extend the standard definition of propensity score by denoting three propensity scores that match the treated group in the post-treatment period with each of the other three remaining groups, i.e. controls in pre and post-treatment periods and treated before treatment.

Under this approach, that we call triple inverse probability weighting regression adjusted (3IPWRA), the propensity score is computed as the probability of belonging to the treated group in the post-treatment period. More precisely, the initial sample is split according to the four groups defined by the interaction of time and treatment, and the propensity score is separately computed in the three sub-samples obtained by merging the treated group in the post-treatment period with each one of the three remaining sub-populations one at a time. In this way, a specific propensity score for each of the four groups is defined. The propensity score is not used for matching but for inverse probability weighting. In fact, 3IPWRA uses the estimated propensity score to calculate the Horvitz-Thompson inverse probability weights for the ATT that are finally employed in the standard TWFE specification with covariates and their interactions with the time and treatment group dummy as in Equation (12). In this way, the weighting scheme aims at balancing the distribution of the covariates among the four different groups.

The weighted least square estimation corresponds to:

$$\underset{\phi}{argmin} \underbrace{\left(D_i T_i - \frac{(1 - D_i T_i)\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right)}_{= w_i} (Y_i - \bar{X}_i{}' \phi)^2 \tag{27}$$

where w_i represents the weighting scheme based on the propensity score $\hat{\pi}(X_i)$ which is estimated $\forall i \in (t, d) = \{(0,1), (0,0), (1,0)\}$ by merging the treatment group in the post-treatment period, where $(t, d) = (1, 1)$, with the group (t, d) where i belongs and then computing propensity scores. When $(t, d) = (1, 1)$ the weights are equal to one, while in all other three sub-populations the weights correspond to $\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}$. The notation $\bar{X}_i$ refers to the full vector of controls such as in Equation (12), namely T, D, TD, X', TX', DX'. Such weighted regression approximates the more general semi-parametric version of the IPWRA estimator (Imbens, 2004). Both logistic and machine learning estimators for the propensity score will be analyzed and discussed in the simulations of Section 4.

3.2.3 Triple Weighting Doubly Robust Difference-in-Difference

Our third proposed estimator is the triple weighting doubly robust Difference-in-Difference (3WDRDiD) estimator that adapts the weighting scheme used in the DRDiD (Sant'Anna and Zhao, 2020) to handle the weights employed in the 3IPWRA estimator.

In the context of repeated cross-sections data, define again $m^{rc}_{d,t}(x) \equiv E[Y|D = d, T = t, X = x]$, $d, t \in \{0, 1\}$ and for $d \in \{0, 1\}$, $\mu^{rc}_{d,Y}(T, X) \equiv T \cdot \mu^{rc}_{d,1}(X) + (1 - T) \cdot \mu^{rc}_{d,0}(X)$

and $\mu^{rc}_{d,\Delta}(X) \equiv \mu^{rc}_{d,1}(X) - \mu^{rc}_{d,0}(X)$. Denote $p(X) = P(DT = 1|X)$, i.e. the probability to belong to the treated group in the post-treatment period, then the DRDiD estimator is defined as:

$$\delta_1^{dr,rc} = E[(\omega_1^{rc}(D) - \omega_0^{rc}(D,T,X;p))(Y - \mu^{rc}_{0,Y}(T,X))] \tag{28}$$

where:

$$E[(\omega_1^{rc}(D)] = E[(\omega_{1,1}^{rc}(D)] - E[(\omega_{1,0}^{rc}(D)] \tag{29}$$

$$\omega_0^{rc}(D,T,X;p) = \omega_{0,1}^{rc}(D,T,X;p) - \omega_{0,0}^{rc}(D,T,X;p) \tag{30}$$

and for $t \in 0, 1$:

$$E[(\omega_{1,1}^{rc}(D,T)] = \frac{D \cdot 1\{T = t\}}{E[D \cdot 1\{T = t\}]} \tag{31}$$

$$E[(\omega_{1,0}^{rc}(D,T)] = \frac{(1-D)p(X) \cdot 1\{T = t\}}{1 - p(X)} \Bigg/ E\left[\frac{(1-D)p(X) \cdot 1\{T = t\}}{1 - p(X)}\right] \tag{32}$$

$$E[(\omega_{0,t}^{rc}(D,T,X;p)] = \frac{(1-D)p(X) \cdot 1\{T = t\}}{1 - p(X)} \Bigg/ E\left[\frac{(1-D)p(X) \cdot 1\{T = t\}}{1 - p(X)}\right] \tag{33}$$

Compared to the original estimator proposed by Sant'Anna and Zhao (2020), here the main difference is the use of the "triple-matching weights", which are the propensity scores computed by separately "matching" the treated with each of the three remaining groups, one at a time, as in 3IPWRA. To allow for this weighting scheme, the weights for the treated group in the pre-treatment period are not equal to one, as in Equation (24), but follow the adjustment employed for two untreated sub-populations as in Equation (32), since the idea is to adjust these three groups for both the heterogeneity in time and treatment.

4 Monte Carlo Simulations

The semi-parametric estimators presented above are designed for overcoming the typical problems encountered when estimating the causal effect of interest in a DiD setting. In this section, we conduct a series of Monte Carlo simulations in order to investigate the finite sample properties of the proposed estimators in a repeated cross-sections scenario. The different methodologies are tested across three different experimental settings. Each design is characterized by two repeated cross-sections, one observed at $t = 0$ and another at $t = 1$, with a total sample size of $n = 1000$ observations. The Monte Carlo simulation consists of 500 randomly generated datasets and estimation results are stored at each repetition.

Each of the three experiments is characterized by X-specific trends, but they refer to three distinct scenarios, as summarized in Table 1. Experiment 0 assumes a randomized selection into treatment, time-invariant covariates, and homogeneous treatment effects in X, in addition to X-specific trends. Following Section 2.4, in this setting we expect all estimation methods to perform relatively well, including the TWFE, since its assumptions are satisfied by construction. Experiment 1 allows instead for non-random selection into treatment, similarly to Experiment 2. However, in addition, the latter relaxes both the assumption of absence of compositional changes in the covariates between the pre and post-treatment periods and the homogeneity of treatment effects. For this reason, Experiment 2 reproduces the most realistic setting where heterogeneity is imposed along different dimensions of the data generating process (DGP, henceforth).

The choice of the functional forms of our DGPs, presented below, is aimed at preserving the comparibility with the work of Sant'Anna and Zhao (2020) and Kang and Schafer (2007), which employed the same functional specifications. Indeed, Experiment 1 closely reproduces the Monte Carlo simulations in Sant'Anna and Zhao (2020), while Experiment 0 and 2, building on the same general framework, test the specified alternative conditions.

Overall, each of our three experiments considers four different DGPs. The aim is to assess whether the researcher can correctly specify the propensity score and the outcome models in different scenarios. First of all, for a generic variable $W = (W_1, W_2, W_3, W_4)'$, we define the underlying true outcome and propensity score model as:

$$f_{reg}(W) = 210 + 25.4 \cdot W_1 + 13.7 \cdot (W_2 + W_3 + W_4) \tag{34}$$

$$f_{ps}(W) = 0.75 \cdot (-W_1 + 0.5 \cdot W_2 - 0.25 \cdot -0.1 \cdot W_4) \tag{35}$$

The function $f_{ps}(W)$, which determines selection into treament, is modelled through the inverse of the logit function, i.e. $expit(f_{ps}(W)) = \frac{exp(f_{ps}(W))}{1+exp(f_{ps}(W))}$, which has the desirable property of producing an average propensity score of 0.5. In other words, assuming parametrically a logit model for the propensity score (with all the relevant covariates) will lead to a correct estimation of the probability of being treated, by construction.

In the context of each of the three experiments, the baseline function for the outcome $f_{reg}(W)$ produces a mean of $E(Y) = E[f_{reg}(W)] = 210.0$ and, when combined with $f_{ps}(W)$, leads to $E(Y|D=0) = 200.0$ and $E(Y|D=1) = 220.0$. As outlined in Kang and Schafer (2007), the selection bias in this DGP is not severe because the difference between the average outcome of the treated units and the average outcome of the full population is only a one-quarter of a population standard deviation. Nevertheless, this difference is large enough to invalidate the performance of naive estimators.

The generic vector W can either represent vector Z, which is the set of variables observed by the researcher, or vector X, which is not observable. The idea is that the unique generic DGP (expressed in terms of W) leads, for each experiment, to four cases depending on whether W is replaced by the observed vector Z or by the unobservable vector X. When the modelling functions are defined as $f_{ps}(Z)$ and $f_{reg}(Z)$, i.e. they are functions of the observed Z, then both the propensity score and outcome regression models will be correctly specified since the variables we observe coincide with those affecting the outcome. We call this scenario DGP A. When the data are generated by $f_{ps}(X)$ and $f_{reg}(X)$, then the researcher, who has only access to Z, will misspecify both models. We call this scenario DGP D. Typically, such a scenario is the most realistic since researchers do not know have an *a priori* knowledge of the phenomenon under analysis.

We also consider the two cases in which just one of the two models is correctly specified. We call this scenarios DGP B (when the outcome model is correctly specified) and DGP C (when only the propensity score model is correctly specified). This is achieved by assuming that Z is a highly non-linear transformation of X and its interactions. Assume $X = (X_1, X_2, X_3, X_4)'$ is distributed as $N(0, I_4)$ with I_4 representing the 4×4 identity matrix. For $j = 1, 2, 3, 4$ define the standardized variable $Z_j = (\tilde{Z}_j - E[\tilde{Z}_j])/\sqrt{Var(\tilde{Z}_j)}$ where $\tilde{Z}_1 = \exp(0.5X_1)$, $\tilde{Z}_2 = 10 + X_2/(1 + \exp(X_1))$, $\tilde{Z}_3 = (0.6 + X_1X_2/25)^3$, and $\tilde{Z}_4 = (20 + X_2 + X_4)^2$. Note that the choice of the non-linear transformation that relates the individual variables of Z and X includes a wide range of functional forms, such as quadratic, cubic and exponential. In addition, such transformations include interactions of the Xs in order to achieve additional complexity in the relationship that links Z and X. Therefore, when the propensity score and outcome regression models are misspecified, i.e. when the DGPs are built from X whereas we observe only Z, the fact of not observing the actual true regressors is likely to cause a bias in the estimation. In this case, allowing for non-parametric first-stage estimates, which may better capture the non-linearities between Z and X, can minimize the bias. For example, random forest will build a non-parametric relation between Z and the outcome, approximating more closely the relation between the true X (which is a non-linear transformation of Z) and the outcome rather than simply assuming a linear relationship that may not hold in the data. Similarly, lasso allows for more flexibility with respect to traditional methods, as explained in the Appendix.

Table 2 summarizes the different estimation methods that are tested in each experiment. They are evaluated in terms of average bias, root mean square error (RMSE), variance, and computational time required for the estimation. When not otherwise specified, all estimators employ a logit model for the propensity score and a linear regression model for the outcome. Therefore, the first is estimated using maximum likelihood and the second by ordinary least squares. Note that the choice of using a logit model for the propensity score is required to perfectly match, by construction, the functional form of the probability of being treated in our DGPs. When the DGPs are built from $f_{ps}(Z)$ and $f_{reg}(Z)$, the models are therefore correctly specified.[3]

For the DRDiD and 3IPWRA we also allow for the possibility of first-stage estimates using machine learning methods. Such non-parametric methods should better capture the non-linearities under investigation when the working models are misspecified. When lasso is used, the outcome and the treatment model are designed as a penalized linear and a penalized logistic regression, respectively. Lasso is performed in R using the 'glmnet' package (Friedman et al., 2010) and the shrinkage parameter λ is selected through 10-fold cross-validation and represents the largest value of λ whose cross-validation error is

[3] This would not be true if, for example, a linear probability model would be used for the propensity score. That model has also the issue of fitting values outside the range 0-1.

within 1 standard deviation from the minimum. This allows to select the sparsest model with performances approximately equal to the optimum.

Since lasso implicitly performs variable selection and can therefore handle a large set of covariates, in the simulations we allow lasso to employ an expanded set of covariates that include all third order terms and interactions of the original variables. As a result, lasso performs a selection from a wider set of variables and may more precisely capture the non-linearities in the functional forms related to the phenomenon under study. Since functional forms are unlikely to be known by the researcher, lasso represents a more flexible estimation alternative. On the contrary, in real data analysis, when employing traditional estimators, the researcher may be constrained by the risk of including a number of predictors p that is too large. In some extreme cases this number may be close or even higher than the number of observations n, invalidating the estimation. Despite being technically possible in our synthetic dataset to include all interactions terms for the traditional estimators as well (since we have just four regressors, the expanded set of covariates $p = 34$ would be still lower than n), we limit the inclusion of these higher order terms and interactions only to the lasso model to emulate the more recurrent real-word scenario where it is not possible for traditional estimators to do so.

When random forest is used, the estimation is implemented using the 'cforest' R package (Hothorn et al., 2006; Strobl et al., 2007, 2008). The number of trees is set to 100, as suggested by Oshiro et al. (2012), in order to obtain a good balance between accuracy and computational effort. At each node, as common practice, the number of randomly sampled input variables is restricted to $\sqrt{p}$, where p is again the number of predictors (James et al., 2013).

When using machine learning tools, we do not perform sample splitting, even if generally suggested in Chernozhukov et al. (2018) and Bach et al. (2021). In fact, as noted in Farrell et al. (2021), 'sample splitting may be used to obtain valid inference in cases where the parameter of interest itself is learned from the data'. For the causal estimands in the DRDiD and 3IPWRA, the regression functions and propensity score must be estimated, but these are first-stage estimates. The second stage of the DRDiD and 3IPWRA are not estimated by means of method of moments, but they are just the expected value of the estimand of the ATT. Therefore, in this case sample splitting does not provide any advantage. Such a result is also confirmed by additional Monte Carlo simulations provided in the appendix. Finally, it is also important to note that, when using sample splitting, the computational time required is increased by a factor of k in case of k-fold sample splitting. If the standard errors are calculated by bootstrap, this would lead to an increase of computational time in the order of $k \times n$ where n is the number of the repetitions in the bootstrap.

In what follows we discuss each experiment under different assumptions for the DPG.

4.1 Experiment 0: X-Specific Trends and Randomized Selection

Our first experiment is characterized by randomized selection into treatment, with X-specific trends, time-invariant covariates, and homogeneous treatment effects in X. Since in this case the selection of treated individuals is assumed to be random, DGP.A coincides with DGP.B and DGP.C is equivalent to DGP.D because the true propensity score is a constant and there is no need to specify a correct propensity score model. For this reason, in the case of Experiment 0 we use the notation DGP.AB and DGP.CD in order to maintain a certain degree of consistency with the notation used in the following experiments below.

The DGPs are therefore specified as in Table 3, where ϵ_0, $\epsilon_1(d)$, $d = 0, 1$, are independent standard normal random variables representing the stochastic error term of the potential outcomes, p is the constant probability of being treated, λ is the proportion of treated units and U_d and U_t are independent standard uniform stochastic variables used to randomly select individuals into treatment and post-treatment period, respectively. For a generic variable W, $\upsilon(W, D)$ is an independent normal random variable with mean $D \cdot f_{reg}(W)$ and unit variance which represents the time-invariant unobserved group heterogeneity between treated and untreated populations. The trend is specified as $\tau(W) = f_{reg}(W)$, and therefore in the post-treatment period $T = 1$ it sums to the standard function of the outcome model f_{reg}. This explains the presence of the factor 2 that multiplies the term $f_{reg}(W)$ in the formula of the potential outcome Y_1^d.

Note that at $t = 0$ only the potential outcome of no treatment Y_0^0 exists for both treated and controls, while at $t = 1$ the possible potential outcomes are two, so that in Table 3 we adopt the more concise notation Y_1^d. The available data to the researcher are $\{Y_0, D, Z\}$ if $T = 0$ and $\{Y_1, D, Z\}$ when $T = 1$, where $Y_0 = Y_0^0$ and $Y_1 = DY_1^1 + (1-D)Y_1^0$. In the aforementioned DGPs, the true ATT is zero.

It is important to note that the trend here is a function that depends on the covariates. As a consequence, the unconditional parallel trend does not hold and a correct inclusion of the covariates is required to satisfy conditional parallel trends. As a consequence, in this setting, a TWFE specification without covariates would be biased. Since the simulation replicates a randomized experiment, and therefore the mean of the distribution of the covariates is the same among treated and controls (see Figure 1), we expect that the inclusion of time-invariant covariates affecting the trend would eliminate the bias. The results of the simulations are displayed in Table 4 and Table 5.

In line with our expectations, when the outcome regression model is correctly specified, all estimators, including the standard specification of the TWFE with covariates, perform well and are approximately unbiased. In this case, the most efficient estimators are the doubly robust DRDiD estimator of Sant'Anna and Zhao (2020) ($RMSE = 0.184$), the standard and lasso versions of 3IPWRA ($RMSE = 0.185$ and $RMSE = 0.186$, respectively) and the TWFE that includes both time and treatment group interactions with the covariates ($RMSE = 0.184$). When instead the outcome model is incorrectly specified, a small degree of bias is present for all estimates. Overall, all estimators perform relatively well, and methods that employs machine-learning first stage estimates tend to be more efficient, being characterized by lower variance, since they are better at capturing the non-linearities needed to reproduce the true DGP. The lowest bias (-0.074) is displayed by the IPW estimator of Abadie (2005), but the DRDiD and 3IPWRA have an overall better performance in terms of RMSE. The lowest RMSE is the one attached to the random forest specification of the 3IPWRA ($RMSE = 2.875$).

4.2 Experiment 1: X-specific Trends and Non-Randomized Selection

Experiment 1 closely replicates the simulation presented by Sant'Anna and Zhao (2020). Besides X-specific trends, here the selection into treatment is not randomized, causing additional obstacles to the identification of the causal parameter. In a non-randomized experiment, selection into treatment may be associated to some individual characteristics X and is therefore likely to cause heterogeneity in the distribution of the covariates between treated and controls. In addition, in Experiment 1 covariates are assumed to be time-invariant, so that compositional changes in the independent variables are ruled out, and treatment effects are homogeneous in X. In a non-randomized experiment, a propensity score model can be usefully employed for the estimation of the causal parameter and therefore four different DGPs are specified as in Table 6, where the notation closely follows the one in Experiment 0 (see Section 4.1).

The major difference with respect to the previous experiment is that now selection into treatment depends on a propensity score which is specified as a logistic transformation of the generic function $f_{ps}(W)$, where W can be either Z or X depending on whether our model is correctly specified or not. When the selection into treatment is driven by the propensity score, treatment and control groups are generally heterogeneous in terms of covariate distribution. Figure 2 shows the distribution of X_2 between the two groups in the pre- and post-treatment periods (plots with similar characteristics can be shown for the remaining three regressors X_1, X_3, and X_4). Within each treatment group category there are no significant changes between $t = 0$ and $t = 1$ because the covariates are time-invariant. However, the plot displays heterogeneity in the distribution of the covariates among treated and controls. Note that this heterogeneity is not severe, since, as explained in Kang and Schafer (2007), selection into treatment leads to $E(Y|D = 0) = 200.0$ and $E(Y|D = 1) = 220.0$ (with respect to a population average of $E(Y) = 210.0$). The difference between the mean of the treated and the mean of the full population is thus only one-quarter of a population standard deviation. If an estimator performs poorly in such a scenario, it may even be more biased when treated and control groups are more heterogeneous in terms of covariate distribution.

Therefore, if traditional estimators fail under a moderate selection bias like this one, such a result streghtens the need of adopting alternative more flexible estimation techniques. The results of the simulation can be found in Table 7, Table 8, Table 9 and Table 10.

In Experiment 1, the weaknesses of the TWFE specification with covariates are noticeable. Independently of the correct or incorrect specification of the propensity score and outcome regression models, TWFE is typically severely biased, with a bias of -20.686 even in the most favourable scenario embodied by DGP A. However, as outlined by Zeldow and Hatfield (2019), when the conditional independence assumption is satisfied by observing time-invariant covariates, the TWFE model can be corrected by including the interactions between these stationary covariates and the time dummy. In Experiment 1A, where the propensity score and the outcome regression are correctly specified, this correction works properly. However, its performance, despite offering a major improvement, gradually worsens when the models are not correctly specified. In such a scenario, alternative semi-parametric estimators achieve better results. In Exp.1D, the random forest version of the 3IPWRA has the lowest bias (-1.470) and RMSE (3.333), followed by the lasso specification of the DRDiD with -1.720 and 4.116. Overall the different version of the 3IPWRA and DRDiD tend to outperform other estimators, in particular the IPW and RA which are typically less efficient.

4.3 Experiment 2: X-specific Trends and Non-Randomized Selection under Compositional Changes

Experiment 2 tests the different estimators when, in addition to a X-specific trend and non-randomized selection, there are compositional changes in the distribution of the covariates between the pre and post-treatment periods. In addition, in this design the treatment effects are allowed to vary for different values of X. As discussed in Section 2.4, such setting is a real threat to identification. Indeed, the inclusion of time-varying covariates may cause a bias in the TWFE estimates if either (i) the variation in X between time periods is not the same for treated and control units, or (ii) the effect of the covariates on the outcome varies over time. As a consequence, time-varying covariates cannot be used to satisfy the conditional parallel trend assumption, since in presence of a X-specific trend, the effect of the covariate that determines the trend is time-varying, causing a bias. Similarly, allowing for heterogeneous treatment effects in X in the DGP causes a bias as well when using the traditional TWFE estimator, as discussed in section 2.4.

Table 11 describes the four DGPs In Experiment 2. In this design, the DGPs are subject to two main changes. The first is the creation of a selection parameter through time: individuals that have certain characteristics X are more likely to be observed in the post-treatment period. In practice, we defined a 'propensity score' which does not refer to the treatment variable D as usual, but to the time variable T. As a consequence, each observation has an individual probability $\lambda(W)$ of belonging to $t = 1$ which is based on the value of its covariates W modeled through the function $f_{ps}(W)$. On the contrary, in Experiment 0 and Experiment 1, all individuals were equally likely (with a probability $\lambda = 0.5$) to be observed at $t = 1$. Such a new design of Experiment 2 causes the distribution of the covariates to vary over time for both treated and controls. Using a more technical terminology, we allow for time-varying covariates which cause compositional changes in X between the pre- and post-treatment periods. Figure 3 clearly elucidates the implications of the new DGPs in terms of the covariate distribution among the different groups. We can see that the distribution of X_2, in addition to being heterogeneous among treated and controls, varies within each of the two treatment statuses in the time dimension as well, e.g. between $t = 0$ and $t = 1$.

The second change consists in allowing the treatment effect to vary with X, instead of assuming the same treatment effect for each individual. This is achieved by denoting the treatment effect as $\overline{\delta}(W) = -10W_1 + 10W_2 - 10W_3 - 10W_4$. Such a functional choice of the treatment effects, despite being realistic in magnitude, is arbitrary, and therefore other functional forms are tested in the Appendix. In addition, to guarantee that the ATT is zero as in the two previous experiments, we use the demeaned transformation of $\overline{\delta}(W)$, e.g. $\delta(W) = \overline{\delta}(W) - E_{i|D=1}[\overline{\delta}(W)]$, where $E_{i|D=1}[\overline{\delta}(W)]$ denotes the ATT before demeaning. The results of the simulations are displayed in tables 12 to 15.

Experiment 2 represents the most realistic setting for a typical researcher, and therefore its implications are particularly relevant. In all DGPs, the traditional TWFE specification is severely biased, as well as the version including the time dummy and covariates interactions. However, when we control for the interaction between covariates and the treatment group dummy as well, the estimates are characterized by a substantial correction. For example, in Exp.2D, when both the propensity score and the outcome regression models are incorrectly specified, the model with full correction has a limited bias (−1.707) and RMSE (4.897). However, as shown in the Appendix, such a correction accounts for the heterogeneity in the covariate distribution for the four groups defined by the time and treatment group dimension, but it is still weak to the issue of heterogeneity in treatment effects. The IPW and RA, since they do not handle time-varying covariates, show relevant bias and variance, especially in the case of misspecified models. On the contrary, the doubly robust estimator DRDiD, despite relying on the same assumption of the two previous estimators, is shown to be particularly robust to compositional changes when the underlying models are correctly specified. In the optimistic scenario of Exp.2A, DRDiD is approximately unbiased (−0.006) and has the lowest RMSE. In the more realistic scenario of Exp.2D, its bias, despite being contained, is instead sizeable (−4.402). Here indeed, other estimators work better. In addition to the lasso version of the DRDiD, whose bias is close to zero (0.387), the 3WDRDiD, namely the modified version of the original DRDiD by Sant'Anna and Zhao (2020) with the triple matching propensity score weights, achieves the lowest bias (−0.285). Similar or even better performances are obtained by the 3IPWRA estimator and its random forest version, that have an almost identical degree of bias (−0.385 and 0.390 respectively) but are more efficient, being characterized by the lowest RMSE overall (3.913 and 2.893 respectively). Overall, the evidence shows that in Experiment 2 the 3IPWRA is the most robust estimator in terms of bias and RMSE in case of misspecified propensity score and outcome regression models, which is the likely scenario in which a typical researcher operates.

In general, considering the results of the three experiments, we recommend using the DRDiD and 3IPWRA estimators over the traditional TWFE with corrections. 3IPWRA evolves over the TWFE correction by including a semi-parametric propensity score model, while DRDiD allows for heterogeneous treatment effects in X. In addition, the semi-parametric form of these two new classes of estimators enables the researcher to flexibly deal with non-linearities in the outcome and propensity score models and with heterogeneous effects. A useful empirical strategy may be therefore to compare and combine the estimations of the different versions of the DRDiD and the 3IPWRA, since they rely on different assumptions. While the first is not built to handle time-varying covariates, the second seems to be weak to strong heterogeneous effects, as shown in the Appendix. Consequently, if the two estimators converge on comparable results, this may constitute an indication of the validity of their empirical findings. However, when using methods that do not rely exclusively on pre-treatment levels of the covariates, such as 3IPWRA and TWFE with corrections, researchers should pay attention not to include bad controls since conditioning on variables already affected by the treatment invalidates identification and is likely to generate a bias (Zeldow and Hatfield, 2019).

5 Empirical illustration: the effect of tariff reduction on corruption behaviors

We illustrate the implications of using alternative estimation methods by reproducing the analysis in Sequeira (2016) who investigate the effect of tariff reduction on corruption behaviors by using bribe payment data on the cargo shipments transiting from South Africa into the ports in Mozambique. This contribution adds to a rich debate on whether a decrease in tariff rates disincentives corruption. On the one side, tariff rates decreases are expected to lower the incidence of bribing behavior since they reduce the marginal advantage to evade taxes (Allingham and Sandmo, 1972; Poterba, 1987; Fisman and Wei, 2001). On the other side, lower tariff levels have also an income effect, increasing private agents' resources to pay higher bribes (Slemrod and Yitzhaki, 2002; Feinstein, 1991).

In 1996, a trade agreement between South Africa and Mozambique paced a series of tariff reductions that took place between 2001 and 2015, with the largest of them occurring in 2008 and entailing an average nominal tariff rate of about 5 percentage points. In this context, Sequeira (2016) collected primary data on the bribe payments of shipments imported from South Africa to Mozambique from 2007 to 2013 through an audit study. As previously documented in Sequeira and Djankov (2014), it was common for cargo owners, in exchange for tariff evasion, or simply to avoid the threat of being cited for real or fictious irregularities, to bribe border officials in charge of collecting all tariff payment and of providing clearance documentation. For example, prior to 2008, approximately 80 percent of the random sample of tracked shipments were linked with sizeable bribe payment during the clearing process (mean bribes reached USD 128 per tonnage). As a consequence, Sequeira (2016) exploits the exogenous change in tariffs induced by the trade agreement to examine the effect of changes in tariffs on corruption levels. Since not all products experienced a variation in tariff rates during this period, the author adopts a Difference-in-Difference design to isolate the causal relationship between tariffs and corruption, on pooled cross sectional data collected between 2007 and 2013, for a total of 1084 observations. More specifically, the design is based on the canonical TWFE estimator in the following specification:

$$\begin{aligned} y_{it} = & \gamma_1(TariffChangeCategory_i \times POST) + \mu POST \\ & + \beta_1 TariffChangeCategory_i + \beta_2 BaselineTariff_i \\ & + \Gamma_i + p_i + +\omega_i + \delta_i + \epsilon_{it} \end{aligned} \tag{36}$$

where y_{it} represents the natural log of the amount of bribe paid for shipment i in period t, conditional on paying a bribe, $TariffChangeCategory_i \in \{0,1\}$ takes value one if the commodity was subject to tariff reduction, $POST \in \{0,1\}$ denotes the years following 2008, and $BaselineTariff_i$ is a control for the pre-treatment tariff for product i. The specification also accounts for a vector of product, shipment, clearing agent, and firm-level characteristics Γ_i which include the elements summarised in Table 16. Industry, year, and clearing agent fixed effects are included, denoted by p_i, ω_t, and δ_i respectively. The parameter of interest is the coefficient of the interaction between the time and treatment dummies, namely γ_1.

The main result of Sequeira (2016) is that the tariff reduction led to a significant drop in the amount of bribe paid. Chang (2020) replicated the paper comparing the results to the one obtained by his proposed estimator, the debiased machine learning Difference-in-Difference (DMLDiD) estimator. DMLDiD, which builds on Abadie (2005), is an IPW estimator whose score function is adapted to satisfy the Neyman orthogonality conditions. This allows researchers to flexibly use a rich set of machine learning methods in the first-step estimation. Table 17 summarizes the results obtained in the two papers, where

TWFE is the standard specification in Sequeira (2016), which is found in Equation 1 of Table 9 in the paper, while TWFE ($\Gamma_i\times$ POST) is the specification that also includes the interactions between the covariates Γ_i and $POST$, which is equation 2 of the same table. DMLDiD refers instead to the estimates obtained by Chang (2020) and is either estimated by using a first-stage kernel estimation or a lasso. Chang's semiparametric estimates indicate that the effect of the reduction was larger than originally thought.

However, both classes of estimators utilized in the paper may be characterized by a substantial degree of bias. As shown in Section 2.4 and Section 4, under a non-randomized treatment scenario the standard TWFE is likely to be biased because of the possible presence of X-specific trends, compositional changes, heterogeneous effects, and non-linearities. Likewise, the DMLDiD estimator has weaknesses that hinder the reliability of its findings, as well. First, simulations in Section 4 showed that the IPW class of estimators can be severely biased under realistic settings and was outperformed by other estimators. Second, DMLDiD estimates in Table 17 suffer from very high standard errors which blur the interpretation of its findings. For example, the 95 percent confidence interval lies approximately between 0.318 and -14.306 for the kernel DMLDiD, and the same applies to its lasso version, even if to a smaller degree.

To gain additional evidence on the performance of the DMLDiD estimator, we tested it under the most relevant scenarios outlined in the Monte Carlo simulation of Section 4. In particular, since DMLDiD requires a substantially higher amount of computational time with respect to the other analyzed estimators, we ony conducted Experiments 1 and 2. The simulation results presented in Table 18 demonstrate that the degree of bias of the DMLDiD is substantial and far higher than the one experienced by the 3IPWRA and DRDiD classes of estimators. Even when both the propensity score and outcome models are correctly specified, in Experiment 1 and 2 the DMLDiD estimator has a bias of -36.052 and -31.348, respectively.

When both models are instead misspecified, its performance, in terms of RMSE, considerably worsens. The bias in Experiment 1 reaches a value of -117.857, while it decreases to -0.807 in Experiment 2. However, this last result is likely to be driven by different sorts of biases balancing out since the estimator shows a significant degree of bias in all the other settings, even when the models are correctly specified, an instance where we would expect an appropriate estimator to be approximately unbiased. Overall, the variance of the DMLDiD is much larger than the one of the 3IPWRA and DRDiD estimators (8100.95 in Experiment 2D in Table 18), hindering the reliability of any practical application of the estimator.

Motivated by these considerations, we employ the IPWRA and DRDiD classes of estimators, which proved to be the most efficient estimators in terms of bias in the Monte Carlo simulations, to the current study of the effect of tariff reduction on bribing behaviour. In such a setting, the low number of observations does not allow for traditional first-stage estimation methods to produce accurate fitted values, favouring the use of lasso and random forests. Indeed, the sample is characterized by a large number of observations for the control group in the post-treatment period, but a limited number of observations for the other three groups, namely the treated in the pre and post-treatment period (120 and 56 observations respectively), and the controls in the pre-treatment period (84). The lasso specification captures non-linearities by allowing for a richer set of covariates: Γ_i is expanded to include all second order terms and interactions, leading to a set of 112 controls. Contrarily to Chang (2020), we stick to the specification in Sequeira (2016) by including all industry, time and clearing agent fixed-effects. In this case, the interactions with Γ_i are not generated for computational tractability. The ATT is estimated using both lasso and random forest 3IPWRA, and the lasso version of DRDiD, since these estimators perform well in the Monte Carlo simulations in Section 4. When operating with these three estimators, standard errors are computed through weighted bootstrap, similarly to Sant'Anna and Zhao (2020). To allow for clusters, the random weights in the bootstrap procedure are defined at the cluster level and not at the individual level so that each observation belonging to a specific cluster receives the same random weight in each of the repetitions of the bootstrap.

Our final estimates are displayed in Table 19. Our results, across different methods and specifications, corroborate the hypothesis that the tariff reduction led to a drop in the amount of bribe paid, but give compelling evidence against the assumption that the effect was higher in magnitude. In fact, the standard TWFE seems to overestimate the ATT (-3.748) since all the methods characterized by the best results in our simulations converge to lower values: the TWFE($\Gamma_i \times POST$+$\Gamma_i \times D$) estimate is -3.667, the lasso

3IPWRA is -3.023, the random forest 3IPWRA is -3.216, and the lasso DRDiD is -2.764. The standard errors are typically lower than those in Chang (2020), allowing for much more precise confidence intervals estimations.

In summary, our findings reveal that the tariff reduction had a significant effect on bribing behavior, but the impact is smaller than originally estimated by the standard TWFE specification and by DMLDiD as in Chang (2020). The average of our estimates is -3.167, which is closer to the TWFE estimate with the correction in Sequeira (2016).

6 Conclusion

Our analysis shows that the commonly-used standard TWFE may be severely biased under recurrent settings. We performed a series of Monte Carlo simulations in order to assess the performance of TWFE with corrections and other semi-parametric estimators. Despite including both time and treatment group interactions with the covariates provides a substantial correction when dealing with heterogeneity in the time and treatment group dimensions, TWFE tends to be outperformed by other semi-parametric methods, such as DRDiD and 3IPWRA. Indeed, the semi-parametric form of these two estimators enables the researcher to flexibly deal with non-linearities in the outcome and propensity score models and with heterogeneous effects.

The DRDiD estimator proposed by Sant'Anna and Zhao (2020) is characterized by better performance in terms of bias and mean square error when both the propensity score and outcome models are correctly specified. However, in the more realistic scenario when they are not, the class of 3IPWRA estimators proposed by the authors, based on Blundell and Dias (2009), outperforms DRDiD in presence of compositional changes. We also propose a number of modifications to the DRDiD estimator, allowing for lasso or random forests first-stage estimates, or applying triple matching propensity score weights (naming the estimator 3WDRDiD in this case), that are also characterized by a relatively good performance. In particular, the lasso version of the DRDID estimator has consistently lower bias than its original version over all the three experiments when the outcome and the propensity score models are both misspecified.

Based on our Monte Carlo simulations' results, we propose a practical empirical strategy for estimating causal effects in a DiD context based on the combination of the best performing estimators, i.e. the different versions of the DRDiD and the 3IPWRA. The two classes of estimators rely on different sets of assumptions. While the first is not built to handle time-varying covariates, the second seems to be weak to particularly strong heterogeneous effects. In addition, 3IPWRA requires particular care in order not to include bad controls (as in TWFE corrections). Therefore, in case of convergence of these methods' findings, this may constitute an indication of the validity of an hypothesis under study.

In addition, we illustrate the implications of using alternative estimation methods by reproducing the analysis in Sequeira (2016) who investigate the effect of tariff reduction on corruption behaviors by using bribe payment data on the cargo shipments transiting from South Africa into the ports in Mozambique. Our estimates show that tariff reduction led to a decrease in bribes paid but the effect is lower in magnitude than the one originally estimated by Sequeira's standard TWFE specification and by the DMLDiD replication in Chang (2020).

As a suggestion for further research, other DiD estimators offer a number of potential advantages and should be evaluated in a Monte Carlo setting. In particular, Nie et al. (2021) and Zimmert (2018) offer interesting alternatives following the debiased machine learning literature of Chernozhukov et al. (2018).

References

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

Michael G. Allingham and Agnar Sandmo. Income tax evasion: a theoretical analysis. *Journal of Public Economics*, 1(3-4):323–338, 1972. URL https://EconPapers.repec.org/RePEc:eee:pubeco:v:1:y:1972:i:3-4:p:323-338.

Philipp Bach, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. Doubleml–an object-oriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603*, 2021.

Richard Blundell and Monica Costa Dias. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3):565–640, 2009.

Richard Blundell, Monica Costa Dias, Costas Meghir, and John Van Reenen. Evaluating the employment impact of a mandatory job search program. *Journal of the European economic association*, 2(4):569–606, 2004.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Scott Cunningham. A tale of time varying covariates. https://causalinf.substack.com/p/a-tale-of-time-varying-covariates, 2021. Accessed: 07/02/2022.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Jonathan S. Feinstein. An econometric analysis of income tax evasion and its detection. *RAND Journal of Economics*, 22(1):14–35, 1991. URL https://EconPapers.repec.org/RePEc:rje:randje:v:22:y:1991:i:spring:p:14-35.

Raymond Fisman and Shang-Jin Wei. Tax Rates and Tax Evasion: Evidence from &quot;Missing Imports&quot; in China. NBER Working Papers 8551, National Bureau of Economic Research, Inc, October 2001. URL https://ideas.repec.org/p/nbr/nberwo/8551.html.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010. URL https://www.jstatsoft.org/v33/i01/.

Bryan S. Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse Probability Tilting for Moment Condition Models with Missing Data. *The Review of Economic Studies*, 79(3):1053–1079, 04 2012. ISSN 0034-6527. doi: 10.1093/restud/rdr047. URL https://doi.org/10.1093/restud/rdr047.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

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

Torsten Hothorn, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro, and Mark Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Michael Lechner et al. *The estimation of causal effects by difference-in-difference methods.* Now Hanover, MA, 2011.

Breed D Meyer. Natural and quasi-experiments in economics. *Journal of business & economic statistics*, 13(2):151–161, 1995.

Xinkun Nie, Chen Lu, and Stefan Wager. Nonparametric heterogeneous treatment effect estimation in repeated cross sectional designs, 2021.

Thais Oshiro, Pedro Perez, and José Baranauskas. How many trees in a random forest? *Lecture notes in computer science*, 7376, 07 2012. doi: 10.1007/978-3-642-31537-4_13.

James M Poterba. Tax Evasion and Capital Gains Taxation. *American Economic Review*, 77(2):234–239, May 1987. URL https://ideas.repec.org/a/aea/aecrev/v77y1987i2p234-39.html.

Jonathan Roth, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*, 2022.

Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

Sandra Sequeira. Corruption, trade costs, and gains from tariff liberalization: Evidence from southern africa. *American Economic Review*, 106(10):3029–63, 2016.

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

Joel Slemrod and Shlomo Yitzhaki. Tax avoidance, evasion, and administration. 3: 1423–1470, 2002.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25), 2007. doi: 10.1186/1471-2105-8-25.

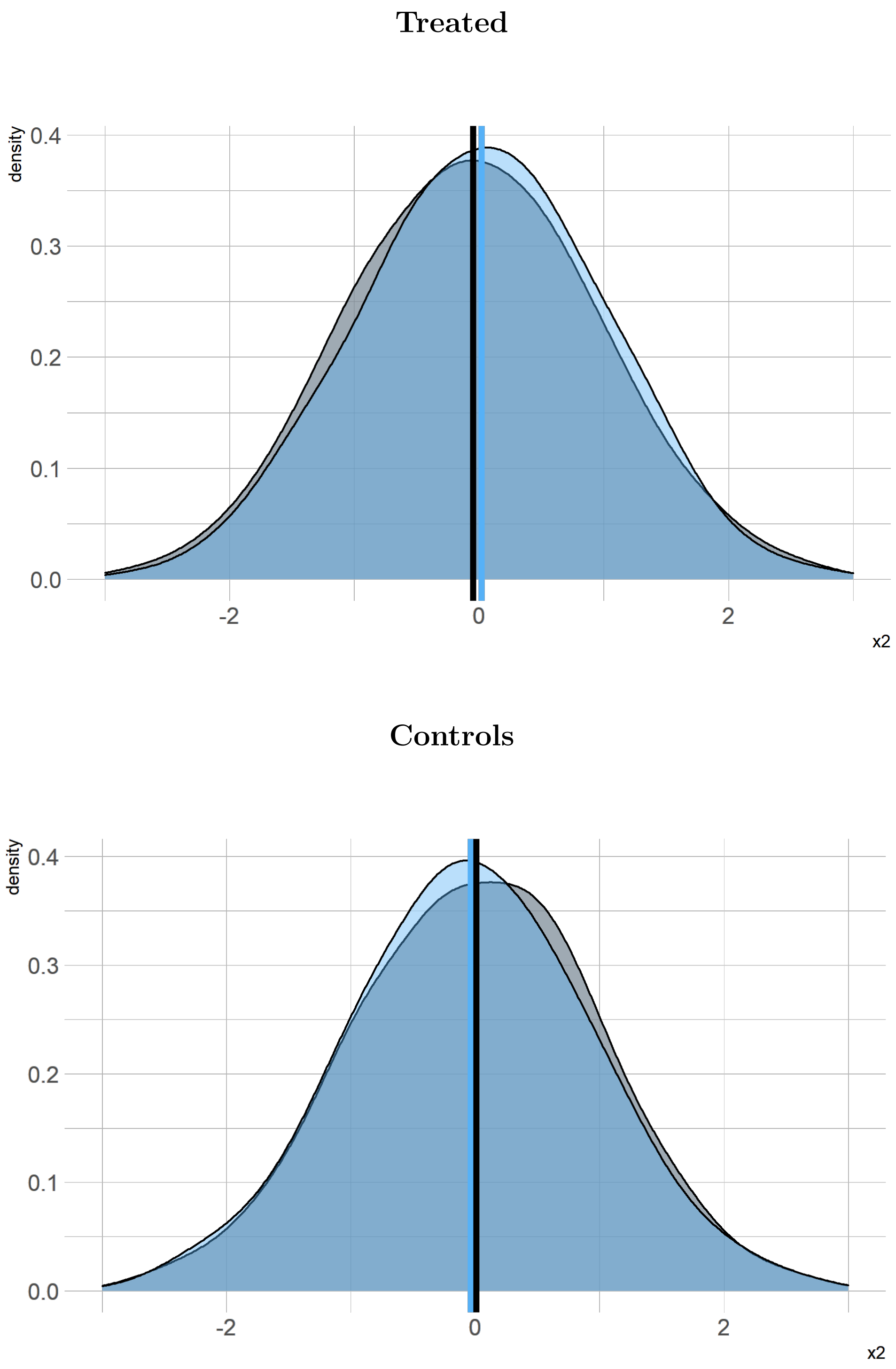
Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9 (307), 2008. doi: 10.1186/1471-2105-9-307.

Bret Zeldow and Laura A Hatfield. Confounding and regression adjustment in difference-in-differences. *arXiv preprint arXiv:1911.12185*, 2019.

Michael Zimmert. Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding. Papers 1809.01643, arXiv.org, September 2018. URL https://ideas.repec.org/p/arx/papers/1809.01643.html.

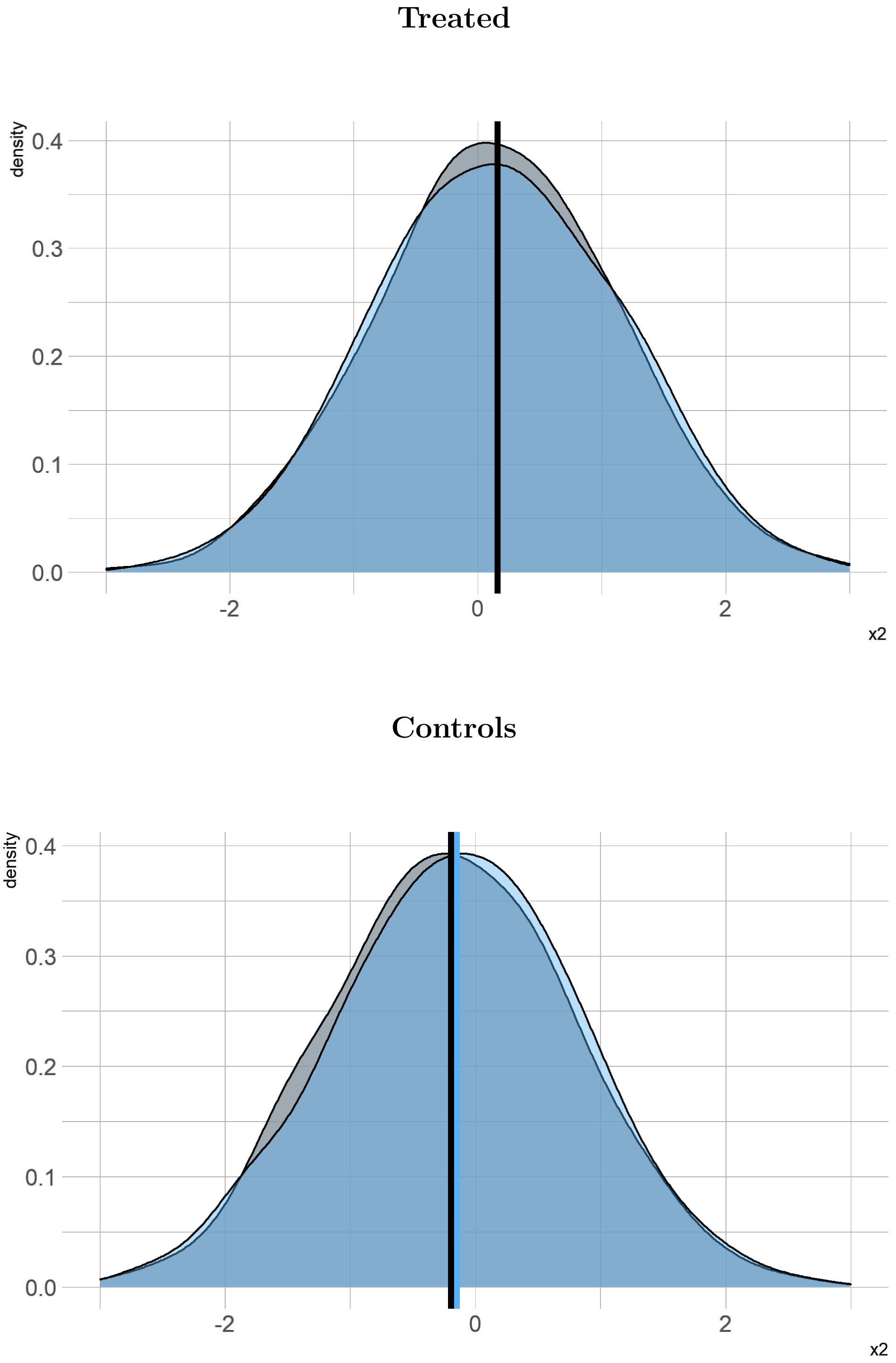
7 Main Figures

Figure 1: Density plot of X_2 among treated in pre- (black) and post- (blue) treatment periods



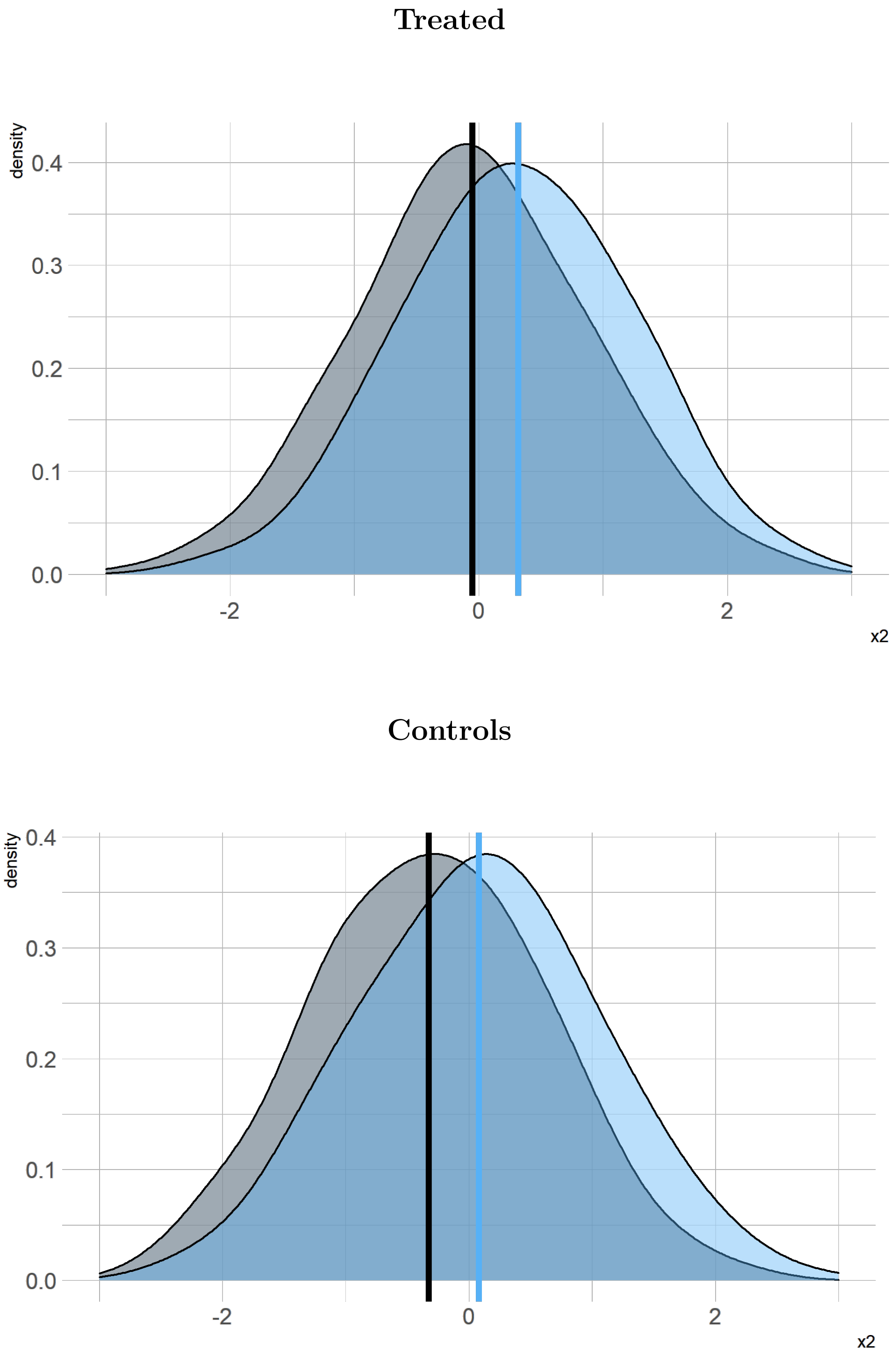
Notes: The graph considers a representative random sample from Exp.0 with DGP CD. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the selected treatment group category, while the blue one is the mean of the distribution of X_2 in the post-treatment period for the same treatment group category. Note that the distribution of X_2 is time-invariant and, because of randomization, the distribution is approximately the same among treated and controls.

Figure 2: Density plot of X_2 among treated in pre- (black) and post- (blue) treatment periods



Notes: The graph considers a representative random sample from Exp.1 with DGP D. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the selected treatment group category, while the blue one is the mean of the distribution of X_2 in the post-treatment period for the same treatment group category. Note that the distribution of X_2 is time-invariant but there is heterogeneity between treated and controls populations, as captured by their difference in means.

Figure 3: Density plot of X_2 among treated in pre- (black) and post- (blue) treatment periods



Treated

Controls

Notes: The graph considers a representative random sample from Exp.2 with DGP D. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the selected treatment group category, while the blue one is the mean of the distribution of X_2 in the post-treatment period for the same treatment group category. Note the heterogeneity in both the time and treatment group dimensions.

8 Main Tables

Table 1: Summary table of the experiments

	Description
EXP.0	Randomized experiment, homogeneous effects in X and time-invariant covariates
EXP.1	Non-randomized experiment, homogeneous effects in X and time-invariant covariates
EXP.2	Non-randomized experiment, heterogeneous effects in X and time-varying covariates

Table 2: Summary table of the estimator utilized in the Monte Carlo simulation

Estimator	Description
TWFE	two-way-fixed-effects regression with covariates as in eq. (8)
TWFE (T·X)	two-way-fixed-effects regression with covariates and their interaction with the time dummy, as in eq. (11)
TWFE (T·X+D·X)	two-way-fixed-effects regression with covariates and their interaction with the time and treatment group dummies, as in eq. (12)
IPW	Inverse probability weighting (Abadie, 2005)
RA	Outcome regression (Heckman et al., 1997)
DRDiD	Improved locally efficient doubly robust estimator, original version (Sant'Anna and Zhao, 2020)
LASSO DRDiD	Locally efficient doubly robust estimator, Sant'Anna and Zhao (2020) modified with lasso
RF DRDiD	Locally efficient doubly robust estimator, Sant'Anna and Zhao (2020) modified with random forest
3IPWRA	Triple propensity score inverse probability weighting regression adjusted estimator with logit regression for the propensity score. Builds on the idea in Blundell and Dias (2009) of a propensity score that balances heterogeneity in both the time and treatment group dimensions and employs it in a weighted regression
LASSO 3IPWRA	Triple propensity score inverse probability weighting regression adjusted estimator with lasso estimation of the propensity score.
RF 3IPWRA	Triple propensity score inverse probability weighting regression adjusted estimator with random forest estimation of the propensity score
3WDRDiD	Doubly robust DiD (Sant'Anna and Zhao, 2020) adjusted with triple propensity score inverse probability weighting

Table 3: DGPs in Experiment 0 (PS=propensity score, OR=outcome regression)

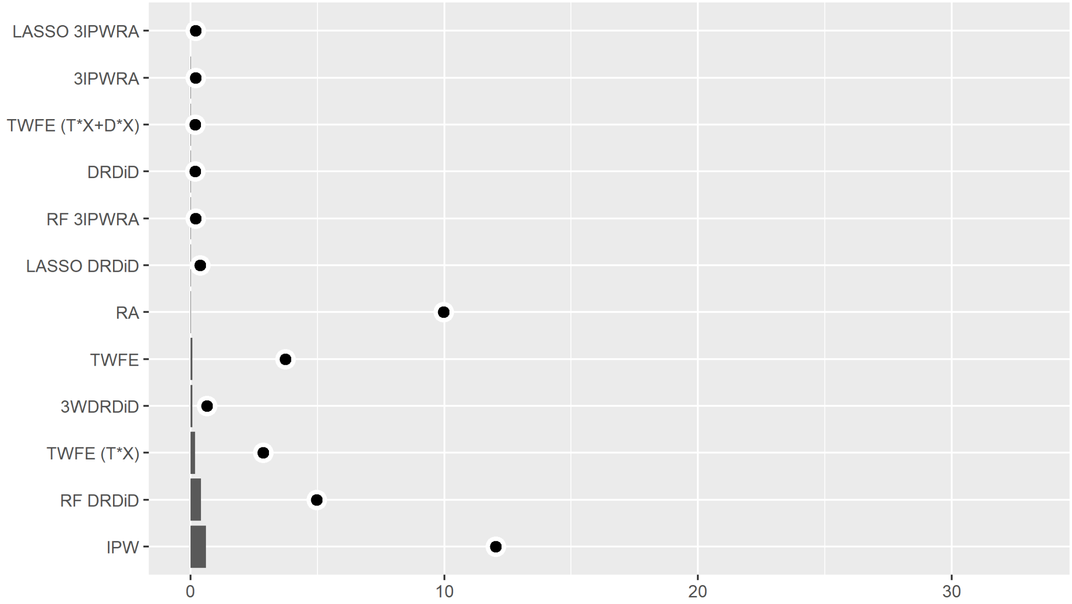
DGP.AB (PS and OR models correct)	**DGP.CD (PS and OR models incorrect)**
$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$	$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$
$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(D)$	$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \epsilon_1(D)$
$p = 0.5$	$p = 0.5$
$\lambda = 0.5$	$\lambda = 0.5$
$D = 1\{p \geq U_d\}$	$D = 1\{p \geq U_d\}$
$T = 1\{\lambda \geq U_t\}$	$T = 1\{\lambda \geq U_t\}$

Notes: EXP.0 assumes a randomized experiment, homogeneous effects in X and time-invariant covariates.

Table 4: Exp.0AB Outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−0.060	3.733	13.929	0.001
Regression, eq. (11)	TWFE (T·X)	−0.178	2.852	8.105	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.009	0.184	0.034	0.001
Abadie (2005)	IPW	0.596	12.032	144.413	0.014
Heckman et al. (1997)	RA	−0.022	9.976	99.530	0.013
Sant'Anna and Zhao (2020)	DRDiD	0.009	0.184	0.034	0.019
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.016	0.375	0.141	1.068
Sant'Anna and Zhao (2020)*	RF DRDiD	−0.406	4.959	24.428	3.429
Authors' work, eq. (27)	3IPWRA	0.008	0.185	0.034	0.018
Authors' work, eq. (27)	LASSO 3IPWRA	−0.005	0.186	0.035	2.285
Authors' work, eq. (27)	RF 3IPWRA	0.012	0.193	0.037	1.287
Sant'Anna and Zhao (2020)*	WDRDiD	0.070	0.632	0.395	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

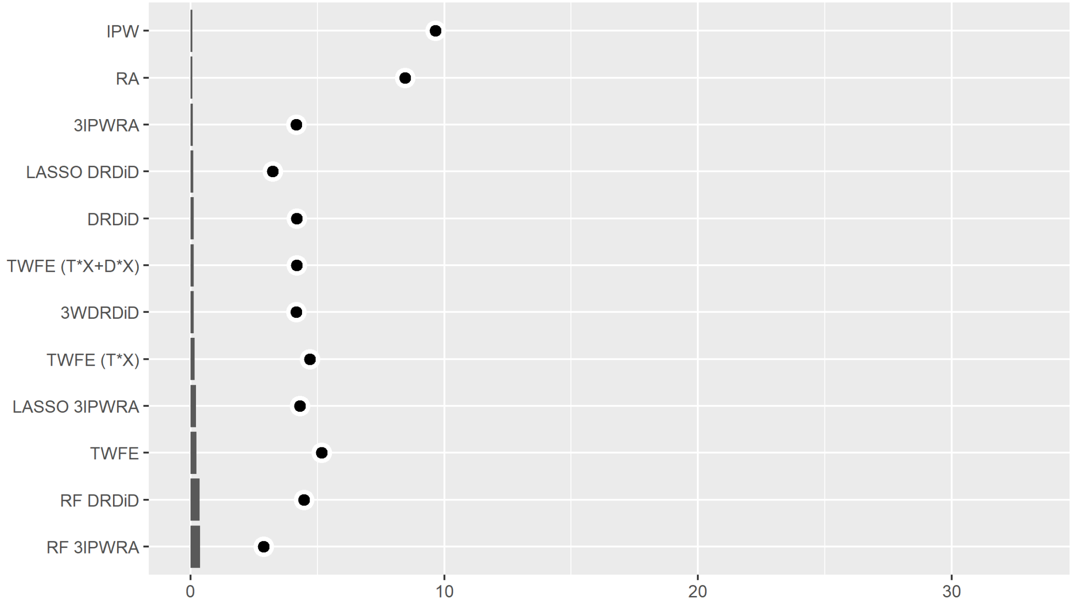


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.0 assumes a randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 5: Exp.0CD Outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	0.228	5.167	26.647	0.001
Regression, eq. (11)	TWFE (T·X)	0.157	4.702	22.083	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.119	4.187	17.517	0.002
Abadie (2005)	IPW	−0.074	9.643	92.988	0.013
Heckman et al. (1997)	RA	−0.075	8.449	71.383	0.012
Sant'Anna and Zhao (2020)	DRDiD	0.114	4.178	17.441	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.105	3.230	10.420	1.148
Sant'Anna and Zhao (2020)*	RF DRDiD	−0.350	4.462	19.792	3.216
Author's work, eq. (27)	3IPWRA	0.089	4.159	17.292	0.053
Author's work, eq. (27)	LASSO 3IPWRA	−0.218	4.305	18.488	2.05
Author's work, eq. (27)	RF 3IPWRA	0.379	2.875	8.122	1.246
Sant'Anna and Zhao (2020)*	3WDRDiD	0.123	4.155	17.252	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator



Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.0 assumes a randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 6: DGPs in Experiment 1 (PS=propensity score, OR=outcome regression)

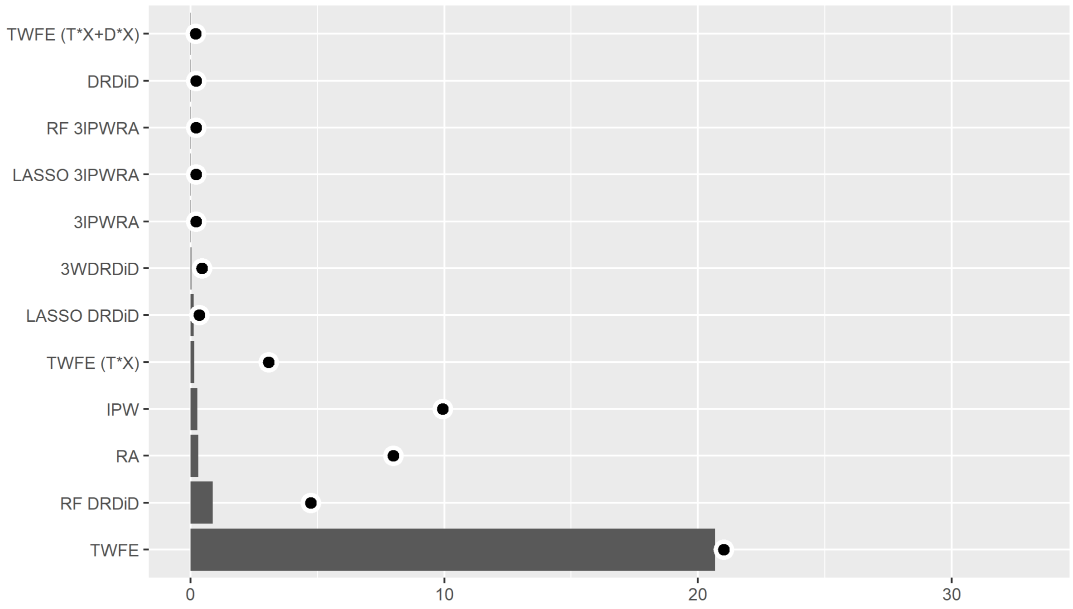
DGP.A (PS and OR models correct)	**DGP.B (PS model incorrect, OR correct)**
$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$	$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0(d)$
$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(d)$	$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(d)$
$p(Z) = \dfrac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$	$p(X) = \dfrac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$
$\lambda = 0.5$	$\lambda = 0.5$
$D = 1\{p(Z) \geq U_d\}$	$D = 1\{p(X) \geq U_d\}$
$T = 1\{\lambda \geq U_t\}$	$T = 1\{\lambda \geq U_t\}$
DGP.C (PS model correct, OR incorrect)	**DGP.D (PS and OR models incorrect)**
$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$	$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$
$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \epsilon_1(d)$	$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \epsilon_1(d)$
$p(Z) = \dfrac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$	$p(X) = \dfrac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$
$\lambda = 0.5$	$\lambda = 0.5$
$D = 1\{p(Z) \geq U_d\}$	$D = 1\{p(X) \geq U_d\}$
$T = 1\{\lambda \geq U_t\}$	$T = 1\{\lambda \geq U_t\}$

Notes: EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates.

Table 7: Exp.1A Propensity score model correct, outcome regression correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−20.686	21.021	13.991	0.001
Regression, eq. (11)	TWFE (T·X)	−0.131	3.066	9.386	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.006	0.185	0.034	0.001
Abadie (2005)	IPW	−0.259	9.927	98.480	0.014
Heckman et al. (1997)	RA	−0.308	7.995	63.833	0.010
Sant'Anna and Zhao (2020)	DRDiD	0.006	0.203	0.041	0.018
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.116	0.345	0.106	1.007
Sant'Anna and Zhao (2020)*	RF DRDiD	−0.871	4.738	21.686	3.147
Author's work, eq. (27)	3IPWRA	0.007	0.203	0.041	0.045
Author's work, eq. (27)	LASSO 3IPWRA	−0.007	0.216	0.047	1.878
Author's work, eq. (27)	RF 3IPWRA	0.007	0.212	0.045	1.240
Sant'Anna and Zhao (2020)*	3WDRDiD	0.027	0.438	0.191	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

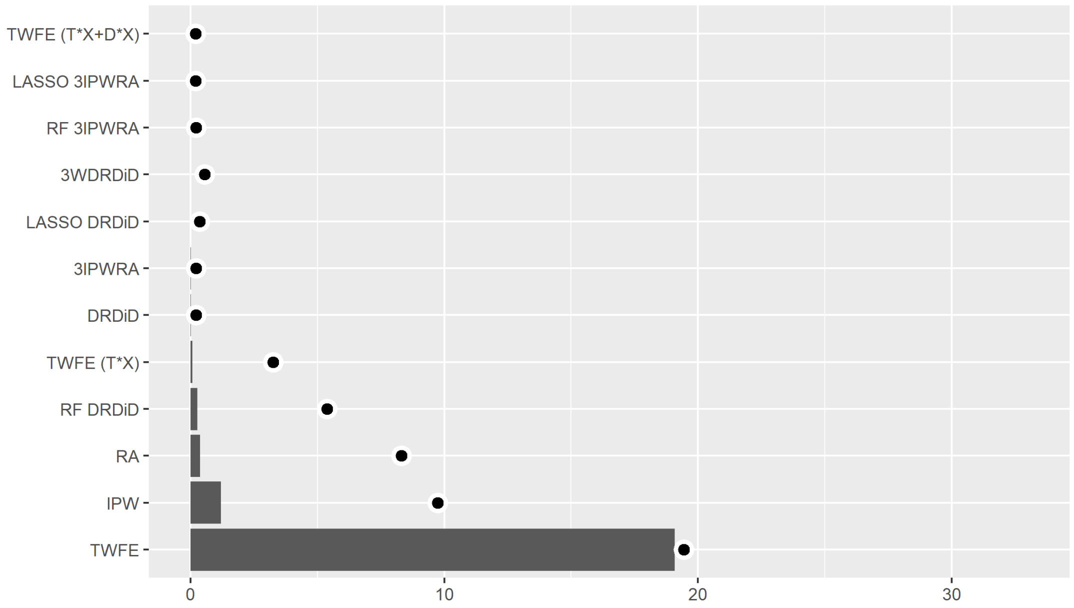


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 8: Exp.1B Propensity score model incorrect, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−19.095	19.449	13.670	0.001
Regression, eq. (11)	TWFE (T·X)	−0.066	3.254	10.586	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.001	0.194	0.038	0.002
Abadie (2005)	IPW	−1.191	9.739	93.436	0.014
Heckman et al. (1997)	RA	−0.369	8.305	68.842	0.010
Sant'Anna and Zhao (2020)	DRDiD	0.008	0.210	0.044	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.004	0.346	0.120	1.119
Sant'Anna and Zhao (2020)*	RF DRDiD	−0.270	5.377	28.839	3.245
Author's work, eq. (27)	3IPWRA	0.006	0.207	0.043	0.039
Author's work, eq. (27)	LASSO 3IPWRA	0.001	0.195	0.038	1.977
Author's work, eq. (27)	RF 3IPWRA	0.002	0.212	0.045	1.273
Sant'Anna and Zhao (2020)*	3WDRDiD	−0.002	0.552	0.304	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

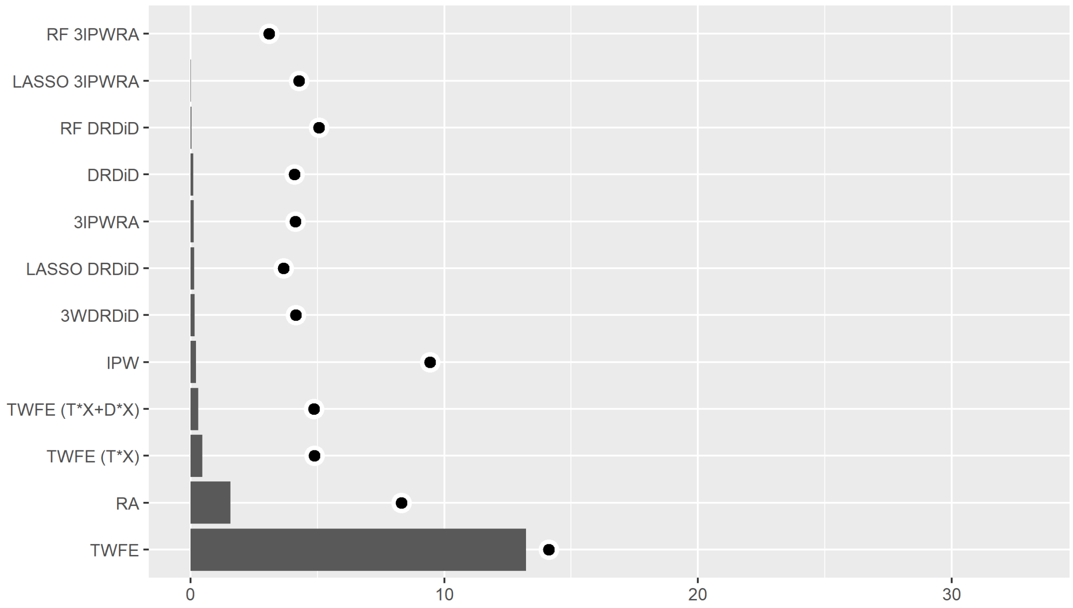


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 9: Exp.1C Propensity score model correct, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−13.220	14.120	24.609	0.001
Regression, eq. (11)	TWFE (T·X)	−0.459	4.875	23.559	0.002
Regression, eq. (12)	TWFE (T·X+D·X)	−0.300	4.852	23.450	0.001
Abadie (2005)	IPW	0.204	9.439	89.050	0.013
Heckman et al. (1997)	RA	−1.563	8.317	66.732	0.010
Sant'Anna and Zhao (2020)	DRDiD	−0.095	4.087	16.698	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.142	3.657	13.351	1.089
Sant'Anna and Zhao (2020)*	RF DRDiD	0.033	5.061	25.610	3.121
Author's work, eq. (27)	3IPWRA	−0.113	4.134	17.081	0.020
Author's work, eq. (27)	LASSO 3IPWRA	0.006	4.265	18.193	1.873
Author's work, eq. (27)	RF 3IPWRA	0.004	3.089	9.543	1.244
Sant'Anna and Zhao (2020)*	3WDRDiD	−0.155	4.144	17.145	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

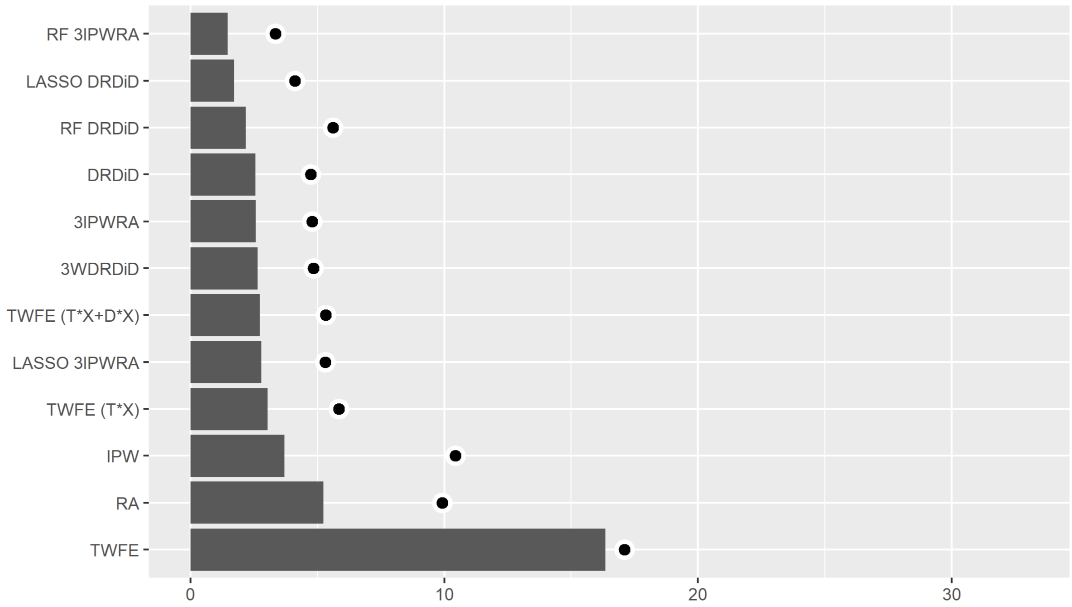


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 10: Exp.1D Propensity score model incorrect, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−16.355	17.106	25.127	0.001
Regression, eq. (11)	TWFE (T·X)	−3.030	5.838	24.904	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−2.727	5.317	20.834	0.001
Abadie (2005)	IPW	−3.702	10.439	95.264	0.014
Heckman et al. (1997)	RA	−5.242	9.909	70.700	0.010
Sant'Anna and Zhao (2020)	DRDiD	−2.555	4.727	15.814	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−1.720	4.116	13.981	1.140
Sant'Anna and Zhao (2020)*	RF DRDiD	−2.175	5.618	26.835	3.117
Author's work, eq. (27)	3IPWRA	−2.578	4.787	16.267	0.026
Author's work, eq. (27)	LASSO 3IPWRA	−2.794	5.301	20.292	1.991
Author's work, eq. (27)	RF 3IPWRA	−1.470	3.333	8.951	1.233
Sant'Anna and Zhao (2020)*	3WDRDiD	−2.634	4.838	16.466	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator



Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 11: DPGs in Experiment 2 (PS=propensity score, OR=outcome regression)

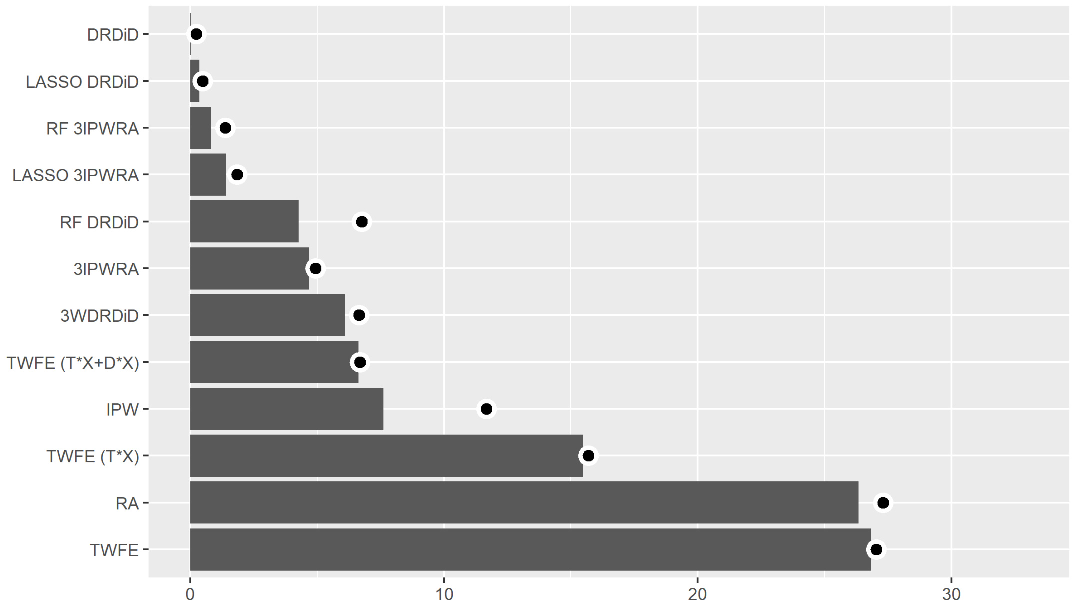
DGP.A (PS and OR models correct)	**DGP.B (PS model incorrect, OR correct)**
$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$	$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0(d)$
$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \delta(Z) \cdot D + \epsilon_1(D)$	$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \delta(Z) \cdot D + \epsilon_1(D)$
$p(Z) = \frac{\exp(f_{ps}(Z))}{(1 + \exp(f_{ps}(Z)))}$	$p(X) = \frac{\exp(f_{ps}(X))}{(1 + \exp(f_{ps}(X)))}$
$\lambda(Z) = \frac{\exp(f_{ps}(Z))}{(1 + \exp(f_{ps}(Z)))}$	$\lambda(X) = \frac{\exp(f_{ps}(X))}{(1 + \exp(f_{ps}(X)))}$
$D = 1\{p(Z) \geq U_d\}$	$D = 1\{p(X) \geq U_d\}$
$T = 1\{\lambda(Z) \geq U_t\}$	$T = 1\{\lambda(X) \geq U_t\}$
DGP.C (PS model correct, OR incorrect)	**DGP.D (PS and OR models incorrect)**
$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$	$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$
$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \delta(X) \cdot D + \epsilon_1(D)$	$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \delta(X) \cdot D + \epsilon_1(D)$
$p(Z) = \frac{\exp(f_{ps}(Z))}{(1 + \exp(f_{ps}(Z)))}$	$p(X) = \frac{\exp(f_{ps}(X))}{(1 + \exp(f_{ps}(X)))}$
$\lambda(Z) = \frac{\exp(f_{ps}(Z))}{(1 + \exp(f_{ps}(Z)))}$	$\lambda(X) = \frac{\exp(f_{ps}(X))}{(1 + \exp(f_{ps}(X)))}$
$D = 1\{p(Z) \geq U_d\}$	$D = 1\{p(X) \geq U_d\}$
$T = 1\{\lambda(Z) \geq U_t\}$	$T = 1\{\lambda(X) \geq U_t\}$

Notes: EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates.

Table 12: 2A Propensity score model correct, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−26.834	27.033	10.691	0.001
Regression, eq. (11)	TWFE (T·X)	−15.475	15.694	6.836	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−6.624	6.688	0.852	0.002
Abadie (2005)	IPW	−7.614	11.675	78.330	0.013
Heckman et al. (1997)	RA	−26.338	27.313	52.334	0.011
Sant'Anna and Zhao (2020)	DRDiD	−0.006	0.227	0.051	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.351	0.474	0.101	1.003
Sant'Anna and Zhao (2020)*	RF DRDiD	−4.274	6.751	27.310	3.125
Author's work, eq. (27)	3IPWRA	4.681	4.937	2.456	0.024
Author's work, eq. (27)	LASSO 3IPWRA	−1.415	1.833	1.357	2.089
Author's work, eq. (27)	RF 3IPWRA	0.811	1.377	1.238	1.364
Sant'Anna and Zhao (2020)*	3WDRDiD	6.098	6.644	6.952	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

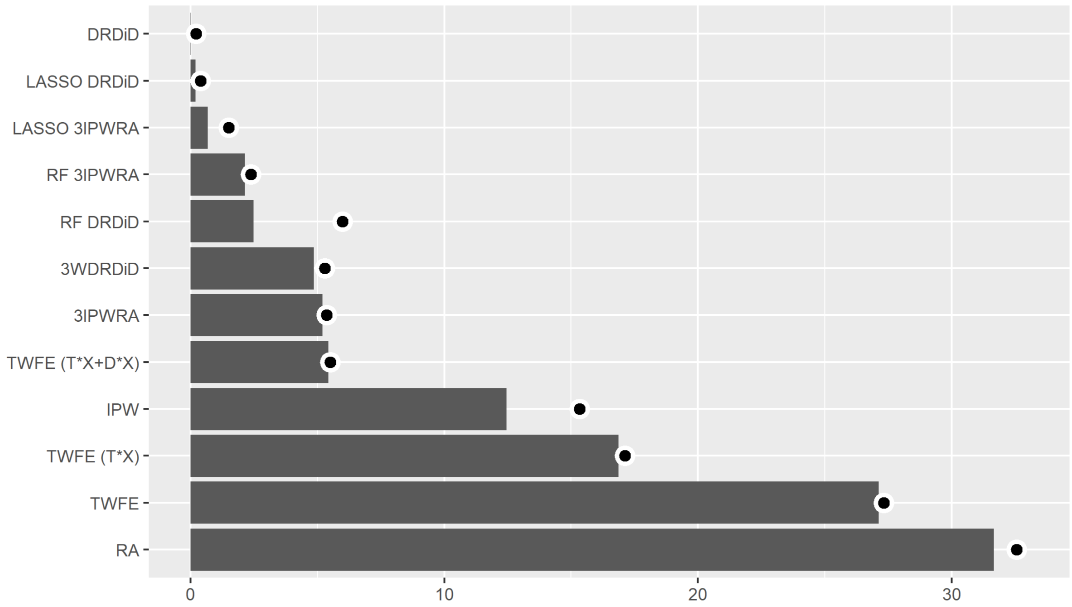


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 13: 2B Propensity score model incorrect, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−27.136	27.329	10.517	0.001
Regression, eq. (11)	TWFE (T·X)	−16.878	17.125	8.400	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−5.427	5.510	0.907	0.002
Abadie (2005)	IPW	−12.451	15.333	80.080	0.013
Heckman et al. (1997)	RA	−31.668	32.569	57.879	0.011
Sant'Anna and Zhao (2020)	DRDiD	0.008	0.216	0.047	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.197	0.387	0.111	1.068
Sant'Anna and Zhao (2020)*	RF DRDiD	−2.473	5.993	29.807	3.114
Author's work, eq. (27)	3IPWRA	5.198	5.368	1.801	0.025
Author's work, eq. (27)	LASSO 3IPWRA	0.683	1.499	1.779	2.141
Author's work, eq. (27)	RF 3IPWRA	2.136	2.376	1.083	1.345
Sant'Anna and Zhao (2020)*	3WDRDiD	4.867	5.281	4.205	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

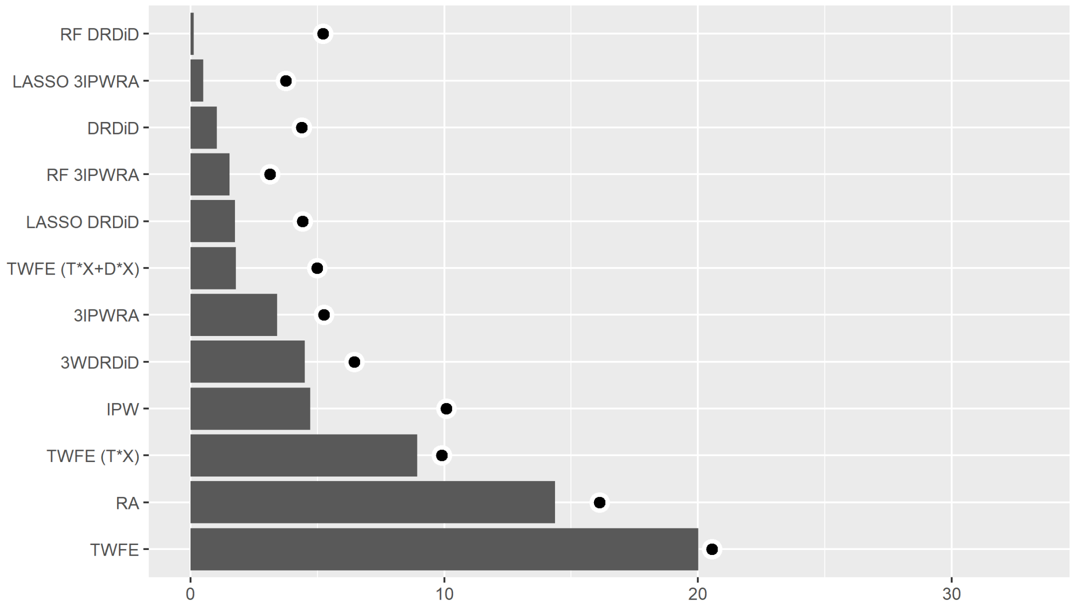


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 14: 2C Propensity score model correct, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−20.013	20.545	21.604	0.001
Regression, eq. (11)	TWFE (T·X)	−8.938	9.891	17.948	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	1.779	4.977	21.607	0.001
Abadie (2005)	IPW	−4.714	10.086	79.510	0.014
Heckman et al. (1997)	RA	−14.373	16.118	53.190	0.010
Sant'Anna and Zhao (2020)	DRDiD	1.028	4.372	18.061	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	1.754	4.408	16.351	1.096
Sant'Anna and Zhao (2020)*	RF DRDiD	−0.127	5.219	27.222	3.115
Author's work, eq. (27)	3IPWRA	3.404	5.244	15.914	0.016
Author's work, eq. (27)	LASSO 3IPWRA	0.493	3.758	13.878	2.078
Author's work, eq. (27)	RF 3IPWRA	1.538	3.125	7.401	1.369
Sant'Anna and Zhao (2020)*	3WDRDiD	4.493	6.445	21.351	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

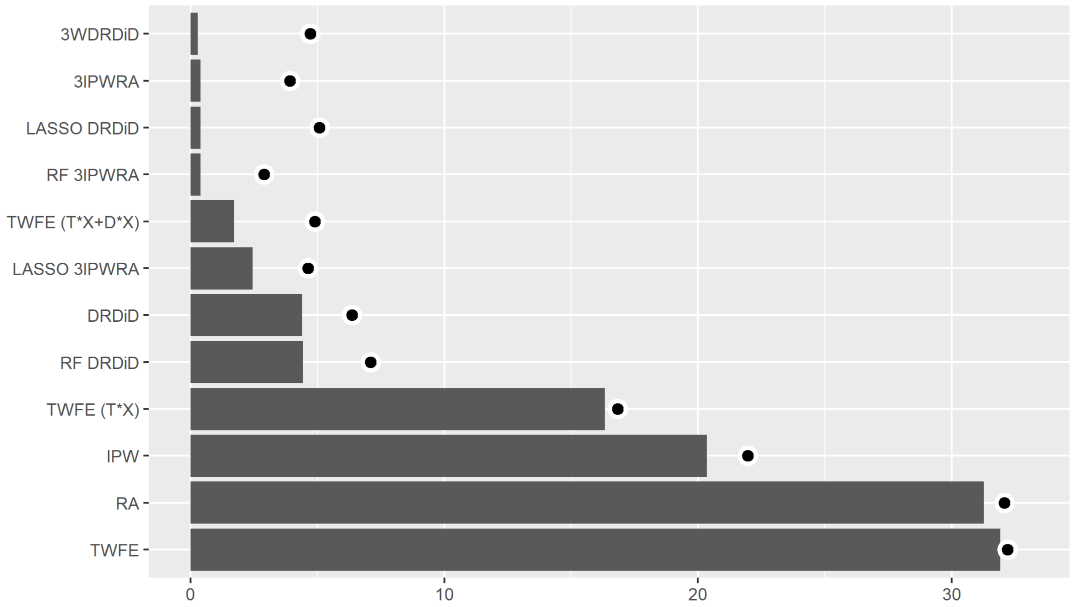


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 15: 2D Propensity score model incorrect, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−31.917	32.212	18.890	0.001
Regression, eq. (11)	TWFE (T·X)	−16.331	16.832	16.641	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−1.707	4.897	21.068	0.001
Abadie (2005)	IPW	−20.356	21.968	68.219	0.012
Heckman et al. (1997)	RA	−31.280	32.087	51.159	0.011
Sant'Anna and Zhao (2020)	DRDiD	−4.402	6.354	20.990	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.387	5.082	25.679	1.141
Sant'Anna and Zhao (2020)*	RF DRDiD	−4.422	7.089	30.706	3.102
Author's work, eq. (27)	3IPWRA	−0.385	3.913	15.167	0.025
Author's work, eq. (27)	LASSO 3IPWRA	−2.553	4.681	15.393	2.142
Author's work, eq. (27)	RF 3IPWRA	0.390	2.893	8.219	1.355
Sant'Anna and Zhao (2020)*	3WDRDiD	−0.285	4.708	22.080	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator



Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 16: Variables included in Γ_i

	Description
diff	If the product have differentiated prices among countries
agri	If the product is an agricultural good
lvalue	The log shipment value per tonnage
perishable	If the product is perishable
largefirm	If the firm has has more than 100 employees
dayarrival	The day of arrival during the week
inspection	If the shipment was pre-inspected at origin
monitor	If the shipment was monitored
SouthAfrica	If the product comes from South Africa
terminal	Terminal of cleareance
hs4group	4-digits Harmonized System (HS) code for product industry classification

Table 17: The effect of tariff reduction on bribes

	TWFE Sequeira (2016)	TWFE ($\Gamma_i \cdot POST$) Sequeira (2016)	DMLDiD (Kernel) Chang (2020)	DMLDiD (lasso) Chang (2020)
ATT	-3.748^{***}	-2.928^{***}	-6.998^{*}	-5.222^{**}
St.Err.	1.075	0.944	3.752	2.647

Notes: TWFE and TWFE($\Gamma_i \times POST$) are eq.1 and eq.2 in Table 9 in Sequeira (2016): the first controls for covariates, while the second adds also the interactions between covariates and the post-treatment dummy. DMLDiD (Kernel) and DMLDiD (lasso) are Column 3 and 5 in Table 2 in Chang (2020). Since the estimator is an IPW method adapted to handle machine-learning first stage estimates, the first uses Kernel in the first stage while the latter employs lasso. The coefficients capture the difference in the log of bribes paid for products that changed tariff level, before and after the tariff change took place. Standard errors are clustered at the level of product's four-digit HS code. Regarding the significance level, *** stands for p-value $p < 0.01$, ** for $p < 0.05$, and * for $p < 0.1$.

Table 18: Monte Carlo simulations for the DMLDiD estimator (Chang, 2020)

	Bias	RMSE	Variance	Time
Experiment 1A	-36.052	71.748	3848.038	28.325
Experiment 1D	-117.857	150.04	8621.81	27.777
Experiment 2A	-31.348	78.794	5225.768	28.178
Experiment 2D	-0.807	90.009	8100.949	27.127

Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. EXP.2 assumes a non-randomized experiment, heterogeneous effects in $X$and time-varying covariates. The letter A refers to the instance where both the propensity score and outcome models are correctly specified, while the letter D to the case of misspecification of both.

Table 19: The effect of tariff reduction on bribes

	TWFE Sequeira (2016)	TWFE($\Gamma_i \cdot POST$) Sequeira (2016)	TWFE($\Gamma_i \cdot POST$+$\Gamma_i \cdot D$)
Coefficient	-3.748^{***}	-2.928^{***}	-3.667^{***}
St.Err.	1.075	0.944	1.071

	lasso 3IPWRA	lasso DRDiD	Random Forest 3IPWRA
Coefficient	-3.023^{***}	-2.764^{***}	-3.216^{***}
St.Err.	0.654	0.905	0.108

Notes: TWFE and TWFE($\Gamma_i \times POST$) are eq.1 and eq.2 in Table 9 in Sequeira (2016): the first controls for covariates, while the second adds also the interactions between covariates and the post-treatment dummy. Instead, TWFE($\Gamma_i \times POST$+$\Gamma_i \times D$) additionally controls for the treatment group and covariates interactions. 3IPWRA (eq. (27)) utilizes both lasso and random forest for first stage estimates, while the doubly robust estimator DRDiD of Sant'Anna and Zhao (2020) is modified to allow for lasso estimates of both the propensity score and outcome regression models. The coefficients capture the difference in the log of bribes paid for products that changed tariff level, before and after the tariff change took place. Standard errors are clustered at the level of product's our-digit HS code and are computed through bootstrap for 3IPWRA and DRDiD. Regarding the significance level, *** stands for p-value $p < 0.01$, ** for $p < 0.05$, and * for $p < 0.1$.

Appendix

A Machine learning first-stage estimates

A.1 Debiased machine learning

The literature on the use of machine learning methods for causal inference is recent and growing. In general, the aim of machine learning methods is to predict Y assuming a model for the predictors X. Since machine learning methods optimize prediction, they are aimed at minimizing the out of sample mean square error (MSE). Since MSE is the sum of the squared bias and the variance of the predictor, the optimum may, and usually does, implicitly allow for some degree of bias. This characteristic makes machine learning estimators not directly applicable to causal inference, where the aim is to obtain unbiased estimates of the causal parameter of interest. However, Chernozhukov et al. (2018) studied a rather flexible approach to employ the potential of machine learning in the field of causal inference. The idea is that in many econometric settings there are intermediate parts of the estimation process that focus on predicting values that are not readily available to the researchers. Chernozhukov et al. (2018) found that, when three main conditions are met, first-stage estimates can be obtained through machine learning predictors without creating bias in the final estimates of the causal parameter.

Suppose we are interested in estimating the causal parameter θ_0 in the presence of nuisance functions g_o and m_0 which depend on high-dimensional functions of the covariates X. For example, Bach et al. (2021) considers a Interactive Regression Model (IRM) in the form:

$$Y = g_0(D, X) + \zeta, \quad E(\zeta|D, X) = 0 \tag{37}$$

$$D = m_0(X) + V, \quad E(V|X) = 0 \tag{38}$$

where Y is the dependent variable, $D \in \{0, 1\}$ is the treatment variable of interest, the high-dimensional vector $X = (X_1, ..., X_p)$ represents the other confounding covariates, and η and V are random errors. In this setting, Equation (37) is the outcome model equation, with the causal parameter of interested being defined as

$\theta_0 = E[g_0(1, X) - g_0(0, X)|D = 1]$, and Equation (38) represents the treatment model. X affects both the the policy variable D, through the function $m_0(X)$, and the outcome variable, via the function $g_0(D, X)$. Such a design generalizes the standard linear regression models, which occurs when both $g_0(D, X)$ and $m_0(X)$ are linear functions of X and D is additively separable. Therefore, machine learning estimators allow for more flexible forms of $g_0(D, X)$ and $m_0(X)$ since they are able to handle the high dimensionality and non-linearity in X.

However, machine learning estimates of the nuisance parameters can be employed only when three conditions are satisfied. The first refers to the score function of the method-of-moments estimator used to infer the casual parameter. Indeed, define the following moment condition:

$$E[\psi(W; \theta_0; \eta)] = 0 \tag{39}$$

where we call ψ is the so-called score function, $W = (Y, D, X)$ is the set of observed variables, θ_0 is the causal parameter, and η denotes nuisance functions with population value η_0.

The first key condition when using machine learning to estimate the nuisance parameter η is employing a score function $\psi(W; \theta_0; \eta)$ that (i) satisfies Equation (39) yielding θ_0 as a unique solution, and (ii) that satisfies the Neyman orthogonality condition defined as:

$$\partial_\eta E[\psi(W; \theta_0; \eta)]|_{\eta=\eta_0} = 0 \tag{40}$$

The Neyman orthogonality expressed in Equation (40) guarantees that the moment condition defined in Equation (39) and utilized to infer θ_0 is insensitive to small perturbations of the nuisance function η when close to η_0. Intuitively, the Neyman orthogonality condition is satisfied when the derivative of the score functions with respect to the parameter η is equal to 0 in the neighborhood of η_0. Since machine learning estimates $\hat{\eta}$ of η are generally biased due to regularization, using a Neyman-orthogonal score eliminates the biases arising from the first-stage estimates.

In the IRM setting, Bach et al. (2021) shows that the IPWRA score function

$$\psi(W;\theta_0;\eta) \equiv \frac{D(Y-g(0,X)}{p} - \frac{m(X)(1-D)(Y-g(0,X)}{p(1-m(X))} - \frac{D}{p}\theta \tag{41}$$
$$\eta = (g,m,p), \quad \eta_0 = (g_0,m_0,p_0), \quad p_0 = P(D=1)$$

satisfies the Neyman orthogonality condition in Equation (40). By substituting Y in Equation (41) with the variation ΔY between pre and post-treatment period and accordingly adjust the outcome model $g(0,X)$, it can be shown that the Equation (41) corresponds to the method of moments version of the DRDiD estimator for panel data proposed by Sant'Anna and Zhao (2020). A similar reasoning applies for the case of repetead cross-sections. As a consequence, in the DRDiD the outcome regression model and the treatment model can be estimated with machine learning estimators without creating bias in the estimates of the causal parameter, as long as other two conditions are matched.

The second condition refers to the rate of convergence of the machine learning estimators used for the nuisance parameters. Formally, in the IRM setting outlined before the machine learning estimators must satisfy:

$$||\hat{m_0} - m_0|| + ||\hat{g_0} - g_0|| \leq o(N^{-1/4}) \tag{42}$$

where $||\cdot||$ indicates the $L^2(P)$ norm operator and $o(\cdot)$ the little-o notation. Chernozhukov et al. (2018) shows that such a condition is generally met by most machine learning estimators such as lasso, ridge, random forests, neural nets, and various hybrids and ensembles of these methods.

Finally, the authors suggest to use a form of sample splitting: the nuisance parameters are estimated on a random partition, while the remaining sample is used for the estimation of the orthogonal score. For instance, when using a 2-fold partition, the dataset is randomly split into two parts, one used for the estimation of the first-stage estimates and the other in the computation of the score function. Such a procedure avoids biases that may arise from the overfitting of the machine-learning estimates. However, as discussed in this paper, this passage is not needed in the case of DRDiD and 3IPWRA since the second stage of the estimation does not employ method of moments.

A.2 Lasso

Lasso is the machine learning method closest to standard linear regression. Following James et al. (2013), consider a regression in the form:

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \epsilon \tag{43}$$

where Y is the outcome variable, $X_1, X_2, ..., X_p$ is the set of covariates, and ϵ is the error term. Assuming that the outcome is linearly related with the predictors, then fitting the least squares to predict the outcome will produce estimates that have low bias. When the number of observations n is much larger than p, least-squares estimates tend to have low variance as well, implying good prediction properties of the estimator. However, in the case p is close to n, then, because of the issue of overfitting, the least-squares fit usually shows high variance, leading to poor predictions out of the training sample. This issue degenerates when $p > n$, since in such a scenario estimates cannot be produced at all since variance becomes infinite. Therefore, a useful approach is to shrink the estimated coefficients to substantially reduce the variance of the estimator when this comes at the cost of a negligible increase in bias. Since this shrinkage, also known as regularization, leads to some of the estimated coefficients to be exactly zero, it can be intuitively interpreted as a form of variable selection.

From a more technical standpoint, lasso coefficients β^L_λ minimize the formulation of the least-squares with a penalty term governed by the parameter λ:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{44}$$

where the first term is the sum of squared residuals (RSS), while the second part, which is multiplied by the tuning parameter $\lambda \geq 0$, is the penalty term. The tuning parameter λ is optimally determined by the use of cross-validation, which is a resampling method that splits the data into test and train portions on different iterations to select the parameter that leads to the lowest MSE. Often, lasso is compared with ridge regression, a similar approach which instead minimizes:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{45}$$

The difference between the two is just the definition of the penalty term: lasso employs the l_1 norm, while ridge regression the l_2 norm. The lasso, when the tuning parameter λ is sufficiently large, leads to some of the coefficient estimates to be exactly equal to zero, which is unlikely in the case of ridge regression. For this reason, the lasso is said to perform variable selection.

A.3 Random forest

The basic idea of tree-based methods for classification and regression is to sequentially segment the predictor space into multiple regions through a recursive binary splitting. As summarized in James et al. (2013), tree-based methods mainly consists of two steps. The first is dividing the set of possible values for $X_1, X_2, ..., X_p$ into J separate and non-overlapping regions, $R_1, R_2, ..., R_J$. The second is, for every observation belonging to region R_j, computing the prediction by taking the mean of the outcome values Y for the training observations in R_j. The aim is therefore to find the regions $R_1, ..., R_J$ that minimize the RSS of:

$$\sum_{j=1}^{J}\sum_{i\in\mathbb{R}_j}^{J}\left(y_i-\hat{y}_{\mathbb{R}_j}\right)^2 \tag{46}$$

where $\hat{y}_{\mathbb{R}_j}$ is the average response for the training observations in the jth region. Since it is not possible to consider each possible partition of the feature space into J boxes, a feasible computational method is recursive binary splitting, which consists of a binary splitting of the predictor space at each step. That is, we consider all predictors $X_1, ..., X_p$, and all possible values of the cutpoint s for each of the predictors, and then select the variable and cutpoint that leads to a tree with lowest RSS. More precisely, for any j and

s, define the pair of half-planes

$$R_1(j,s) = \{X|X_j < s\} \qquad R_2(j,s) = \{X|X_j \geq s\}$$

where the notation for $R_1(j,s)$ indicates the region of predictor space in which X_j takes on a value less than s, and $R_2(j,s)$ for the region characterized by values greater than s. Then, recursive binary splitting looks for the values of j and s that minimize the following equation:

$$\sum_{i:x_i \in \mathbb{R}_1(j,s)} \left(y_i - \hat{y}_{\mathbb{R}_1}\right)^2 + \sum_{i:x_i \in \mathbb{R}_2(j,s)} \left(y_i - \hat{y}_{\mathbb{R}_2}\right)^2 \tag{47}$$

where $\hat{y}_{\mathbb{R}_1}$ represents the average response for the training observations in the region $R_1(j,s)$, and $\hat{y}_{\mathbb{R}_1}$ represents the average response for the training observations in the region $R_2(j,s)$. The process is repeated many times, each time considering the optimal split among the existing regions until a stopping criterion is reached.

To avoid overfitting, tree pruning balances the trade-off between accuracy and complexity of the overall tree. Similarly to regularization, it introduces a penalty term for tuning parameter $\alpha \geq 0$ to improve out-of-sample prediction. In this case, the algorithm minimizes:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in \mathbb{R}_m} \left(y_i - \hat{y}_{\mathbb{R}_m}\right)^2 + \alpha|T| \tag{48}$$

where $|T|$ represents the number of terminal nodes, R_m is the subset of the predictor space corresponding to the mth terminal node, and $\hat{y}_{\mathbb{R}_m}$ is the mean of the training observations in R_m. In this case as well, the tuning parameter α is selected by means of cross-validation.

Random forests is a machine learning estimator that uses decision trees but employs a particular type of bootstrap to drastically reduce the variance of the estimator. Bootstrap is the technique of taking repeated random samples from the training data set and then taking the average of each estimation. Intuitively, bootstrap employs the idea that, given a set of n independent observations $Z_1, ..., Z_n$, each with variance σ^2, the variance of the mean of Z is σ^2/n. So bootstrap when applied to decision trees leads to a final estimator

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}_{bag}(x) \tag{49}$$

where $\hat{f}^{b}_{bag}(x)$ is the prediction obtained from the bth bootstrapped training set and B is the number of repetitions.

However, despite training the trees on different subsets, the predictions are usually highly correlated, limiting the effectiveness of the central limit theorem. Random forest solves this weakness by restricting at each node the split of the predictor space to a random subsample of m predictors instead of the full set of p predictors, with m usually chosen in the order of $\sqrt{p}$. Intuitively, random forest gains prediction efficiency by decorrelating the decision trees and thereby causing the resulting average to have lower variance.

B Sample splitting

B.1 Testing sample splitting in the Monte Carlo simulations

As discussed in the paper, in general Chernozhukov et al. (2018); Bach et al. (2021) suggest to employ sample splitting when using ML first stage estimates: the nuisance (first-stage) parameters are estimated on a random partition of the sample, while the remaining part is used for the estimation of the orthogonal score. For instance, when using a 2-fold partition, the dataset is randomly split into two parts, one used for the estimation of the first-stage estimates and the other in the computation of the score function. Such a procedure avoids biases that may arise from the overfitting. However, we show that there are no benefits of applying sample splitting in the context of the DRDiD and 3IPWRA estimators by testing them in our most indicative settings of the Monte Carlo simulations.

- LASSO DRDID SPLIT uses sample splitting with 4-fold random partition. 75% of the sample is used to estimate with lasso the propensity score and 25% is used in the weighted lasso regression that predicts the values of the outcome model. Then fitted values are calculated for the whole sample and the estimand for the expected value of the ATT is derived.

- RF DRDID SPLIT uses sample splitting with 4-fold random partition. 75% of the sample is used to estimate with random forests the propensity score and 25% is used in the random forest that predicts the outcome model. Then fitted values are calculated for the whole sample and the estimand for the expected value of the ATT is derived.

- 3IPWRA LASSO SPLIT uses sample splitting with 4-fold random partition. 75% of the sample is used to estimate the nuisance parameter of the propensity score and 25% is used in the final weighted least squares regression.

- RF 3IPWRA SPLIT uses sample splitting with 4-fold random partition. 75% of the sample is used to estimate with random forests the nuisance parameter of the propensity score and 25% is used in the final weighted least squares regression.

Table 20 compares the machine learning estimators present in the main text with those employing sample splitting by replicating Experiment 2A and Experiment 2D. In Experiment 2A, sample splitting worsens the performance of each estimator. The bias in the LASSO DRDID moves from -0.018 to 0.253, in the RF DRDID from -4.247 t0 -5.434, and in the LASSO 3IPWRA and RF 3IPWRA from -2.352 to -2.716 and from 0.821 to 2.008 respectively. In Experiment 2D, LASSO DRDID seems to benefit in terms of reduction of bias (from 0.538 to 0.135) but not in terms of RMSE (from 5.004 to 5.280). However, in the other three cases the benefit in terms of bias disappears: for the RF DRDID the bias is almost identical (from -5.131 to -5.104), and for the LASSO 3IPWRA (from -2.517 to -2.720) and RF 3IPWRA (from 0.033 to 0.363) it increases in absolute terms.

In conclusion, we show that not only there are no clear improvements by using sample splitting in the 3IPWRA and DRDiD estimators, but most often sample splitting leads to poorer estimates. In addition, it is important to consider that, when using sample splitting, the computational time required is enlarged by a factor of k in case of k-fold sample splitting. If the standard errors are calculated by bootstrap, this would lead to an increase of computational time in the order of $k \times n$ where n is the number of the repetitions in the bootstrap. We therefore show that sample splitting in the 3IPWRA and DRDiD does not provide any significant improvement

C Robustness checks

C.1 Introduction

In this supplemental appendix, we present a series of robustness checks of the Monte Carlo simulations in Section 4. Our objective is to test alternative specifications of Experiment 1 and Experiment 2, which are the core of the Monte Carlo simulations presented in the paper. Recall that Experiment 1 evaluates a non-randomized experiment with homogeneous treatment effects of the covariates and time-invariant regressors, while Experiment 2 allows for heterogeneous treatment effects and time-varying regressors. Recall also the notation of data generating processes (DPGs) used in the simulations. DPG A assumes that the researcher can correctly specify both the propensity score and outcome models; DPG D assumes that neither of the two models is correctly specified. DPG B and DPG C are intermediate cases that are not commented here because they are less relevant to our evaluation.

C.2 Remove unobserved heterogeneity

As a first robustness check, we remove the term $v(D, X)$ that represents the time-invariant unobserved group heterogeneity between treated and untreated populations. This term is used as in Sant'Anna and Zhao (2020) and represents additional unobserved heterogeneity in the potential outcomes.

In line with our expectations, Table 21 shows that removing $v(D, X)$ in Experiment 1A does not cause any significant change compared to the previous performance analysis of our estimators. The standard TWFE specification is significantly biased (-20.916), but the other estimators perform relatively well in terms of bias. As shown in Table 22, also in Experiment 2A there are no major changes. The only difference is that in this case TWFE with just the interaction of time and covariates works better than including also treatment group and covariates interactions. However, such a result is not robust when changing other parameters, where the double correction works better. This suggests that the best performance of TWFE (T· X) is not structural, and TWFE (T· X+D·X) usually works better in terms of bias. Note also that the presence of $v(D,X)$, as in the original simulations, generates a realistic scenario and therefore is most informative to practitioners.

C.3 Homogeneous effects

The results in Table 23 are obtained by modifying Experiment 2 by assuming that the treatment effects are homogeneous for the covariates X. When the propensity score and outcome models are correctly specified, TWFE (T· X+D· X), the three versions of the 3IPWRA, and the original version of the DRDiD are approximately unbiased. This result, together with those in the following sections, points to the direction that the degree of bias observed in Experiment 2 of the paper, even if relatively contained, of the 3IPWRA and TWFE (T· X+D·X) estimators is mainly caused by the heterogeneity in treatment effects, while compositional changes are not a source of bias. In other words, it seems that these methods correct for compositional changes but are not completely suitable for heterogeneous effects.

C.4 Strong heterogeneous effects

To provide additional evidence on how the heterogeneity in treatment effects influences the degree of bias of our estimators, we increase such heterogeneity to an extreme level (at least with respect to real-case circumstances) as a sort of stress test scenario.

As summarized in Table 24, in the main simulations the heterogeneity in treatment, despite the interquartile range being concentrated between -11.509 and 11.565, varies between a minimum of -66.786 and a maximum of 68.965. The overall distribution is depicted in Figure 4. Knowing that the ATT is approximately zero and the potential outcome for the treated in the post-treatment period is in the order of 500, the degree of heterogeneity in treatment effects is already sizeable.

Nevertheless, we assess the implications of doubling the magnitude of the heterogeneity. To help visualization, Figure 5 displays the density of the ATT among the treated in the new simulation and Table 25 its summary statistics. In this case, the interquartile range varies between −24.710 and 23.430, with a minimum at −148.480 and a maximum at 146.200, suggesting an extreme variability of the treatment effect over different values of the covariates.

By applying these changes to Experiment 2A, Table 26 support the thesis that heterogeneous treatment effects are not completely handled by TWFE corrections, including (T·X+D·X) which reached −13.327 of bias. The 3IPWRA does not fully correct this issue (9.250 of bias), even if it provides a substantial improvement to regression and has a limited degree of bias in its machine learning versions LASSO 3IPWRA (−2.875) and RF 3IPWRA (1.594). In this context, however, the DRDID estimator is approximately unbiased (0.001) and has the best performance. When the propensity score and outcome model are incorrectly specified, as in Table 27, the different versions of the 3IPWRA and DRDID are much closer in bias and the best performance is achieved by the LASSO DRDID (0.795).

C.5 Alternative functional form for the relationship between X and Z

In this section, we test alternative functional forms for the relationship that links the observed vector of covariates Z to X, which is the vector of true regressors affecting the outcome variable when the propensity score and outcome model are incorrectly specified.

Following the main simulations setup, we define $X = (X_1, X_2, X_3, X_4)'$ as being distributed as $N(0, I_4)$ with I_4 representing the 4 × 4 identity matrix. In this robustness check, we assume an alternative functional form for the relationship between Z and X. For $j = 1, 2, 3, 4$ define the standardized variable $Z_j = (\tilde{Z}_j - E[\tilde{Z}_j])/\sqrt{Var(\tilde{Z}_j)}$ where $\tilde{Z}_1 = \exp(0.5X_1X_2)$, $\tilde{Z}_2 = 12 + X_1/(1 + \exp(X_3))$, $\tilde{Z}_3 = (|X_2X_4| + X_1X_3/25)^3$, and $\tilde{Z}_4 = 0.1(ln(|40 + X_1 + X_4|))^2$. The vector $Z = (Z_1, Z_2, Z_3, Z_4)'$ is the set of variables that are observed by the researcher. In addition, we assume homogeneity in treatment effects to isolate the effect of the alternative functional form.

The results of applying these changes to Experiment 2A, which are shown in Table 28, confirm the main results of the paper. TWFE, IPW, and RA are all severely biased (−185.916, −101.557 and −234.339 respectively), while TWFE (T· X+D·X), 3IPWRA and DRDiD are approximately unbiased (0.015, −0.012 and 0.027 respectively). In this case, the LASSO versions of both 3IPWRA and DRDiD, despite having a higher but limited degree of bias when both the propensity score and outcome models are correctly specified, outperform the standard 3IPWRA and DRDiD when both models are misspecified (see Table 29 which reports the modified version of Experiment 2D). More precisely, the bias is 4.209 for the LASSO DRDiD (compared to 5.982 of DRDiD) and −5.366 for LASSO 3IPWRA (compared to 6.653 of 3IPWRA).

C.6 Non-linear outcome model

An additional robustness test is to specify a new functional form for the outcome model. In this case, we specify the following function:

$$f_{reg}(W) = 210 + 37.4 \cdot W_1 W_2 + 10 \cdot W_3 W_4$$

As in Appendix C.5, to isolate the different sources of bias, we assume that homogeneity in treatment effects of the covariates holds, which is another variation from the main simulations' Experiment 2.

As visible in Table 30, the results of our baseline simulations are confirmed: the bias of TWFE (D·X) and TWFE (T· X+D·X), which is −5.315 and 6.059 respectively, is contained compared to the standard TWFE specification (−20.750); however, these methods are clearly outperformed by the DRDiD (0.193)and 3IPWRA (0.751) and its machine learning alternative versions. When the propensity score and outcome models are misspecified, as in Table 31, the lowest bias is achieved by RF 3IPWRA (−0.003) and LASSO DRDiD (−0.383).

C.7 Non-linear trend model

Similarly to Appendix C.6, we modify the functional form specification of another factor, namely the trend. We assume a non-linear trend and we assume homogeneity in treatment effects. In this case, the trend τ is:

$$\tau = 210 + 27.4 \cdot W_1^2 + 19.5 \cdot W_1 W_2 W_3$$

Such changes are then applied to Experiment 2. Table 32 confirms the results in the previous two sections. TWFE (T· X+D·X) reduce the bias of the traditional TWFE (-5.599 versus -18.535), but the 3IPWRA (0.522) and DRDiD(-0.065), including some of their machine learning versions, clearly outperform the other alternatives. Similarly to previous sections, machine learning first-stage estimators, in particular the LASSO DRDiD (-0.312 of bias), perform better in terms of bias when both the propensity score and outcome models are misspecified, as in Table 33.

C.8 Strong compositional changes

Finally, we propose a robustness check that increases the heterogeneity in the time dimension so that there are strong differences in the distribution of the covariates between the pre-and post-treatment periods. Recall that the IPW, RA and DRDiD methods are not built to handle time-varying covariates. In the main simulations, compositional changes are achieved by defining a propensity score that makes it more likely for individuals with some specific covariate characteristics to belong to the post-treatment period. To increase the difference in the covariate distribution between $t = 0$ and $t = 1$, we test the implications of adding a drift term: in addition to selection through a post-treatment-period propensity score, we make the whole distribution of covariates move between the two periods. In practice, this is achieved by summing a constant term to the distribution of covariates in the post-treatment period, with the drift term being a stochastic variable with a normal distribution. The vector Z is therefore now specified as follows:

$$
\begin{aligned}
Z^{new}_{1|t=1} &= Z_{1|t=1} + N(0.8, 0.05)\\
Z^{new}_{2|t=1} &= Z_{2|t=1} + N(-0.3, 0.02)\\
Z^{new}_{3|t=1} &= Z_{3|t=1} + N(0.9, 0.03)\\
Z^{new}_{4|t=1} &= Z_{4|t=1} + N(1.2, 0.08)
\end{aligned}
$$

Where $Z_{1|t=1}$ refers to the distribution of Z in the post-treatment period after the selection operated by the post-treatment-period propensity score.

We start from the case where, in addition to assuming homogeneous treatment effects, we impose non-linear trends. The results in Table 34 show that, while the IPW and RA methods reach huge biases even when the propensity score and outcome model are correctly specified (28.569 and 79.758 respectively), DRDiD seems quite robust to compositional changes since it is characterized by a bias (-0.020) comparable to the one of 3IPWRA (-0.020), LASSO 3IPWRA (0.023), RF 3IPWRA (0.021) and TWFE (T· X+D·X) (0.020).

C.9 Robustness checks outcomes

The robustness checks present in this appendix confirm the results of the main paper. TWFE is in various settings severely biased, even if the TWFE (T· X+D·X) correction partially alleviates such an issue. However, in general, these methods together with the IPW (Abadie, 2005) and RA (Heckman et al., 1997), are consistently outperformed by the semi-parametric estimators DRDiD (Sant'Anna and Zhao, 2020) and 3IPWRA. We, therefore, recommend preferring the use of these two classes of estimators over the others. However, these methods have certain limitations: while the 3IPWRA does not handle well heterogeneity in treatment effects, DRDiD may be biased in the case of time-varying covariates. Therefore, a practical empirical strategy may be to use both methods and check if the results converge. If this is the case, it could be an indication that our empirical findings are robust to such complications. In addition, we recommend considering the use of machine learning versions of the 3IPWRA and DRDiD estimators in real practice, since they tend to have a lower bias when both the propensity score and the outcome model are misspecified, especially the LASSO ones.

D Appendix Figures

Figure 4: Density plot ATT main paper

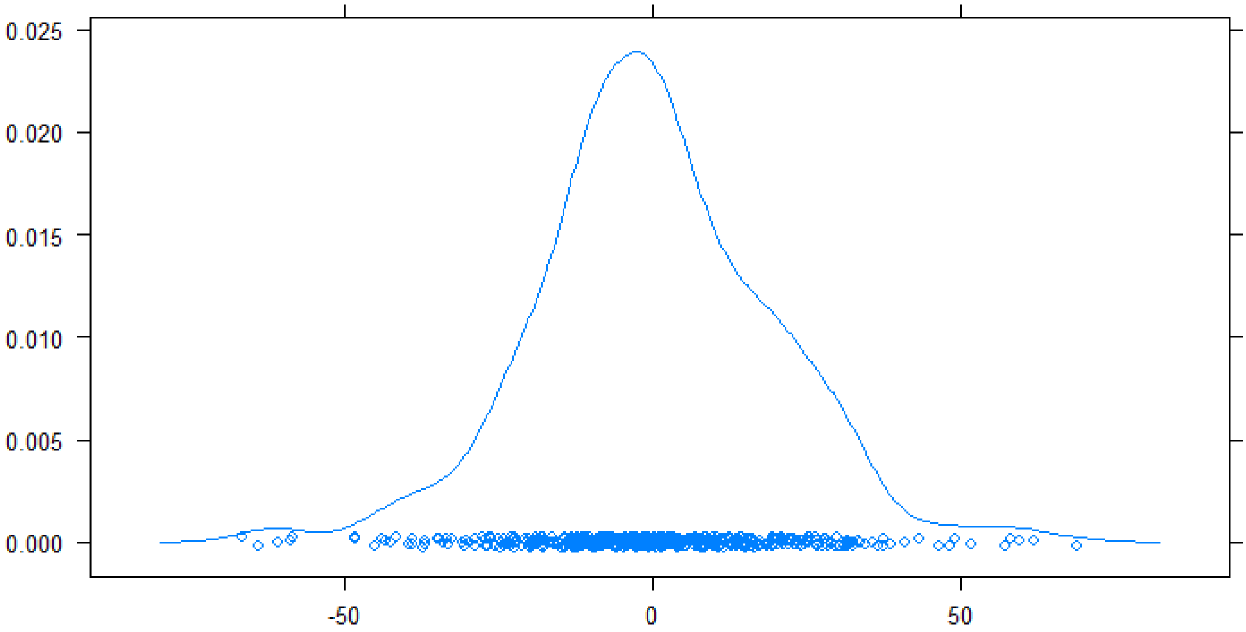
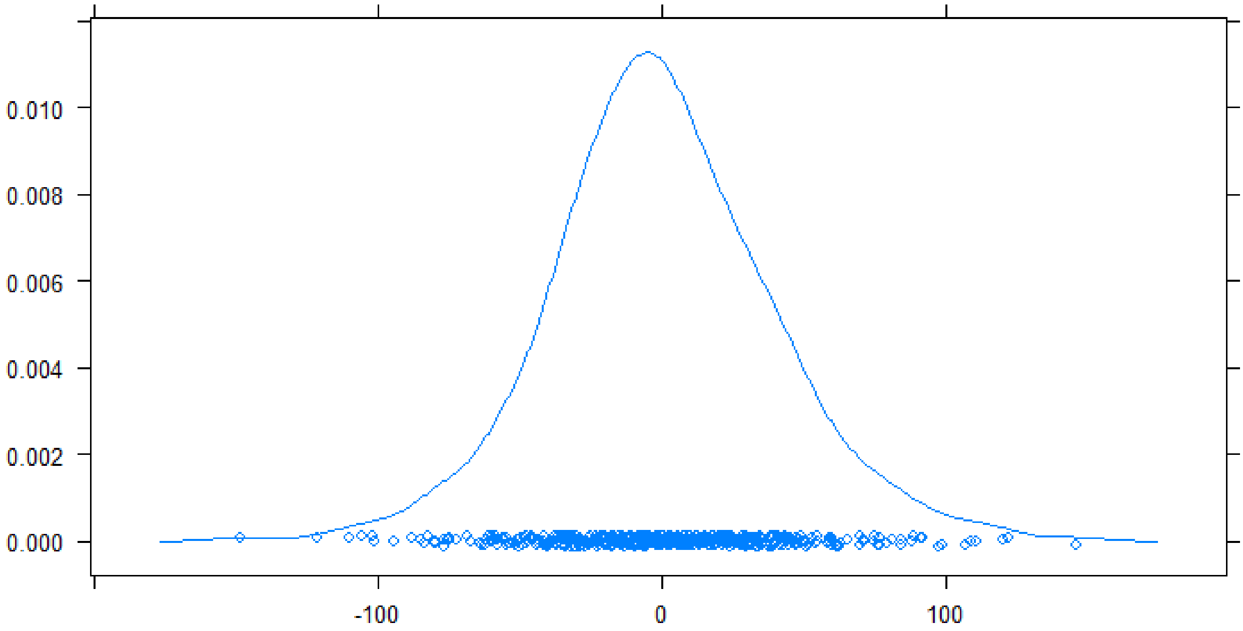


Figure 5: Density plot ATT with strong heterogeneity in treatment effects



E Appendix Tables

Table 20: EXP 2A (Propensity score model correct, outcome regression model correct) including sample-split estimators

Reference	Estimator	Bias	RMSE	Variance	Time
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.018	0.187	0.035	1.339
Sant'Anna and Zhao (2020)*	LASSO DRDiD SPLIT	0.253	0.254	0.001	4.678
Sant'Anna and Zhao (2020)*	RF DRDiD	−4.247	6.734	27.306	3.112
Sant'Anna and Zhao (2020)*	RF DRDiD SPLIT	−5.434	8.214	37.94	6.671
Author's work, eq. (27)	LASSO 3IPWRA	−2.352	2.396	0.205	2.392
Author's work, eq. (27)	LASSO 3IPWRA SPLIT	−2.716	2.724	0.047	9.507
Author's work, eq. (27)	RF 3IPWRA	0.821	1.405	1.301	1.383
Author's work, eq. (27)	RF 3IPWRA SPLIT	2.008	2.396	1.708	1.683

EXP 2D (Propensity score model correct, outcome regression model correct) including sample-split estimators

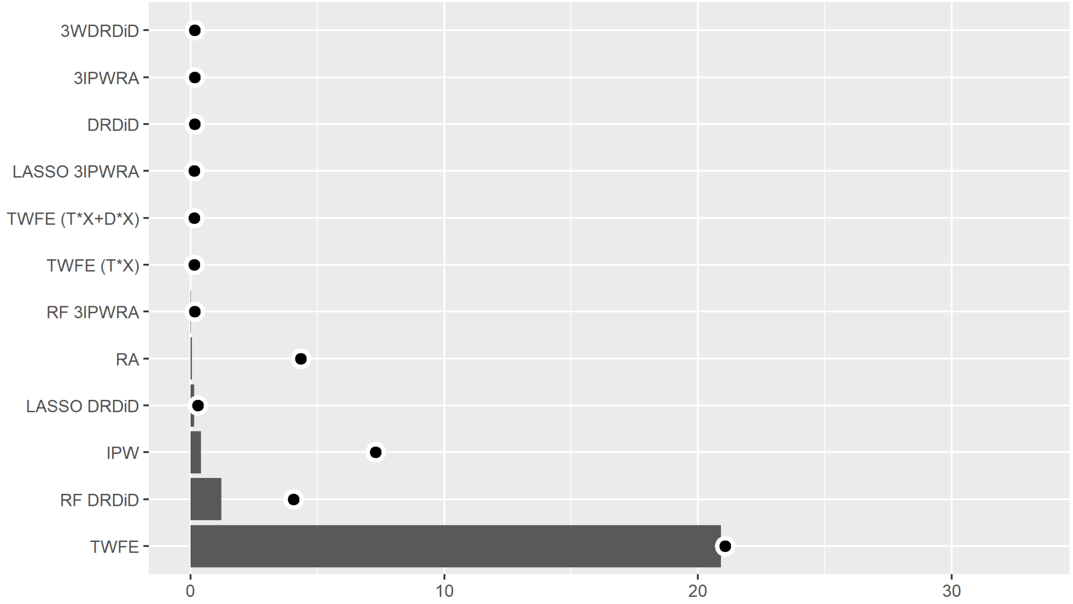
Reference	Estimator	Bias	RMSE	Variance	Time
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.538	5.004	24.754	1.146
Sant'Anna and Zhao (2020)*	LASSO DRDiD SPLIT	0.135	5.280	27.861	4.528
Sant'Anna and Zhao (2020)*	RF DRDiD	−5.131	7.536	30.472	3.047
Sant'Anna and Zhao (2020)*	RF DRDiD SPLIT	−5.104	8.302	42.867	6.517
Author's work, eq. (27)	3LASSO 3IPWRA	−2.517	4.547	14.336	2.079
Author's work, eq. (27)	3LASSO 3IPWRA SPLIT	−2.720	4.765	15.310	7.958
Author's work, eq. (27)	3RF 3IPWRA	0.033	2.981	8.887	1.378
Author's work, eq. (27)	3RF 3IPWRA SPLIT	0.363	3.624	13.000	1.632

Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. Experiment 2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. DRDiD is the doubly robust estimator and it proposed either employing lasso or random forest in the estimation of the propensity score and outcome regression. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with lasso and random forest (eq. (27)). The term "SPLIT" refers to the use of sample splitting. Finally, 'Bias', 'RMSE', 'Variance' and 'Time', stand for the average simulated bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 21: EXP 1A (Propensity score model correct, outcome regression model correct) without unobserved heterogeneity

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−20.916	21.068	6.383	0.002
Regression, eq. (11)	TWFE (T·X)	−0.005	0.138	0.019	0.002
Regression, eq. (12)	TWFE (T·X+D·X)	−0.005	0.138	0.019	0.002
Abadie (2005)	IPW	−0.407	7.293	53.022	0.014
Heckman et al. (1997)	RA	0.042	4.335	18.787	0.010
Sant'Anna and Zhao (2020)	DRDiD	−0.003	0.153	0.023	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.134	0.278	0.059	1.011
Sant'Anna and Zhao (2020)*	RF DRDiD	−1.210	4.052	14.953	3.188
Author's work, eq. (27)	IPWRA	−0.002	0.152	0.023	0.025
Author's work, eq. (27)	LASSO 3IPWRA	−0.004	0.142	0.020	1.860
Author's work, eq. (27)	RF 3IPWRA	−0.006	0.152	0.023	1.256
Sant'Anna and Zhao (2020)*	3WDRDiD	−0.002	0.153	0.023	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

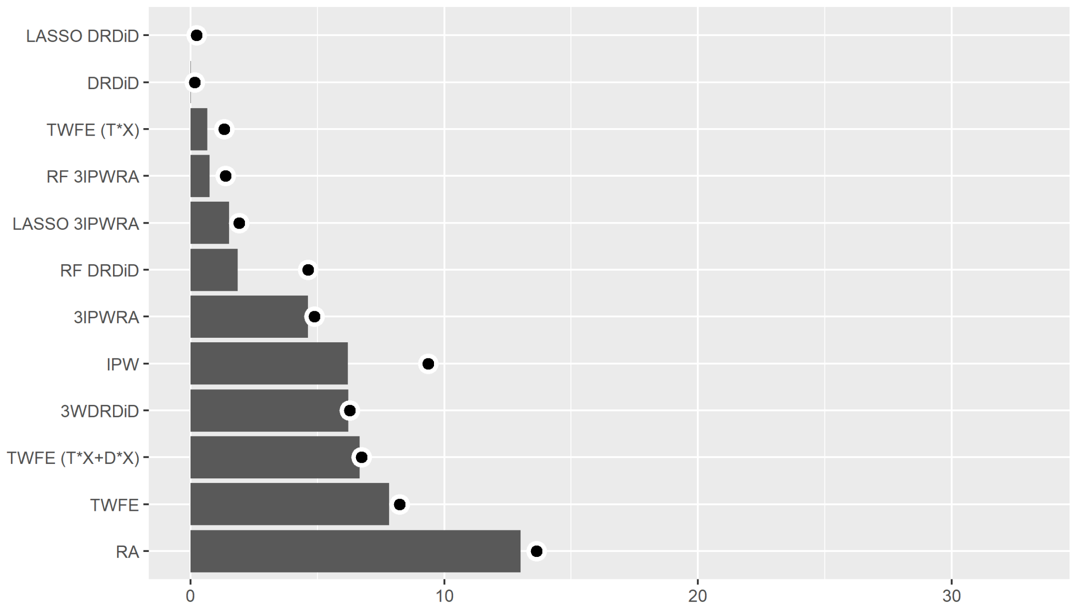


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.1 assumes a non-randomized experiment, homogeneous effects in X and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 22: EXP 2A (Propensity score model correct, outcome regression model correct) without unobserved heterogeneity

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	-7.832	8.243	6.609	0.001
Regression, eq. (11)	TWFE (T·X)	-0.659	1.322	1.312	0.002
Regression, eq. (12)	TWFE (T·X+D·X)	-6.656	6.729	0.982	0.002
Abadie (2005)	IPW	6.204	9.365	49.206	0.014
Heckman et al. (1997)	RA	-13.014	13.634	16.521	0.011
Sant'Anna and Zhao (2020)	DRDiD	0.006	0.160	0.026	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.001	0.233	0.054	1.029
Sant'Anna and Zhao (2020)*	RF DRDiD	-1.855	4.621	17.915	3.176
Author's work, eq. (27)	3IPWRA	4.623	4.874	2.380	0.025
Author's work, eq. (27)	LASSO 3IPWRA	-1.517	1.904	1.327	2.133
Author's work, eq. (27)	RF 3IPWRA	0.746	1.381	1.351	1.377
Sant'Anna and Zhao (2020)*	3WDRDiD	6.214	6.262	0.601	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

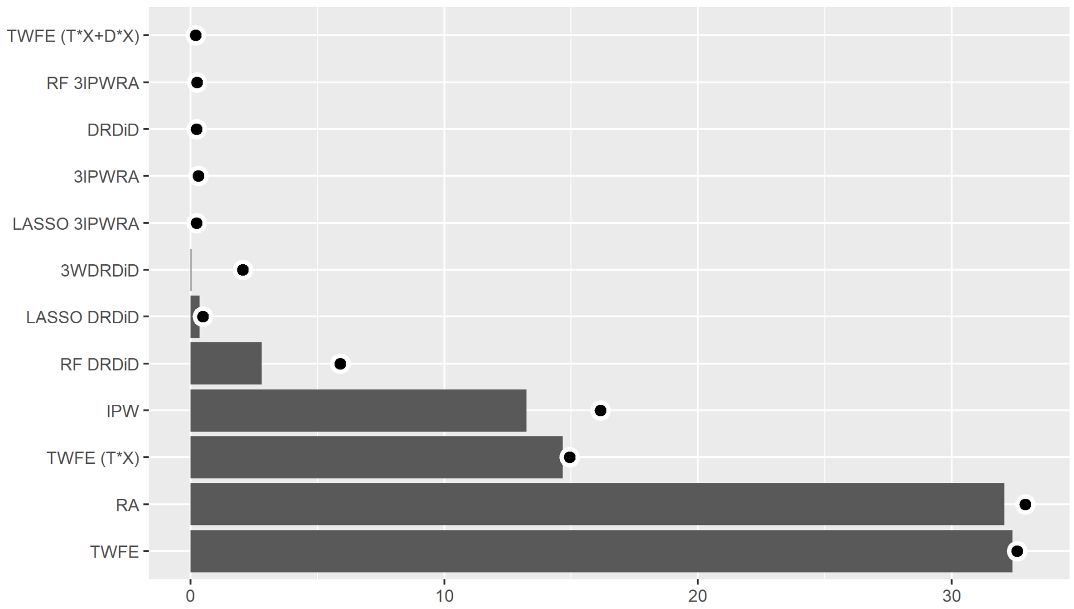


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 23: EXP 2A (Propensity score model correct, outcome regression model correct) with homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−32.408	32.574	10.814	0.001
Regression, eq. (11)	TWFE (T·X)	−14.666	14.944	8.245	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.002	0.192	0.037	0.001
Abadie (2005)	IPW	−13.238	16.161	85.922	0.014
Heckman et al. (1997)	RA	−32.078	32.898	53.237	0.010
Sant'Anna and Zhao (2020)	DRDiD	−0.003	0.225	0.051	0.015
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.352	0.475	0.101	1.033
Sant'Anna and Zhao (2020)*	RF DRDiD	−2.808	5.891	26.818	3.175
Author's work, eq. (27)	3IPWRA	−0.004	0.300	0.090	0.025
Author's work, eq. (27)	LASSO 3IPWRA	−0.005	0.229	0.053	2.159
Author's work, eq. (27)	RF 3IPWRA	−0.003	0.252	0.064	1.380
Sant'Anna and Zhao (2020)*	3WDRDiD	−0.034	2.047	4.191	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator



Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 24: Summary statistics ATT main paper

Min.	-66.786
1st Qu.	-11.509
Median	-1.153
Mean	-0
3rd Qu.	11.565
Max.	68.965

Notes: the descriptive statistics are derived by one realization of the 500 repetitions in the Monte Carlo simulations. They may therefore slightly vary for other realizations.

Table 25: Summary statistics of ATT in presence of strong heterogeneity in treatment effects

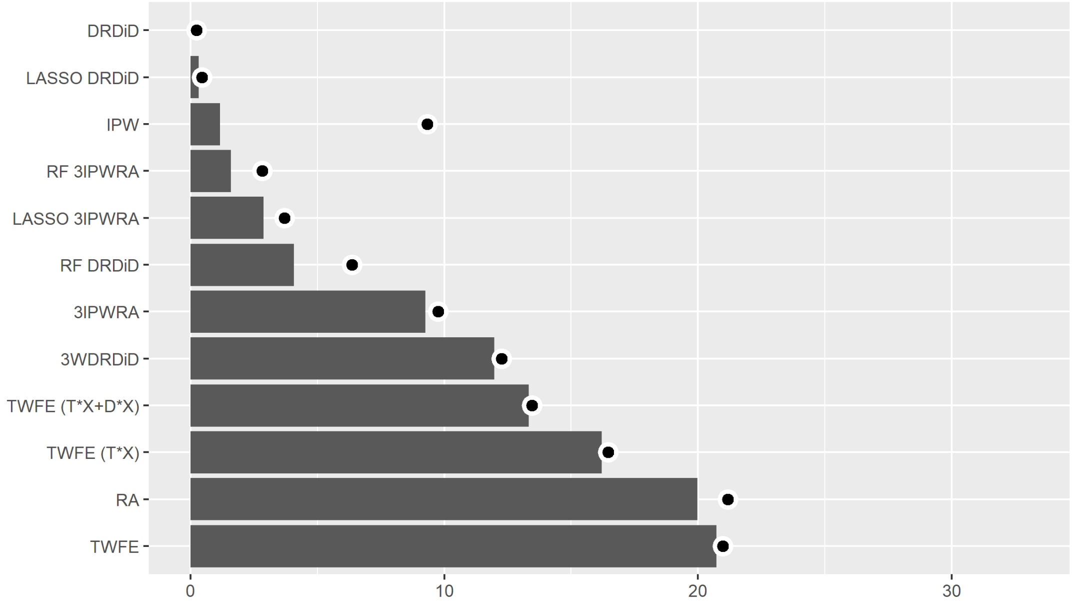
Min.	-148.480
1st Qu.	-24.710
Median	-1.200
Mean	0
3rd Qu.	23.430
Max.	146.200

Notes: the descriptive statistics are derived by one realization of the 500 repetitions in the Monte Carlo simulations. They may therefore slightly vary for other realizations.

Table 26: EXP 2A (Propensity score model correct, outcome regression model correct) with strong heterogeneity in effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−20.724	20.990	11.125	0.001
Regression, eq. (11)	TWFE (T·X)	−16.213	16.461	8.094	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−13.327	13.452	3.331	0.002
Abadie (2005)	IPW	−1.155	9.334	85.798	0.011
Heckman et al. (1997)	RA	−19.982	21.185	49.536	0.008
Sant'Anna and Zhao (2020)	DRDiD	0.001	0.225	0.051	0.013
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.318	0.440	0.093	1.010
Sant'Anna and Zhao (2020)*	RF DRDiD	−4.071	6.357	23.836	3.163
Author's work, eq. (27)	3IPWRA	9.250	9.753	9.557	0.031
Author's work, eq. (27)	LASSO 3IPWRA	−2.875	3.693	5.372	2.113
Author's work, eq. (27)	RF 3IPWRA	1.594	2.813	5.373	1.379
Sant'Anna and Zhao (2020)*	3WDRDiD	11.969	12.257	6.975	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

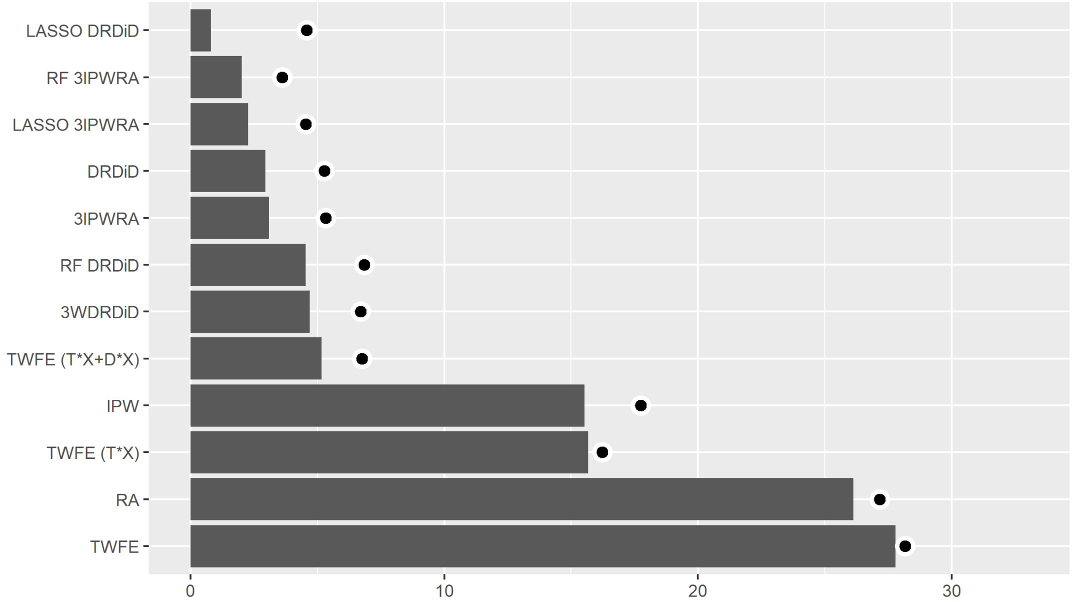


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 27: EXP 2D (Propensity score model incorrect, outcome regression model incorrect) with strong heterogeneity in effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	-27.793	28.163	20.687	0.001
Regression, eq. (11)	TWFE (T·X)	-15.674	16.226	17.629	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	-5.171	6.755	18.888	0.001
Abadie (2005)	IPW	-15.527	17.752	74.057	0.010
Heckman et al. (1997)	RA	-26.132	27.162	54.891	0.007
Sant'Anna and Zhao (2020)	DRDiD	-2.947	5.262	19.009	0.013
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.795	4.573	20.284	1.152
Sant'Anna and Zhao (2020)*	RF DRDiD	-4.539	6.838	26.151	3.123
Author's work, eq. (27)	3IPWRA	3.091	5.319	18.735	0.026
Author's work, eq. (27)	LASSO 3IPWRA	-2.269	4.543	15.490	2.125
Author's work, eq. (27)	RF 3IPWRA	2.015	3.602	8.915	1.340
Sant'Anna and Zhao (2020)*	3WDRDiD	4.705	6.706	22.831	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

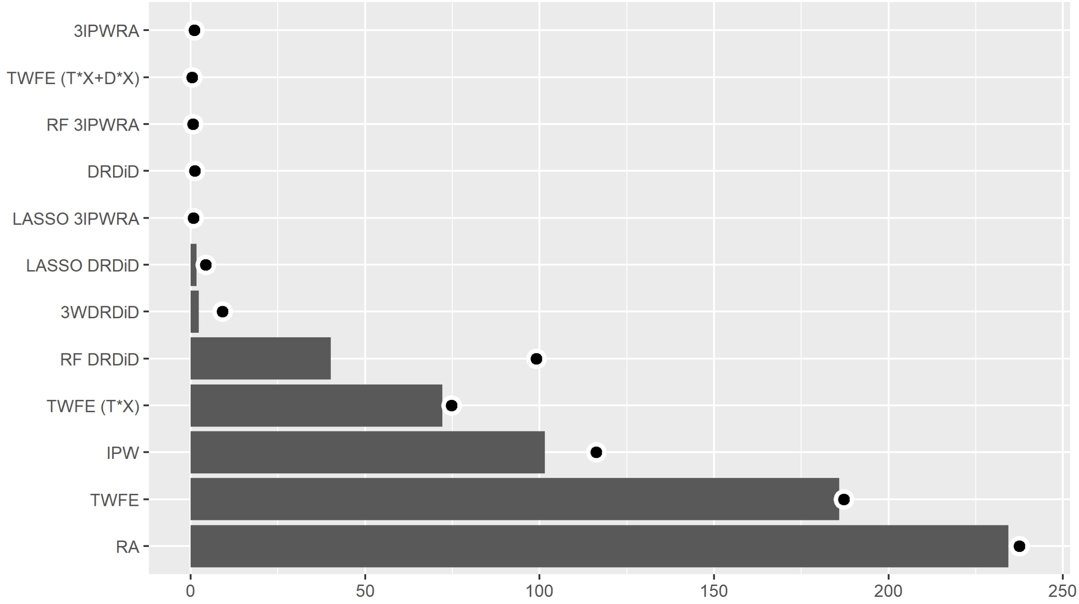


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 28: EXP 2A (Propensity score model correct, outcome regression model correct) with alternative functional form for the relationship between X and Z and homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−185.916	187.140	456.800	0.001
Regression, eq. (11)	TWFE (T·X)	−72.094	74.754	390.587	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.015	0.317	0.101	0.002
Abadie (2005)	IPW	−101.557	116.178	3, 183.341	0.011
Heckman et al. (1997)	RA	−234.339	237.490	1, 486.579	0.007
Sant'Anna and Zhao (2020)	DRDiD	0.027	1.106	1.222	0.021
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−1.677	4.319	15.843	3.997
Sant'Anna and Zhao (2020)*	RF DRDiD	40.202	99.088	8, 202.145	3.093
Author's work, eq. (27)	3IPWRA	−0.012	1.021	1.043	0.021
Author's work, eq. (27)	LASSO 3IPWRA	0.040	0.812	0.658	11.487
Author's work, eq. (27)	RF 3IPWRA	0.019	0.573	0.328	2.982
Sant'Anna and Zhao (2020)*	3WDRDiD	−2.301	9.118	77.845	0.018

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

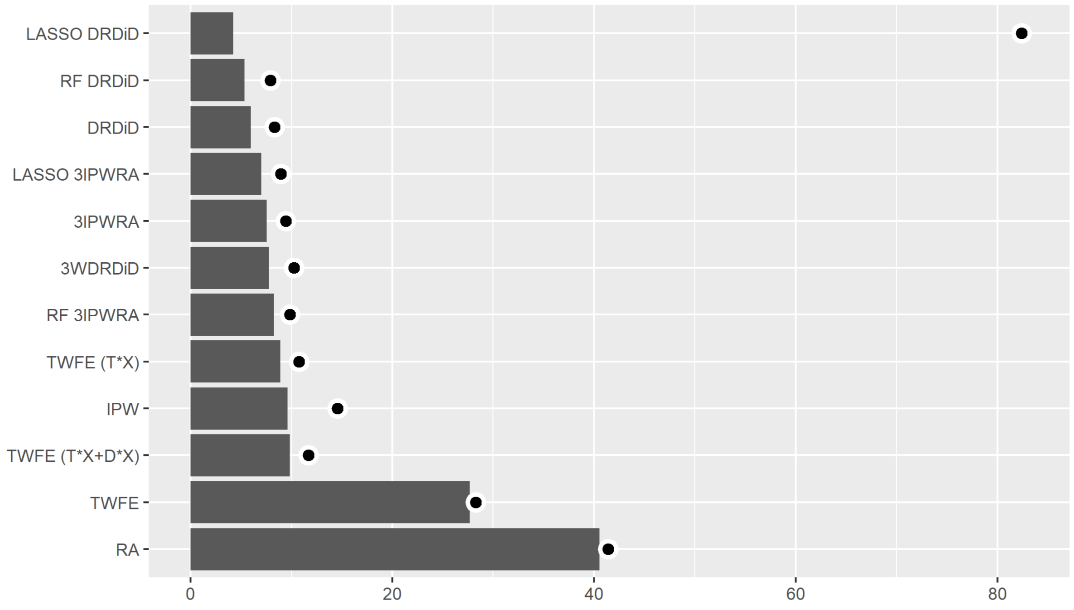


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 29: EXP 2D (Propensity score model incorrect, outcome regression model incorrect) with other functional forms of X and Z and homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	-27.667	28.265	33.464	0.001
Regression, eq. (11)	TWFE (T·X)	-8.904	10.731	35.859	0.002
Regression, eq. (12)	TWFE (T·X+D·X)	9.843	11.698	39.964	0.001
Abadie (2005)	IPW	-9.619	14.546	119.070	0.010
Heckman et al. (1997)	RA	-40.539	41.386	69.384	0.008
Sant'Anna and Zhao (2020)	DRDiD	5.982	8.322	33.474	0.022
Sant'Anna and Zhao (2020)*	LASSO DRDiD	4.209	82.391	6,770.513	4.937
Sant'Anna and Zhao (2020)*	RF DRDiD	5.334	7.921	34.286	3.129
Author's work, eq. (27)	3IPWRA	6.653	10.783	72.013	0.013
Author's work, eq. (27)	LASSO 3IPWRA	-5.366	8.199	38.429	10.756
Author's work, eq. (27)	RF 3IPWRA	11.803	13.827	51.862	1.360
Sant'Anna and Zhao (2020)*	3WDRDiD	7.774	10.250	44.625	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

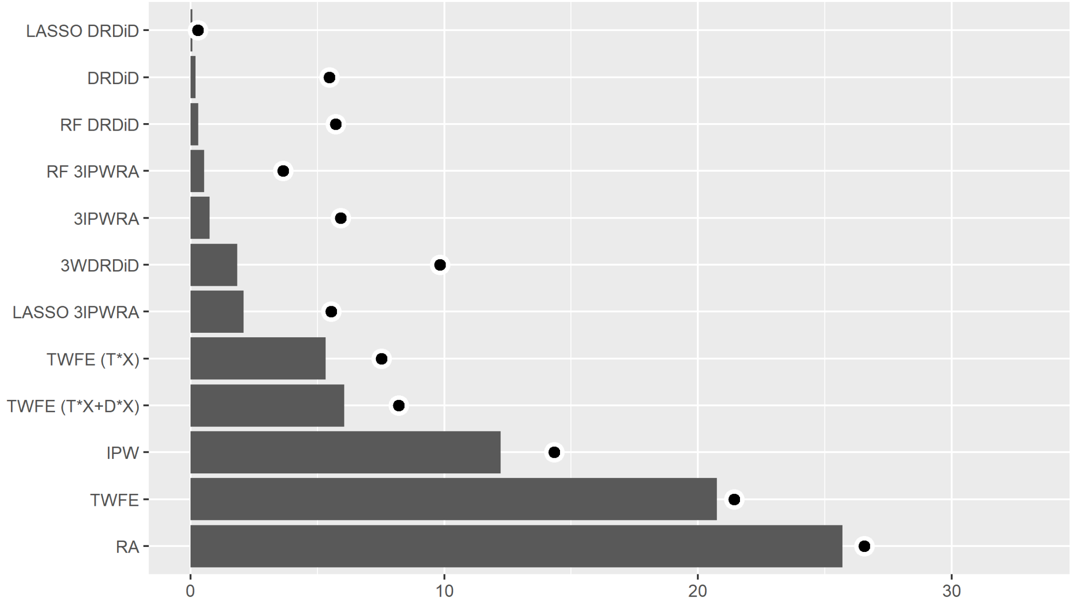


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 30: EXP 2A (Propensity score model correct, outcome regression model correct) with non-linear outcome model

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−20.750	21.431	28.712	0.001
Regression, eq. (11)	TWFE (T·X)	−5.315	7.525	28.387	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	6.059	8.195	30.436	0.001
Abadie (2005)	IPW	−12.225	14.334	56.000	0.013
Heckman et al. (1997)	RA	−25.692	26.557	45.196	0.010
Sant'Anna and Zhao (2020)	DRDiD	0.193	5.468	29.862	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.069	0.290	0.080	1.100
Sant'Anna and Zhao (2020)*	RF DRDiD	0.296	5.716	32.583	3.463
Author's work, eq. (27)	3IPWRA	0.751	5.922	34.500	0.026
Author's work, eq. (27)	LASSO 3IPWRA	2.094	5.543	26.333	2.284
Author's work, eq. (27)	RF 3IPWRA	0.530	3.639	12.959	1.501
Sant'Anna and Zhao (2020)*	3WDRDiD	1.844	9.821	93.059	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

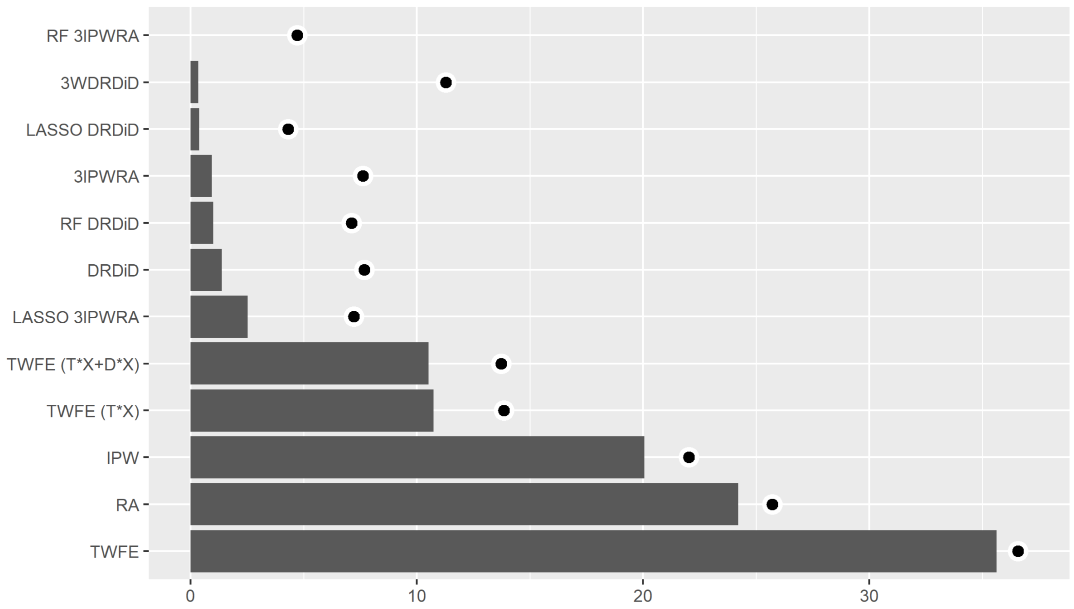


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 31: EXP 2D (Propensity score model incorrect, outcome regression model incorrect) with non-linear outcome model

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	-35.624	36.576	68.769	0.001
Regression, eq. (11)	TWFE (T·X)	-10.738	13.853	76.600	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	10.525	13.729	77.725	0.002
Abadie (2005)	IPW	-20.064	22.020	82.323	0.013
Heckman et al. (1997)	RA	-24.205	25.719	75.604	0.011
Sant'Anna and Zhao (2020)	DRDiD	-1.384	7.678	57.038	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	-0.383	4.309	18.420	1.171
Sant'Anna and Zhao (2020)*	RF DRDiD	-1.002	7.116	49.629	3.218
Author's work, eq. (27)	3IPWRA	-0.948	7.611	57.028	0.031
Author's work, eq. (27)	LASSO 3IPWRA	2.514	7.208	45.631	2.113
Author's work, eq. (27)	RF 3IPWRA	-0.003	4.706	22.148	1.391
Sant'Anna and Zhao (2020)*	3WDRDiD	0.328	11.287	127.300	0.019

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

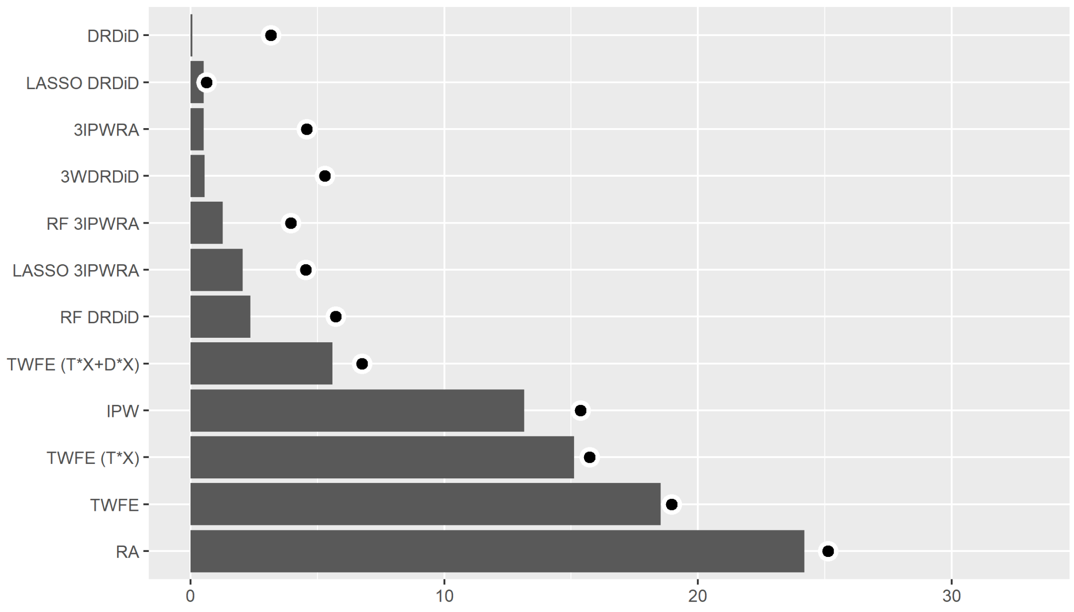


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 32: EXP 2A (Propensity score model correct, outcome regression model correct) with non-linear trend

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−18.535	18.963	16.069	0.001
Regression, eq. (11)	TWFE (T·X)	−15.125	15.732	18.733	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−5.599	6.755	14.286	0.001
Abadie (2005)	IPW	−13.149	15.376	63.526	0.010
Heckman et al. (1997)	RA	−24.200	25.137	46.214	0.008
Sant'Anna and Zhao (2020)	DRDiD	−0.065	3.156	9.956	0.017
Sant'Anna and Zhao (2020)*	LASSO DRDiD	−0.509	0.617	0.121	1.103
Sant'Anna and Zhao (2020)*	RF DRDiD	−2.355	5.721	27.177	3.411
Author's work, eq. (27)	3IPWRA	0.522	4.571	20.620	0.019
Author's work, eq. (27)	LASSO 3IPWRA	−2.049	4.529	16.317	2.247
Author's work, eq. (27)	RF 3IPWRA	−1.270	3.953	14.014	1.508
Sant'Anna and Zhao (2020)*	3WDRDiD	0.554	5.289	27.662	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

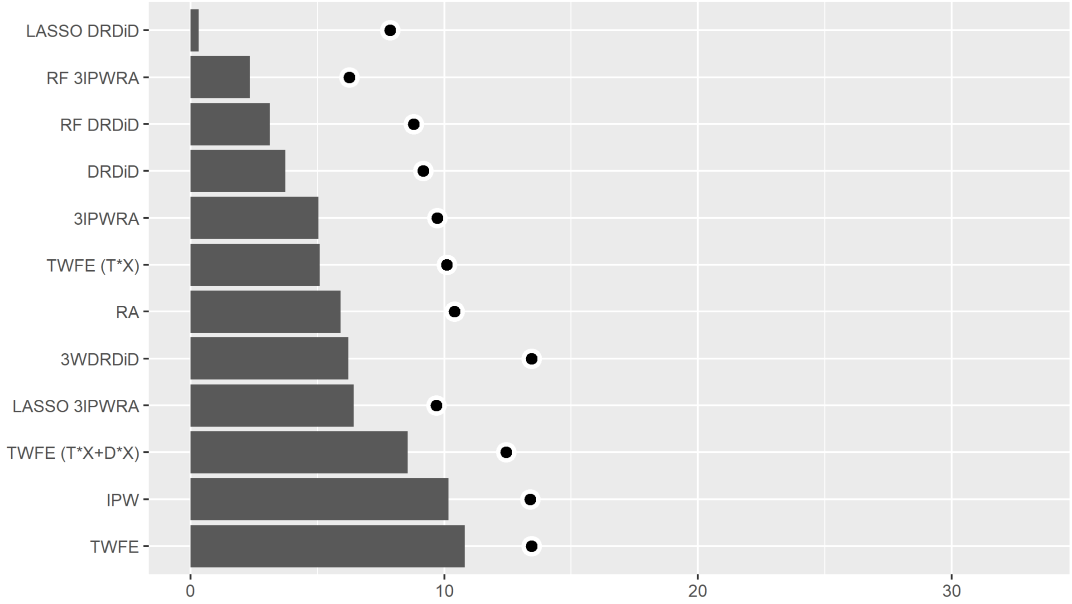


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 33: EXP 2D (Propensity score model incorrect, outcome regression model incorrect) with non-linear trend

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	-10.803	13.439	63.905	0.001
Regression, eq. (11)	TWFE (T·X)	-5.095	10.103	76.121	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	8.554	12.442	81.623	0.002
Abadie (2005)	IPW	-10.168	13.388	75.833	0.011
Heckman et al. (1997)	RA	-5.921	10.398	73.049	0.007
Sant'Anna and Zhao (2020)	DRDiD	3.730	9.175	70.273	0.020
Sant'Anna and Zhao (2020)*	LASSO DRDiD	-0.312	7.855	61.608	1.297
Sant'Anna and Zhao (2020)*	RF DRDiD	3.125	8.798	67.634	3.608
Author's work, eq. (27)	3IPWRA	5.036	9.722	69.157	0.026
Author's work, eq. (27)	LASSO 3IPWRA	6.433	9.678	52.281	2.317
Author's work, eq. (27)	RF 3IPWRA	2.330	6.248	33.609	1.550
Sant'Anna and Zhao (2020)*	3WDRDiD	6.213	13.442	142.096	0.021

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

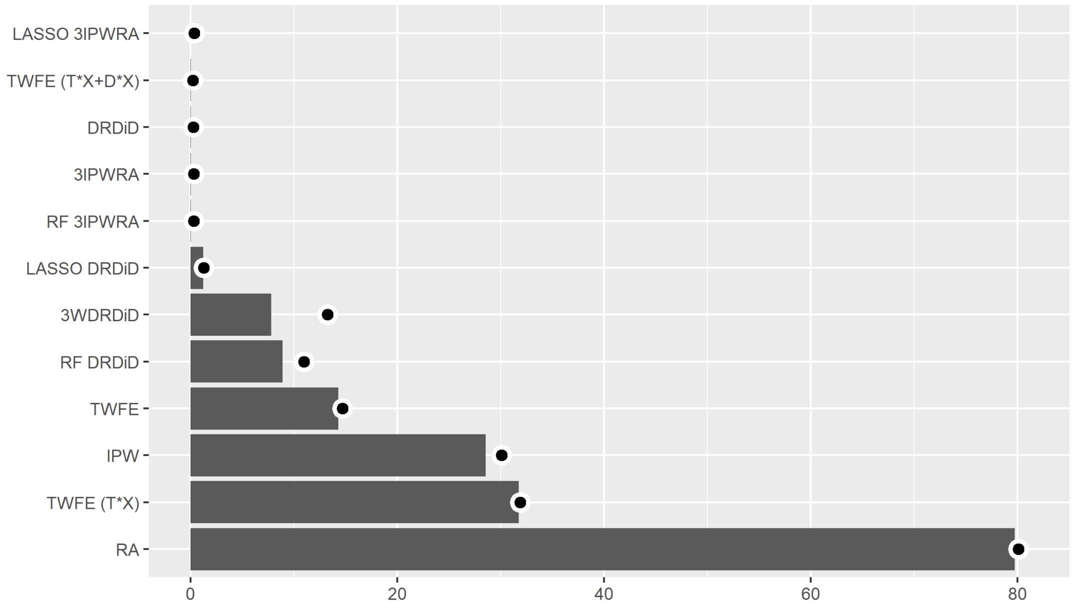


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 34: EXP 2A (Propensity score model correct, outcome regression model correct) with strong compositional changes, linear trend, and homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	14.286	14.674	11.221	0.001
Regression, eq. (11)	TWFE (T·X)	31.772	31.899	8.085	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.020	0.228	0.051	0.001
Abadie (2005)	IPW	28.569	30.070	88.038	0.014
Heckman et al. (1997)	RA	79.758	80.117	57.440	0.011
Sant'Anna and Zhao (2020)	DRDiD	0.020	0.255	0.064	0.016
Sant'Anna and Zhao (2020)*	LASSO DRDiD	1.213	1.278	0.164	1.063
Sant'Anna and Zhao (2020)*	RF DRDiD	8.911	10.949	40.471	3.097
Author's work, eq. (27)	3IPWRA	−0.020	0.305	0.093	0.032
Author's work, eq. (27)	LASSO 3IPWRA	0.023	0.304	0.092	0.682
Author's work, eq. (27)	RF 3IPWRA	0.021	0.302	0.090	1.354
Sant'Anna and Zhao (2020)*	3WDRDiD	−7.791	13.235	114.471	0.017

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

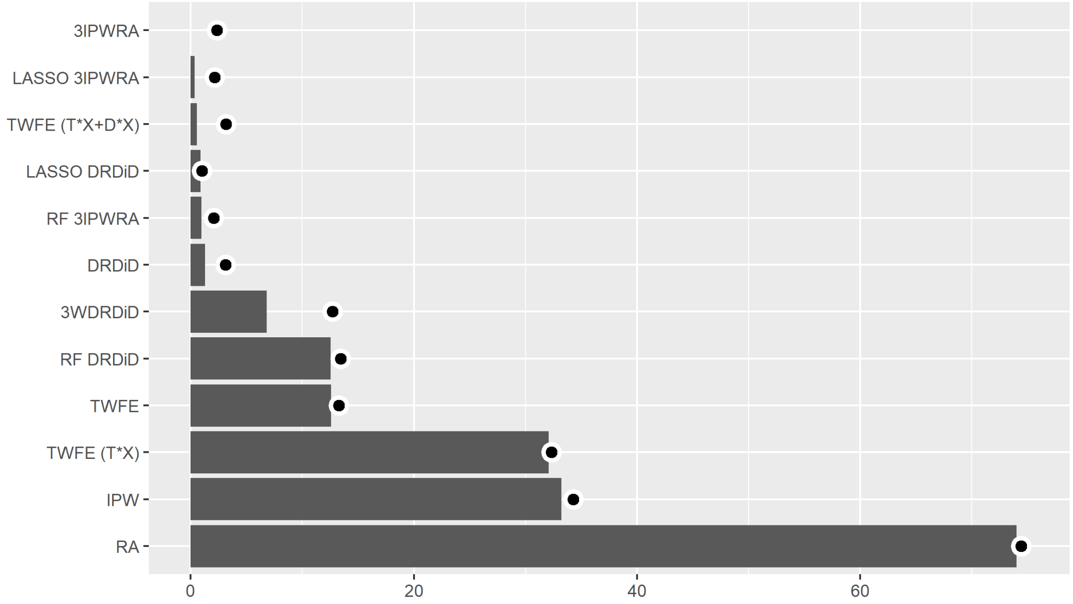


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 35: EXP 2A (Propensity score model correct, outcome regression model correct) with strong compositional changes, non-linear trend, and homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	12.581	13.295	18.452	0.001
Regression, eq. (11)	TWFE (T·X)	32.104	32.319	13.877	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	0.542	3.160	9.694	0.001
Abadie (2005)	IPW	33.233	34.297	71.812	0.011
Heckman et al. (1997)	RA	73.997	74.410	61.374	0.008
Sant'Anna and Zhao (2020)	DRDiD	1.310	3.107	7.937	0.019
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.885	1.010	0.237	1.231
Sant'Anna and Zhao (2020)*	RF DRDiD	12.538	13.421	22.929	3.601
Author's work, eq. (27)	3IPWRA	−0.012	2.356	5.548	0.024
Author's work, eq. (27)	LASSO 3IPWRA	−0.345	2.139	4.456	4.561
Author's work, eq. (27)	RF 3IPWRA	−0.979	2.084	3.383	1.583
Sant'Anna and Zhao (2020)*	3WDRDiD	−6.818	12.703	114.892	0.021

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

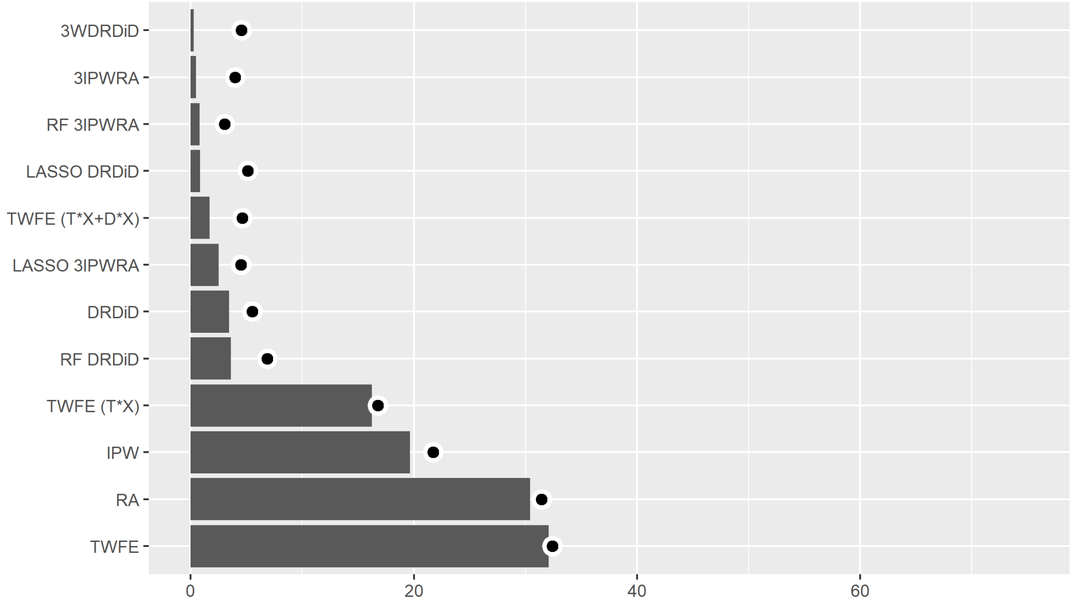


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 36: EXP 2D (Propensity score model incorrect, outcome regression model incorrect) with strong compositional changes, non-linear trend, and homogeneous effects

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	−32.096	32.406	19.991	0.001
Regression, eq. (11)	TWFE (T·X)	−16.237	16.781	17.972	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−1.684	4.631	18.609	0.002
Abadie (2005)	IPW	−19.643	21.724	86.110	0.011
Heckman et al. (1997)	RA	−30.406	31.440	63.930	0.019
Sant'Anna and Zhao (2020)	DRDiD	−3.431	5.523	18.740	0.019
Sant'Anna and Zhao (2020)*	LASSO DRDiD	0.852	5.112	25.403	1.455
Sant'Anna and Zhao (2020)*	RF DRDiD	−3.611	6.862	34.055	3.932
Author's work, eq. (27)	3IPWRA	−0.461	3.989	15.699	0.022
Author's work, eq. (27)	LASSO 3IPWRA	−2.529	4.486	13.728	2.689
Author's work, eq. (27)	RF 3IPWRA	0.788	3.052	8.695	1.735
Sant'Anna and Zhao (2020)*	3WDRDiD	0.258	4.543	20.573	0.022

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator

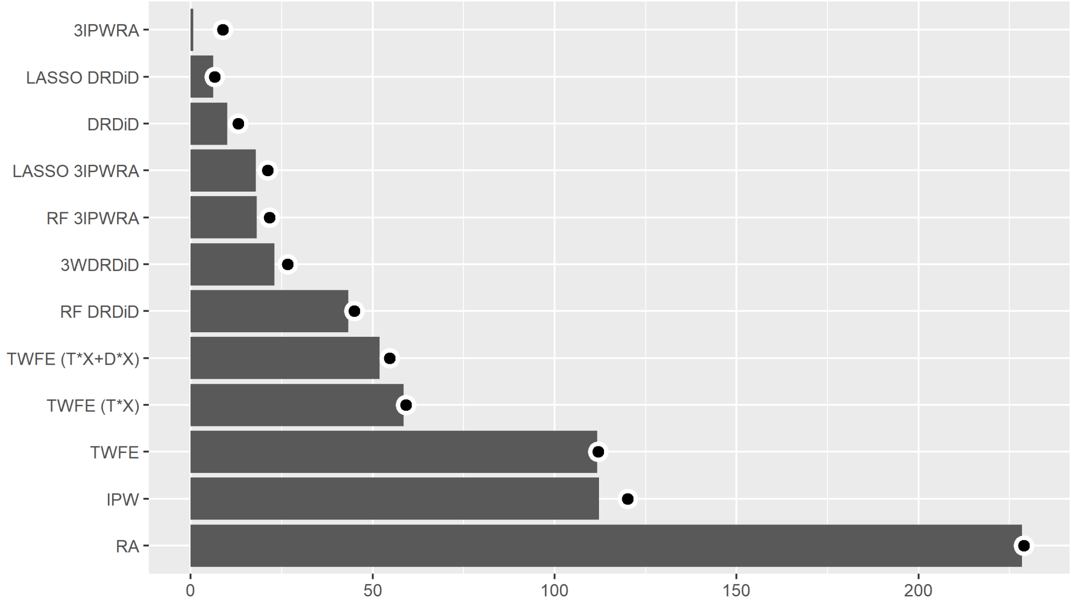


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.

Table 37: EXP 2A (Propensity score model correct, outcome regression model correct) with strong compositional changes, linear trend, homogeneous effects and different specification of the propensity score

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (8)	TWFE	111.716	111.945	51.235	0.001
Regression, eq. (11)	TWFE (T·X)	58.536	59.192	77.189	0.001
Regression, eq. (12)	TWFE (T·X+D·X)	−51.923	54.656	291.255	0.001
Abadie (2005)	IPW	112.165	120.128	1,849.706	0.011
Heckman et al. (1997)	RA	228.507	228.965	209.399	0.009
Sant'Anna and Zhao (2020)	DRDiD	−10.006	13.116	71.906	0.018
Sant'Anna and Zhao (2020)*	LASSO DRDiD	6.159	6.615	5.831	1.526
Sant'Anna and Zhao (2020)*	RF DRDiD	43.367	44.959	140.617	3.724
Author's work, eq. (27)	3IPWRA	−0.676	8.803	77.042	0.021
Author's work, eq. (27)	LASSO 3IPWRA	−17.891	21.156	127.497	9.644
Author's work, eq. (27)	RF 3IPWRA	−18.187	21.638	137.460	1.621
Sant'Anna and Zhao (2020)*	3WDRDiD	23.054	26.583	175.167	0.020

Visualization of the average absolute bias (bar) and RMSE (point) for each estimator



Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (eq. (8)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (11)). IPW is the inverse probability weighting (eq. (17)), RA is the regression adjustment approach (eq. (14)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (26)) and two versions that employ lasso and random forest for the propensity score and outcome regression respectively. 3IPWRA is the weighted regression with weights specified as in and the propensity score is estimated with logit, lasso and random forest (eq. (27)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (28)). Refer to the main text for further details.