UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

CORSO DI LAUREA MAGISTRALE / SPECIALISTICA IN ECONOMICS AND FINANCE

TESI DI LAUREA

"BEYOND REGRESSION: EVALUATING DIFFERENT SEMI-PARAMETRIC APPROACHES AND MACHINE LEARNING TOOLS IN THE DIFFERENCE-IN-DIFFERENCE DESIGN"

RELATORE:

CH.MO PROF. LUCA NUNZIATA

LAUREANDO/A: TOMMASO MANFÈ

MATRICOLA N. 1239202

ANNO ACCADEMICO 2021 – 2022

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Firma (signature) Tommaso Manfè

Contents

Abstract

The contribution of the thesis to the repeated cross-sections Difference-in -Difference (DiD) literature is threefold: first, it shows that the commonly-used DiD regression is severely biased under realistic scenarios and proposes alternative corrections; second, it presents a semi-parametric estimator robust to heterogeneity both in the treatment group and time dimensions; finally, it compares through Monte Carlo simulations the empirical performance of the proposed estimators with those suggested by the literature, in particular with 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 different semi-parametric estimators outperform regression, even if corrections provide substantial benefits. Following Sequeira (2016), the thesis estimates 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), the thesis provides substantial proof that the effect is close and even lower in magnitude than the one of the original paper. Still, the contribution reinforces the evidence that tariff reductions tend to weaken bribing behavior.

1 Introduction

Difference-in-Difference (DiD) is a widespread research design aimed at estimating the causal effects of a policy treatment that affects only a subgroup of the entire population, called the treated group, while leaving unaffected the other remaining part, referred to as the control group. Since observations are taken before and after the treatment, DiD compares four different groups of objects: the treated in the pre and post-treatment period, and the controls in the pre and post-period. The rationale of this empirical strategy is that if treated and control groups are subject to the same time trends, the control group can be used to estimate the counterfactual potential outcome of the absence of treatment for the group of people who have instead received the treatment. Indeed, DiD calculates the mean changes of the outcome variables for the non-treated over time and adds them to the mean level of the outcome variable for the treated before treatment to obtain the mean outcome the treated would have experienced if they had not been subjected to the treatment. By estimating this counterfactual outcome, the researcher is then able to retrieve the causal effect by taking the difference with the realized observed outcome in presence of the treatment.

The thesis 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 3 implements a Monte Carlo simulations under different scenarios to test the performance of the various estimators; Section 4 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 5 concludes with the most relevant findings.

Available coding material can be found at: https://github.com/tommaso-manfe.

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 left 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 time period before the treatment (pre-treatment period) and one in the period after the treatment took place (post-treatment period). Since the treatment is assumed to happen 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. Thus, it is useful to 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 is observable. The outcome that is realized is denoted by Y_t (not indexed by d). Finally, denote the observable covariates by X. Initially, we assume they do not vary over time but later on we are going to analyze the implications of relaxing such an assumption. The object we are interested

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

in estimating is the average effect on the treated (ATT), which is defined as follows:

$$ATT_t = E(Y_t^1 - Y_t^0)|D = 1) \tag{1}$$

$$= E[E(Y_t^1 - Y_t^0|X = x, D = 1)|D = 1]$$

$$= E_{X|D=1}\delta_t(x)$$

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. 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 and therefore the DiD setting usually focus on the estimation of the ATT.

2.2 Assumptions

Also in this case, the notation and examples follow closely Lechner et al. (2011). The examples refer to the setting where the researcher is evaluating the effect of participation in a training program for unemployment on earnings. For both treated and control groups, databases suitable for DiD contain information on the periods before and after training. The results of DiD design, with just a few differences, hold for both panel data and repeated cross-sections, even if the latter scenario will be the main focus of this paper.

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 assignment of treatment 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 time point 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. For example, when we consider the case of the effect of the training program on earnings, if the program is offered to a sizeable subpopulation, then equilibrium wages in the labor market may be altered. Since the training helps individuals in developing some specific skills, non-participants with comparable ability will be less likely to find a job after training occurs because the supply of labor in this skill group is now larger compared to the hypothetical scenario without the training program. Therefore, the outcome of the non-participants is not the same as the counterfactual world without the program. Consequently, SUTVA does not hold.

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. Otherwise, estimates are invalidated by the issue of reverse causality. Applying outcome notation to the explanatory variable X^d, it 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. For instance, if post-treatment job satisfaction is included in the job training model, the variable may be influenced by the treatment, causing endogeneity bias. However, measuring variables before the treatment does not automatically ensure exogeneity: individuals are forward-looking agents and may alter their behavior according to expectations about the future evolution of some variable. If such anticipatory behavior affects also the outcome variable, then the assumption of exogeneity may be violated as well. It is also worth noting that variables that are constant over time are exogenous by construction since treatment is a time-varying variable.

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 in Equation (3) also rules out the possibility of the anticipation effect of a future treatment on the pre-treatment outcome for the treated population. For instance, in the training example when the dependent variable is unemployment, NEPT would not hold if agents postpone their search for a job to the future because they may plausibly anticipate (in a way not captured by the covariates) that participation in an attractive training program is likely to increase both their probability of being hired and their wage received.

2.2.4 Common Trend

The assumption key for identification of causal effects in the DiD design requires that the differences over time in the expected potential outcomes of no treatment is independent to

belonging to the treated or control group:

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

$$= E(Y_1^0) - E(Y_0^0) \tag{5}$$

This is also known as the unconditional parallel trend (UCP) assumption. The hypothesis is essential because the trend of untreated potential outcomes for units belonging to the treated group is not known, but the path of untreated potential outcomes for controls is instead observable. However, the parallel trends assumption is considerably more plausible to hold only after conditioning on a set of observed covariates X:

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

$$= E(Y_1^0|X=x) - E(Y_0^0|X=x); \qquad \forall x \in \chi \tag{6}$$

The conditional parallel trend assumption (CPT) implies that if the treated group had not been subjected to the treatment, it would have evolved, conditional on X, with the same trend measured in the controls sub-population. Therefore, the inclusion of the covariates X should be driven to capture all variables that cause different time trends. For example, pretend that unemployed workers from sectors that are shrinking have more probability of being selected into

the training program. Consequently, the share of unemployed individuals who have previously worked in these declining sectors is over-represented in the training program group. Since such workers possess sector-specific experience and skills, their probability of reemployment is likely to lower while the reemployment chances of unemployed individuals in rising sectors is likely to increase. Since the respective presence of these groups of unemployed is unbalanced in the treated and control groups of the sample, the common trend assumption does not hold until we include the sector of the last employment as a control variable. In addition, the conditional parallel trend assumption is violated in the canonical example of Ashenfelter's dip. Ashenfelter (1978) introduced the idea that earnings often fall just before entering a training program due to negative idiosyncratic temporary shocks. If trainees experience a larger drop in earnings before the program with respect to non-trainees, then, because wages have a natural tendency to mean reversion, the DiD overestimates the causal effect of the participation in the training.

2.2.5 Common support

The conditional parallel trend assumption implies that it is necessary that observations with characteristics x exist in all four sub-samples determined by the treatment and time dummies.

This is guaranteed by the so-called common support (CS) assumption:

$$P[D|X] < 1 - \epsilon \quad and \quad P[D] > 0 \tag{7}$$

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 that 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. An example of a violation of common support assumption would be if the training program were mandatory for a specific age group, for instance those younger 25, and so there would not be any non-participants for this sub-population.

2.2.6 Proof of the Identification of the ATT

Recall that the conditional-on-X effect is defined as:

$$\delta_1(x) = E(Y_1^1 - Y_1^0)|X = x, D = 1)$$

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

Note that the first term $E(Y_1|X = x, D = 1)$ is identified because it coincides with the potential observed outcome of the treated sub-population at time $t = 1$, while the second term is the counterfactual outcome at $t = 1$ of no treatment for the treated sub-population, which instead is not observable. However, the common trend hypothesis enables us to estimate the counterfactual outcome:

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

$$= E(Y_1|X = x, D = 0) - E(Y_0|X = x, D = 0) + E(Y_0|X = x, D = 1) \quad (8)$$

which amounts to sum the common trend, estimated by taking the difference between the observed outcome at $t = 1$ and $t = 0$ for the control group, to the observed outcome at $t = 0$ for

the treated sub-population. Putting all pieces together:

$$\delta(x) = E(Y_1|X = x, D = 1) - E(Y_0|X = x, D = 1)$$

$$- (E(Y_1|X = x, D = 0) + E(Y_0|X = x, D = 0)) \tag{9}$$

where we simplified the notation writing $\delta_1(x) = \delta(x)$ since $\delta_0(x) = 0$ because of there is no causal effect before the treatment takes place. Since computed the missing counterfactual value $E(Y_1^0|X = x, D = 1)$, the causal effect can be therefore identified by taking the difference in outcomes from $t = 0$ to $t = 1$ in both the treated and control groups, and then taking a further difference between these two quantities, hence the name Difference-in-Difference.

2.3 DiD Regression: Two-Way-Fixed Effect

The key for using DiD regression, usually referred to as Two-Way-Fixed Effect (TWFE), is assuming an additive linear structure for potential outcomes. Using regression implicitly

assumes that the conditional expectation function (CEF) follows:

$$E(Y_0^0|D=0) = \alpha$$

$$E(Y_1^0|D=0) = \alpha + \gamma$$

$$E(Y_0^1|D=1) = \alpha + \beta$$

$$E(Y_1^1|D=1) = \alpha + \gamma + \beta + \delta$$

where α represents the expected value of 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 effect, namely differential in the potential outcome between the treated and control population in both 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 regression. Usually the Two-Way-Fixed-Effects (TWFE), in the case of absence of covariates, takes the following form:

$$Y_i = \alpha + \gamma T_i + \beta D_i + \delta(T_i \cdot D_i) + \epsilon_i \tag{10}$$

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 of the unit belongs to the treated individuals, 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. However, 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.

2.4 TWFE With 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:

$$E(Y_0^0|X, D = 0) = \alpha + X'\theta$$

$$E(Y_1^0|X, D = 0) = \alpha + \gamma + X'\theta$$

$$E(Y_0^1|X, D = 1) = \alpha + \beta + X'\theta$$

$$E(Y_1^1|X, D = 1) = \alpha + \gamma + \beta + \delta + X'\theta$$

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:

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

$$= \alpha + \gamma + \beta + X'\theta$$

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:

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

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.

2.4.1 Dealing with X-specific trends

Under many scenarios, the naive inclusion of covariates in the TWFE model is 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:

$$E(Y_0^0|X, D = 0) = \alpha_0 + \theta_0 X_0^0$$

$$E(Y_1^0|X, D = 0) = \alpha + \gamma + \theta_1 X_1^0$$

$$E(Y_0^0|X, D = 1) = \alpha + \beta + \theta_0 X_0^1$$

$$E(Y_1^0|X, D = 1) = \alpha + \gamma + \beta + \theta_1 X_1^1$$

Then, assuming that the conditional parallel trend assumption holds implies:

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

$$\alpha + \gamma + \beta + \theta_1 X_1^1 - (\alpha + \beta + \theta_0 X_0^1) = \alpha + \gamma + \theta_1 X_1^0 - (\alpha + \theta_0 X_0^0)$$

$$(\theta_1 X_1^1 - \theta_0 X_0^1) - (\theta_1 X_1^0 - \theta_0 X_0^0) = 0$$

$$\theta_1(X_1^1 - X_1^0) - \theta_0(X_0^1 - X_0^0) = 0 \tag{12}$$

where in the third passage the right-hand side is subtracted to the left one and the last line rearranges the terms. However, the computed quantity may be different from zero under many 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 (12) becomes:

$$\theta_1(X_1^1 - X_1^0) - \theta_0(X_0^1 - X_0^0) = (\theta_1 - \theta_0) \cdot (X^1 - X^0) \tag{13}$$

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:

$$ATT = (\alpha + (\gamma + \phi X^1) + \beta + \delta + \theta_0 X^1 - (\alpha + \beta + \theta_0 X^1)) - (\alpha + (\gamma + \phi X^0) + \theta_0 X_0 - (\alpha + \theta_0 X^0))$$

$$= \delta + \phi(X^1 - X^0)$$

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 experiments.

Instead, when we allow for time-varying covariates, by replacing Equation (12) and $\tau(X)$

in the ATT, the following result is obtained:

$$ATT = (\alpha + (\gamma + \phi X_1^1) + \beta + \delta + \theta_0 X_1^1 - (\alpha + \beta + \theta_0 X_0^1)) - (\alpha + (\gamma + \phi X_1^0) + \theta_0 X_1^0 - (\alpha + \theta_0 X_0^0))$$

$$= \delta + (\phi + \theta_0)(X_1^1 - X_1^0) - \theta_0(X_0^1 - X_0^0)$$

Consequently, when allowing for time-varying covariates, two conditions must be both satisfied to guarantee that $ATT = \delta$: the relationship of the covariates to the outcome is constant ($\phi = 0$) and the difference in the mean of the covariates among the two evolves equally between pre and post-treatment periods ($X_1^1 - X_1^0 = X_0^1 - X_0^0$) (Zeldow and Hatfield, 2019). As a consequence, a time-varying covariate is a confounder if its relationship to the outcome is time-varying or the covariate evolves differently in the treated and comparison 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 and the model can be written as:

$$Y_i = \alpha + \gamma T_i + \beta D_i + \delta(T_i \cdot D_i) + X_i'\theta + (T_i \cdot X_i')\omega + \epsilon_i \tag{14}$$

Zeldow and Hatfield (2019) outlines that this version eliminates bias in case 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 Section 3, the correction only partially works in the case of time-varying covariates. Therefore, the thesis analyzes another correction, which adds the interactions between the covariates and the treatment group dummy:

$$Y_i = \alpha + \gamma T_i + \beta D_i + \delta(T_i \cdot D_i) + X_i'\theta + (T_i \cdot X_i')\omega + (D_i \cdot X_i')\mu + \epsilon_i \tag{15}$$

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 Section 3.

2.4.2 Heterogeneous effects

In most realistic settings, the effect of the treatment likely varies 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, as in Cunningham (2021), let the treatment effect be heterogeneous in X, namely redefine the potential outcomes for the treated in the post period as $E(Y_1^1|X, D = 1) = \alpha + \gamma + \beta + (\delta + \omega X_1^1) + \theta X_1^1$. Then even assuming time-invariant covariates and $\theta_1 = \theta_0$ yields:

$$ATT = \delta + ((\theta + \omega)X^1 - \theta X^1) - \theta(X^0 - X^0)$$

$$= \delta + \omega X_1$$

Therefore, whenever $\omega \neq 0$ and thus the treatment is heterogeneous in X the estimate obtained by means of regression does not identify the true ATT, even when restricting covariates to be time-invariant.

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

Since in most of the settings it is not possible to use a fully saturated model in X for regression, TWFE assumes a CEF that is a linear function of X, then regression equation might differ from the true CEF. Indeed, including the control variables in a linear fashion implies the assumption of common trends 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

$$E(Y_t^d|X) = f(\alpha + \gamma T + \beta D + \delta TD + \theta X)$$

$$\neq \alpha + \gamma T + \beta D + \delta TD + \theta X$$

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

2.5 Semi-parametric DiD

To overcome the main limitations of TWFE, various semi-parametric estimators have been proposed in the literature. In the following section, we start from the related estimators of Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020). All three of the settings work properly under the assumption of time-invariant covariates. More precisely, these models assume that the pooled repeated cross-section data $\{Y_i, D_i, X_i, T_i\}_i^n$, where i refers to individual i and n is the number of observations, consist of independent and identically distributed (IID)

draws from the mixture distribution:

$$P(Y \leq y, D = d, X \leq x, T = t) = t \cdot \lambda \cdot P(Y_1 \leq y, D = d, X \leq x | T = 1) \tag{16}$$

$$+ (1 - t) \cdot d \cdot (1 - \lambda) \cdot P(Y_0 \leq y, D = d, X \leq x | T = 0)$$

where $(y, d, x, t) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}^k \times \{0, 1\}$, with the joint distribution of (D, X) being invariant to T and λ representing the proportion of individuals at $t = 1$. Such an hypothesis, together with those described in Section 2.2, allows the three estimators to identify the ATT in the case of repeated cross-sections. When panel data are available data, the authors instead assume that data $Y_{i0}, Y_{i1}, D_i, X_i{}_{i=1}^{n}$ are independent and identically distributed (IID).

In all three estimators, covariates are thus not allowed to change over time, ruling out the possibility of compositional changes. However, a few papers have tried to study the implications of allowing compositional changes in the covariates between the pre and post-treatment periods. For instance, Hong (2013) showed that, when the distribution of X varies over time, the traditional propensity score does not properly balance the covariates between the treated and untreated groups. To relax such an assumption, the idea that the author proposes is to define a multivariate propensity score that allows for both selection in treatment and time.

Alternatively,Blundell et al. (2004) and Blundell and Dias (2009) define the propensity score as the probability of belonging to the treated group at the post-treatment period, $P(T \cdot D|X) = 1$, instead of the probability of just belonging to the treatment group, $P(D = 1|X)$. Their idea is to therefore match the four different groups created by the time and treatment dimensions. All these ideas are developed further in this section and the performance of the estimators will be tested empirically through Monte Carlo simulations, which will be the main content of Section 3.

2.5.1 Outcome Regression

The outcome regression (OR) approach, mainly employed in the form of regression adjustment (RA), relies on researchers' ability to specify the model for the outcome evolution. Indeed, Heckman et al. (1997) starting from the definition of the ATT under conditional parallel trends and using the law of iterated expectations yields:

$$\begin{aligned} ATT &= E\left[E(Y_1 - Y_0|X, D = 1) - E(Y_1 - Y_0|X, D = 0)|D = 1\right] \\ &= E(Y_1 - Y_0|D = 1) - E\left[E(Y_1 - Y_0|X, D = 0)|D = 1\right] \end{aligned} \tag{17}$$

where the first term in Equation (17) can be computed by taking sample averages, while the second expected value must be estimated. To retrieve the missing term, the expected value can

be estimated by fitting a regression in the controls group and taking predictions based on the empirical distribution of X_i among treated units. More formally:

$$\delta^{OR} = \bar{Y}_{1,1} - \bar{Y}_{1,0} - \left[\frac{1}{n_{treat}} \sum_{i|D_i=1} (\hat{\mu_{0,1}}(X_i) - \hat{\mu_{0,0}}(X_i)) \right] \tag{18}$$

where $\bar{Y}_{d,t} = \sum_{i|D_i=1} Y_{it}/n_{d,t}$ is the sample average outcome among units in treatment group d and 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 (15) that includes also interactions between X_i with both treatment group and time dummies, even if they would differ due to the fact that the outcome regression approach re-weights based on the distribution of X_i among units with $D_i = 1$ (Roth et al., 2022). In addition, the available methods in the outcome regression approach for the estimation of $\hat{\mu_{d,t}}(X)$ are not limited to linear regression and include more flexible semi-/non-parametric methods. For example, many papers employ nearest neighbor matching to associate treated with untreated units with approximately identical covariate values. The estimation, in this case, consists of a simple DiD estimator between treated

units and the matched comparison group. However, the condition for the consistency of the ATT of the outcome regression is the correct specification of $\hat{\mu_{d,t}}(X)$.

2.5.2 Inverse Probability Weighting

The Inverse Probability Weighting (IPW) approach proposed by Abadie (2005) avoids directly modeling the outcome evolution while focusing on the treatment model, namely the conditional probability of being in the treatment group given covariates, $p(X) \equiv P(D = 1|X)$. In the case of panel data, under the standard assumptions expressed in section 2.2, the ATT can be expressed as

$$\delta^{IPW} = \frac{1}{E(D)} \cdot E\left[\frac{D - p(X)}{1 - p(X)} \cdot (Y_1 - Y_0)\right] \tag{19}$$

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

treated and untreated groups. The IPW is then estimated by using the sample analog

$$\delta^{IPW} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n}(D_j)} \cdot \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \cdot (Y_{i1} - Y_{i0})\right] \tag{20}$$

where $\hat{\pi}(X)$ is an estimator of the true, unknown propensity score $p(x) = P(D = 1|X)$. While in the case of repeated cross-sections the ATT is estimated by :

$$\delta^{IPW} = \frac{1}{E(D) \cdot \lambda} \cdot E\left[\frac{D - p(X)}{1 - p(X)} \cdot \frac{T - \lambda}{1 - \lambda} \cdot Y\right] \tag{21}$$

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

$$\delta^{IPW} = \frac{1}{\lambda \cdot \frac{1}{n}\sum_{j=1}^{n}(D_j)} \cdot \sum_{i=1}^{n}\left[\frac{D_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \cdot \frac{T_i - \lambda}{1 - \lambda} \cdot Y_i\right] \tag{22}$$

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 in the case the propensity score model is correctly specified.

2.5.3 Doubly Robust Methods

Sant'Anna and Zhao (2020) combine the OR and the IPW approaches into a doubly robust estimand for the ATT. Double robustness is the property that if either (but not both) the propensity

score model or the outcome regression models are misspecified, the resulting estimand still identifies the ATT. Intuitively, the doubly robust Difference-in-Difference (DR DiD) 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 from the case of panel data since it gives an easier intuition. Denote $\Delta Y = Y_1 - Y_0$ and $\mu^p_{d,\Delta}(X) = \mu^p_{d,1}(X) - \mu^p_{d,0}(X)$ where $\mu^p_{d,t}(X)$ being a model for the true, unknown outcome regression $E[Y_t|D = d, X = x]$ with $d, t \in \{0, 1\}$. Then the DRDiD estimand is:

$$\delta^{dr,p} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{(1-D)p(X)}{1-p(X)}}{E\left[\frac{(1-D)p(X)}{1-p(X)}\right]}\right)\left(\Delta Y - E[Y_1 - Y_0|D = 0, X = x]\right)\right] \tag{23}$$

The estimand can be decomposed in two parts. The first parenthesis is the IPW part of the estimator, namely the weighting scheme. For the treated group, $D = 1$ and $1 - D$ reduces to zero. The weight is therefore $\frac{1}{E(D)}$, where the denominator is there just to guarantee that 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 $\frac{p(X)}{1-p(X)}$. Likewise, the denominator has the function to let the weights to sum up to one. On the other hand, the right parenthesis shows 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 in the control group and fitting $Y_1 - Y_0$ based on the empirical distribution of X_i among the treated individuals. Similarly, the sample analog can be written as:

$$\delta^{dr,p} = \sum_{i=1}^{n} \left[\left(\frac{D_i}{\sum_{j=1}^{n} D_j} - \frac{\frac{(1-D_i)\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}}{\sum_{j=1}^{n} \left[\frac{(1-D_j)\hat{\pi}(X_j)}{1-\hat{\pi}(X_j)} \right]} \right) \left(\Delta Y - \hat{\mu}^{p}_{0,\Delta}(X) \right) \right] \tag{24}$$

where $\hat{\pi(X)}$ be an arbitrary model for the true, unknown propensity score $p(X)$.

Instead, when dealing with repeated cross-section data, define $m^{rc}_{d,t}(x) \equiv E[Y|D = d, T = t, X = x]$, $d, t \in \{0, 1\}$ 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 $\mu^{rc}_{d,\Delta}(X) \equiv \mu^{rc}_{d,1}(X) - \mu^{rc}_{d,0}(X)$. Then the DR DiD estimator is defined as:

$$\delta^{dr,rc}_{1} = E[(\omega^{rc}_{1}(D) - \omega^{rc}_{0}(D, T, X; p))(Y - \mu^{rc}_{0,Y}(T, X))] \tag{25}$$

where:

$$E[(\omega_1^{rc}(D)] = E[(\omega_{1,1}^{rc}(D)] - E[(\omega_{1,0}^{rc}(D)] \tag{26}$$

$$\omega_0^{rc}(D,T,X;p) = \omega_{0,1}^{rc}(D,T,X;p) - \omega_{0,0}^{rc}(D,T,X;p) \tag{27}$$

and for $t \in 0, 1$:

$$E[(\omega_{1,t}^{rc}(D,T)] = \frac{D \cdot 1\{T=t\}}{E[D \cdot 1\{T=t\}]} \tag{28}$$

$$E[(\omega_{0,t}^{rc}(D,T,X;p)] = \frac{(1-D)p(X) \cdot 1\{T=t\}}{1-p(X)} \Bigg/ E\left[\frac{(1-D)p(X) \cdot 1\{T=t\}}{1-p(X)}\right] \tag{29}$$

In addition, Sant'Anna and Zhao (2020) present also a locally semi-parametrically efficient version of the above estimator, which means that asymptotic variance achieves the semi-parametric efficiency bound when the working models for the nuisance functions are correctly specified.

$$\begin{aligned}\delta_2^{dr,rc} = \delta_1^{dr,rc} &+ (E[\mu_{1,1}^{rc}(X) - \mu_{0,1}^{rc}(X)|D=1] - E[\mu_{1,1}^{rc}(X) - \mu_{0,1}^{rc}(X)|D=1,T=1]) \\ &- (E[\mu_{1,0}^{rc}(X) - \mu_{0,0}^{rc}(X)|D=1] - E[\mu_{1,0}^{rc}(X) - \mu_{0,0}^{rc}(X)|D=1,T=0])\end{aligned} \tag{30}$$

By replacing $p(x)$ with $\hat{\pi}$ and the expectation with sample means, the sample analog is obtained. The outcome equation and the propensity score can be modeled either parametrically, for instance with a linear and logistic regression respectively, or non-parametrically, including machine learning methods. Indeed, the score function of the DRDiD satisfies the Neyman

orthogonality condition that will be defined in Equation (42) and that is key for the debiased machine learning literature (Chernozhukov et al., 2018). The authors instead opt to 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 that are present in their paper, this latter version outperforms others using the traditional parametric models. This last locally-efficient version of the DRDiD, because of its advantages, is the one used in section 3 for the Monte Carlo simulations. In addition, also modified versions using lasso and random forest will be tested as well. DRDiD will generally be consistent if either one of the propensity score and outcome models is correctly specified.

2.5.4 Triple Inverse Probability Weighting Regression Adjusted Estimator

The triple inverse probability weighting regression adjusted (3IPWRA) is an estimator for repeated cross-sections that builds on the idea of Blundell et al. (2004) and Blundell and Dias (2009). 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 observations do not follow 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 match 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.

Likewise, 3IPWRA computes the propensity score 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. However, the propensity score is not used for matching but in the context of inverse probability weighting. In fact, 3IPWRA uses the estimated propensity score to calculate the Horvitz-Thompson inverse prob-

ability 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 (15). 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 the following moment condition and sample analog:

$$E\Bigg[\underbrace{\left(DT - \frac{(1-DT)p(X)}{1-p(X)}\right)}_{=w_i}(Y_i - \bar{X}_i'\omega)\bar{X}_i\Bigg] = 0 \tag{31}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left(D_iT_i - \frac{(1-D_iT_i)\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\right)(Y_i - \bar{X}_i'\omega)\bar{X}_i\right] = 0 \tag{32}$$

where $\hat{\pi}(X_i)$ 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 (d,t) where i belongs and then computing propensity scores. For the group such that $(t,d) = (1,1)$ 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 full vector of controls such as in Equation (15), namely T, D, TD, X', TX', DX'. Such weighted regression, according to Imbens (2004), approximates the more general semi-parametric version of the IPWRA estimator. Both logistic and machine learning estimators for the propensity score will be analyzed in the simulations present in Section 3.

2.5.5 Triple Weighting Doubly Robust Difference-in-Difference

The rationale of the triple weighting doubly robust Difference-in-Difference (3WDRDiD) estimator is to adapt the weighting scheme in the DRDiD of Sant'Anna and Zhao (2020) to handle the weights utilized in the 3IPWRA estimator. Since it refers to 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 $p(X) = P(DT = 1|X)$, then the DR DiD estimator is defined as:

$$\delta_1^{dr,rc} = E[(\omega_1^{rc}(D) - \omega_0^{rc}(D, T, X; p))(Y - \mu^{rc}_{0,Y}(T, X))] \tag{33}$$

where:

$$E[(\omega_1^{rc}(D)] = E[(\omega^{rc}_{1,1}(D)] - E[(\omega^{rc}_{1,0}(D)] \tag{34}$$

$$\omega_0^{rc}(D, T, X; p) = \omega^{rc}_{0,1}(D, T, X; p) - \omega^{rc}_{0,0}(D, T, X; p) \tag{35}$$

and for $t \in 0, 1$:

$$E[(\omega^{rc}_{1,1}(D, T)] = \frac{D \cdot 1\{T = t\}}{E[D \cdot 1\{T = t\}]} \tag{36}$$

$$E[(\omega^{rc}_{1,0}(D, T)] = \frac{(1 - D)p(X) \cdot 1\{T = t\}}{1 - p(X)} \Bigg/ E\left[\frac{(1 - D)p(X) \cdot 1\{T = t\}}{1 - p(X)}\right] \tag{37}$$

$$E[(\omega^{rc}_{0,t}(D, T, X; p)] = \frac{(1 - D)p(X) \cdot 1\{T = t\}}{1 - p(X)} \Bigg/ E\left[\frac{(1 - D)p(X) \cdot 1\{T = t\}}{1 - p(X)}\right] \tag{38}$$

On an intuitive level, with respect to the original Sant'Anna and Zhao (2020), the main difference is the use of the "triple-matching weights", which are the propensity scores computed by separately "matching" the treated and 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 (28), but follow the adjustment employed for two untreated sub-populations (see Equation (37)), since the idea is to adjust these three groups for both the heterogeneity in time and treatment group dimensions.

2.6 Machine learning first-stage estimates

2.6.1 Debiased machine learning

There is growing literature on the use of machine learning for causal inference. 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 to minimize the mean square error (MSE) of the observations out of the sample. 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:

$$Y = g_0(D, X) + \zeta, \quad E(\zeta|D, X) = 0 \tag{39}$$

$$D = m_0(X) + V, \quad E(V|X) = 0 \tag{40}$$

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 (39) 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 (40)

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:

$$E[\psi(W; \theta_0; \eta)] = 0 \tag{41}$$

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 (1) satisfies Equation (41) yielding θ_0 as a unique

solution, and (2) that satisfies the Neyman orthogonality condition defined as:

$$\partial_\eta E[\psi(W;\theta_0;\eta)]|_{\eta=\eta_0} = 0 \tag{42}$$

The Neyman orthogonality expressed in Equation (42) guarantees that the moment condition defined in Equation (41) 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

$$\psi(W;\theta_0;\eta) \equiv \frac{D(Y-g(0,X)}{p} - \frac{m(X)(1-D)(Y-g(0,X)}{p(1-m(X))} - \frac{D}{p}\theta \tag{43}$$

$$\eta = (g, m, p), \quad \eta_0 = (g_0, m_0, p_0), \quad p_0 = P(D=1)$$

satisfies the Neyman orthogonality condition in Equation (42). By substituting Y in Equation (43) with the variation ΔY between pre and post-treatment period and accordingly adjust

the outcome regression estimates $g(0, X)$, it can be shown that the Equation (43) corresponds to the DRDiD estimator for panel data proposed by Sant'Anna and Zhao (2020). 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:

$$||\hat{m_0} - m_0|| + ||\hat{g_0} - g_0|| \leq o(N^{-1/4}) \tag{44}$$

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 is 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.

2.6.2 Lasso

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

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_p X_p + \epsilon \tag{45}$$

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 related linearly 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_λ^L minimize the formulation of the least-squares with a penalty term governed by the parameter λ:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{46}$$

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:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{47}$$

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.

2.6.3 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:

$$\sum_{j=1}^{J} \sum_{i \in \mathbb{R}_j} \left(y_i - \hat{y}_{\mathbb{R}j} \right)^2 \tag{48}$$

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 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 greater than s for $R_2(j, s)$. Then recursive binary splitting looks for the value of j and s that minimize the following equation:

$$\sum_{i:x_i \in \mathbb{R}_1(j,s)} \left(y_i - \hat{y}_{\mathbb{R}_1} \right)^2 + \sum_{i:x_i \in \mathbb{R}_2(j,s)} \left(y_i - \hat{y}_{\mathbb{R}_2} \right)^2 \tag{49}$$

where $\hat{y}_{\mathbb{R}_1}$ represents the average response for the training observations in the region $R_1(j, s)$, and $\hat{y}_{\mathbb{R}_1}$ is represents the average response for the training observations in the region $R_2(j, s)$. Then 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:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in \mathbb{R}_m} \left(y_i - \hat{y}_{\mathbb{R}_m} \right)^2 + \alpha |T| \tag{50}$$

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. Also in this case, 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

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{bag}^{b}(x) \tag{51}$$

where $\hat{f}_{bag}^{b}(x)$ 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. For example, m is usually chosen in the order of $\sqrt{p}$. Intuitively, random forest gains prediction efficiency by decorrelating the decision trees and thereby causing the average to have lower variance.

3 Monte Carlo Simulations

This section conducts a series of Monte Carlo experiments in order to study the finite sample properties of the proposed estimators in the case of repeated cross-sections. The different methodologies are tested in three different experiments. In each design, one cross-section is observed at $T = 0$ and the other at $T = 1$, constituting a total sample size of $n = 1000$. The Monte Carlo simulation consists of 500 random generations of the dataset and for each repetition the estimation results are stored. Each of the three experiments considers a trend that depends on the covariates under different scenarios. Indeed, in addition to X-specific trends, Experiment 0 assumes randomized selection into treatment, time-invariant covariates,

and homogeneous treatment effects in X. Following Section 2.4, in this setting we expect that all estimation methods perform relatively well, including the TWFE, since its all assumptions are satisfied. Conversely, Experiment 1 allows for non-random selection into treatment, testing the robustness of the traditional TWFE under new circumstances. Finally, Experiment 2 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. Thus, Experiment 2 reproduces the more realistic setting since different dimensions of hetereogeneity are allowed.

Following the notation as in Sant'Anna and Zhao (2020) and Kang and Schafer (2007), each experiments considers four different data generating processes (DPGs) which are aimed to model whether or not the researcher can correctly specify the propensity score and the outcome models. First of all, for a generic variable $W = (W_1, W_2, W_3, W_4)'$, define the underlying true outcome and propensity score model:

$$f_{reg}(W) = 210 + 25.4 \cdot W_1 + 13.7 \cdot (W_2 + W_3 + W_4) \tag{52}$$

$$f_{ps}(W) = 0.75 \cdot (-W_1 + 0.5 \cdot W_2 - 0.25 \cdot -0.1 \cdot W_4) \tag{53}$$

The function $f_{ps}(W)$, which determines selection into treament, is modeled through the inverse of the logit function $expit(f_{ps}(W)) = \frac{exp(f_{ps}(W))}{1+exp(f_{ps}(W))}$ and is studied to produce an average response rate of 0.5. As consequence, estimators parametrically assuming a logit distribution will conform to the true DPG. The baseline function for the outcome $f_{reg}(W)$, that will be adapted to the context of each of the three experiments, 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 DPG is not severe because the difference between the average of the treated units and the average 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 estimates.

The generic W represents two possible variables, X and Z. 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. 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 $\tilde{Z}_4 = (20 + X_2 + X_4)^2$. The vector $Z = (Z_1, Z_2, Z_3, Z_4)'$ is the set of variables that are observable by the researcher.

The unique generic DPG (expressed in terms of W) for each experiment leads to four cases depending on whether W is replaced by the observed variable Z or by the unobservable variable X. When both modeling functions are $f_{ps}(Z)$ and $f_{reg}(Z)$, since the researcher observes Z, this will lead to correctly specified models for the estimation of both the propensity score and the outcome regression (DPG A). However, when data are generated from $f_{ps}(X)$ and $f_{reg}(X)$, the researcher, who has only access to Z, will misspecify both models (DPG D). Since Z is a highly non-linear transformation of X and its interactions, the misspecification is likely to cause a sizeable bias in the estimation. However, such a scenario is the most realistic since researchers do not know *a priori* the form of phenomenon they are analyzing. Finally, also the two cases in which just one of the models has the correct specification is analyzed (DPG B and DPG C).

Table 1 summarizes the different methods that are tested in each experiment, which will be 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 consider a logistic propensity score working model and a linear regression working model for the out-

come evolution. Therefore, the first is estimated using maximum likelihood and the second by ordinary least squares.

For the DRDiD and 3IPWRA we also allow for the possibility of first-stage machine learning estimates. Such non-parametric methods should better capture the non-linearities when the working models are misspecified. When lasso is utilized, both 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 such that error is within 1 standard error of the minimum the average cross-validation error. Since lasso implicitly performs variable selection and can therefore handle a multitude of covariates, in the simulation this advantage is reflected by enabling lasso to employ an expanded set of covariates that include all second order terms and interactions. Despite being technically possible in this synthetic dataset to include all interactions terms also for the traditional estimators, the choice is driven to emulate a more realistic scenario where it is not completely feasible for the researcher to methodically include an expanded set of covariates under traditional estimation methods.

Otherwise, there is the risk that the number of predictors n will be close or even higher than the number of observations p, invalidating the estimation.

When random forest is utilized, the estimation is implemented referring to 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 utilizing machine learning tools, I do not perform sample splitting, as suggested in Chernozhukov et al. (2018) and Bach et al. (2021), because of the limited dimension of the dataset. Indeed, many of the analyzed methodologies have intermediate stages that rely on estimations in each of the four sub-groups of the population defined by the time and treatment group dimensions, and threfore sample-splitting would further reduce the number of available observations. Preliminary results with 2-fold random partition of the observations did not seem to improve results but, as indicated by Bach et al. (2021), future works should analyze, when feasible, 4-fold and 5-fold random partitions which demonstrated to work better in a

variety of empirical simulations. Despite of this, the final results on techniques that rely on machine learning first-stages are encouraging and the issue overfitting seems limited in our context.

We performed preliminary simulations of other estimators that utilized machine learning estimates in the DiD, such as the one proposed by Chang (2020) and Nie et al. (2021), but the results showed significantly higher variance that the ones studied in the simulation. Further analysis, encompassing also Zimmert (2018), may be an interesting exercise for future contributions.

Table 1: Summary table of the estimator utilized in the Monte Carlo simulation

Estimator	Description
TWFE	two-way-fixed-effects regression with covariates as ineq. (11)
TWFE (T·X)	two-way-fixed-effects regression with covariates and their interaction with the time dummy, as in eq. (14)
TWFE (T·X+D·X)	two-way-fixed-effects regression with covariates and their interaction with the time and treatment group dummy, as in eq. (15)
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
LASSO 3IPWRA	Triple propensity score inverse probability weighting regression adjusted estimator with lasso
RF 3IPWRA	Triple propensity score inverse probability weighting regression adjusted estimator with random forest
3WDRDiD	Doubly robust DiD Sant'Anna and Zhao (2020) adjusted with triple propensity score inverse probability weighting

3.1 Experiment 0: X-Specific Trends and Randomized Selection

Experiment 0 is randomized experiment with X-specific trends, time-invariant covariates, and homogeneous treatment effects in X. Since in Experiment 0 the selection of treated individuals is assumed to be random, DGP.A coincides with DPG.B and DPG.C is equivalent to DPG.D because the true propensity score is a constant and there is no need to specify a correct propensity score model. Because of that, Experiment 0 uses the notation DPG.AB and DPG.CD to signal the aforementioned equality but to maintain the consistency in notation used in further experiments. The DPGs are therefore specified as:

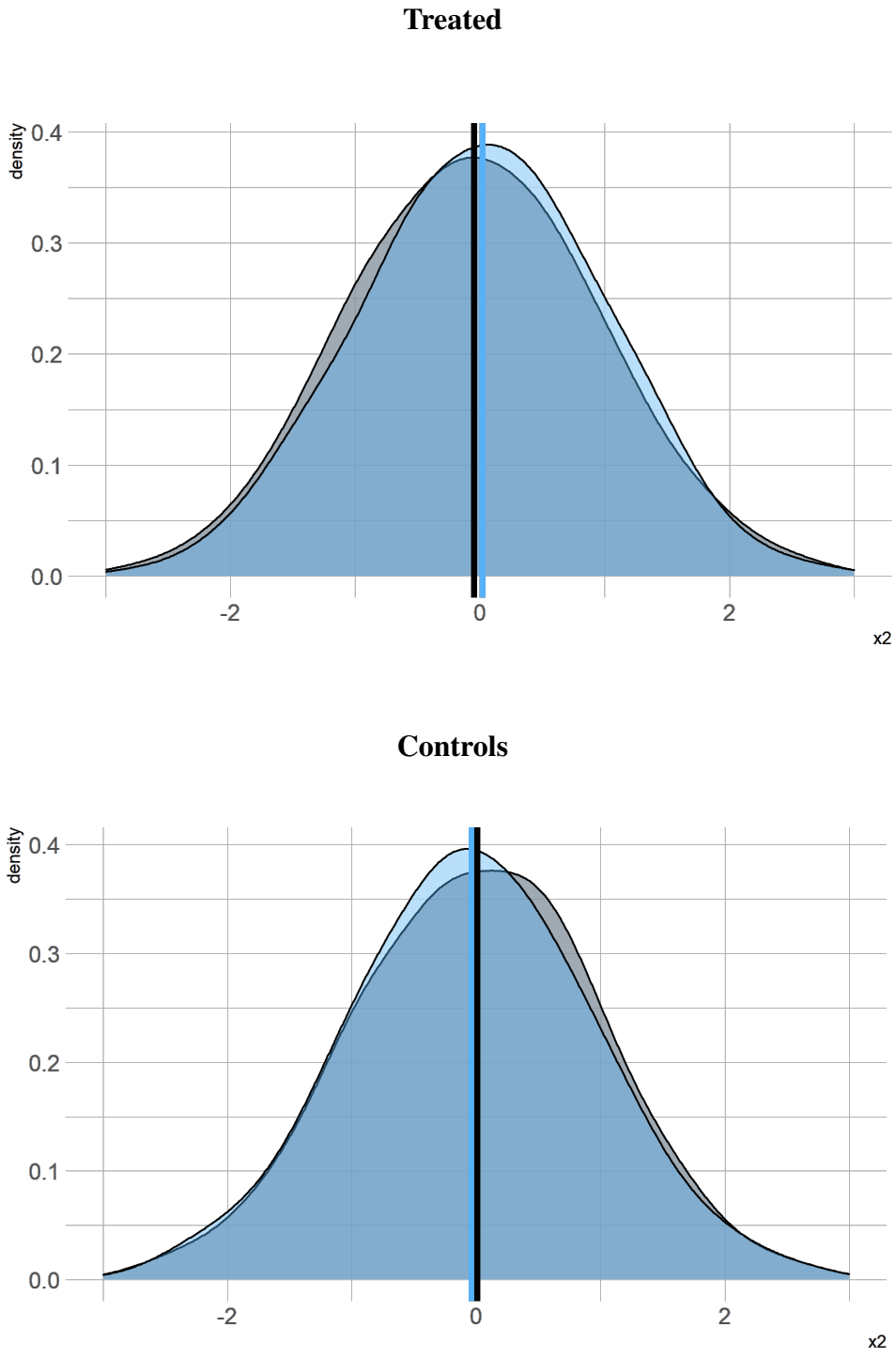
DPG.AB (PS and OR models correct)	**DPG.CD (PS and OR models incorrect)**
$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$	$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$
$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(d)$	$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \epsilon_1(d)$
$p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

where ϵ_0, $\epsilon_1(d)$, $d = 0, 1$ are independent standard normal random variables and represent the stochastic error term of the potential outcomes, p is the constant probability of being treated, λ is the proportion of the sample when $T = 1$ and U_d and U_t are independent standard uniform stochastic variables used to randomly select individuals into treatment and into 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 variance one and represents the time-invariant group heterogeneity between treated and untreated populations. The trend is for simplicity 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 $2 \cdot f_{reg}(W)$in 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 bias. The results of the simulations are displayed in Table 2 and Table 3 .

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

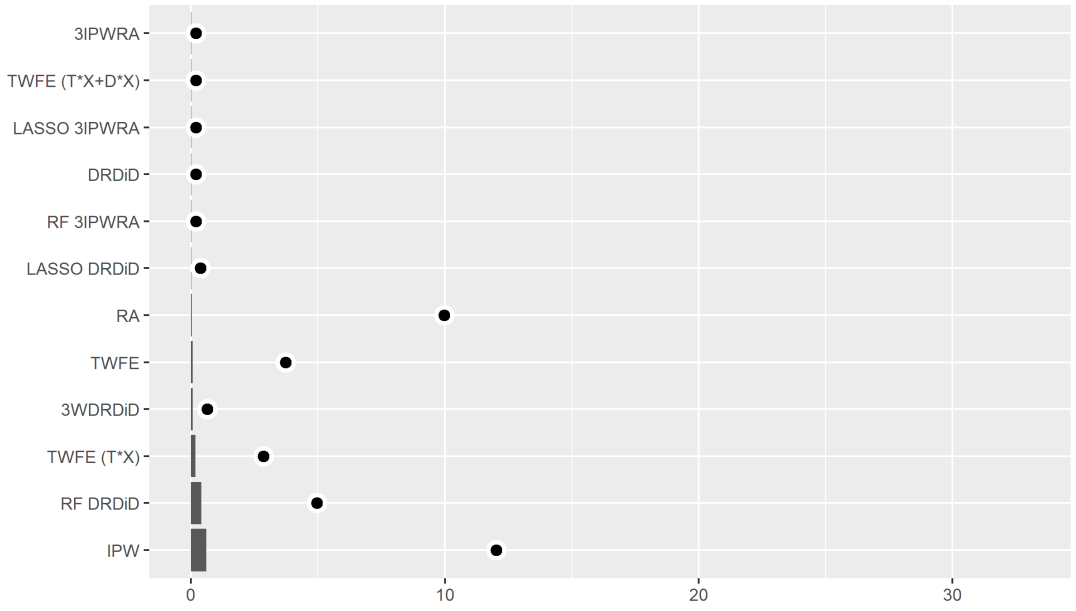


Notes: The graph considers a representative random sample from Exp.0 with DPG CD. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the 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.

Table 2: Exp.0AB Outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-0.060	3.733	13.929	0.001
Regression, eq. (14)	TWFE (T·X)	-0.178	2.852	8.105	0.001
Regression, eq. (15)	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
author's work, eq. (32)	3IPWRA	0.008	0.185	0.034	0.018
author's work, eq. (32)	LASSO 3IPWRA	0.009	0.184	0.034	0.458
author's work, eq. (32)	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

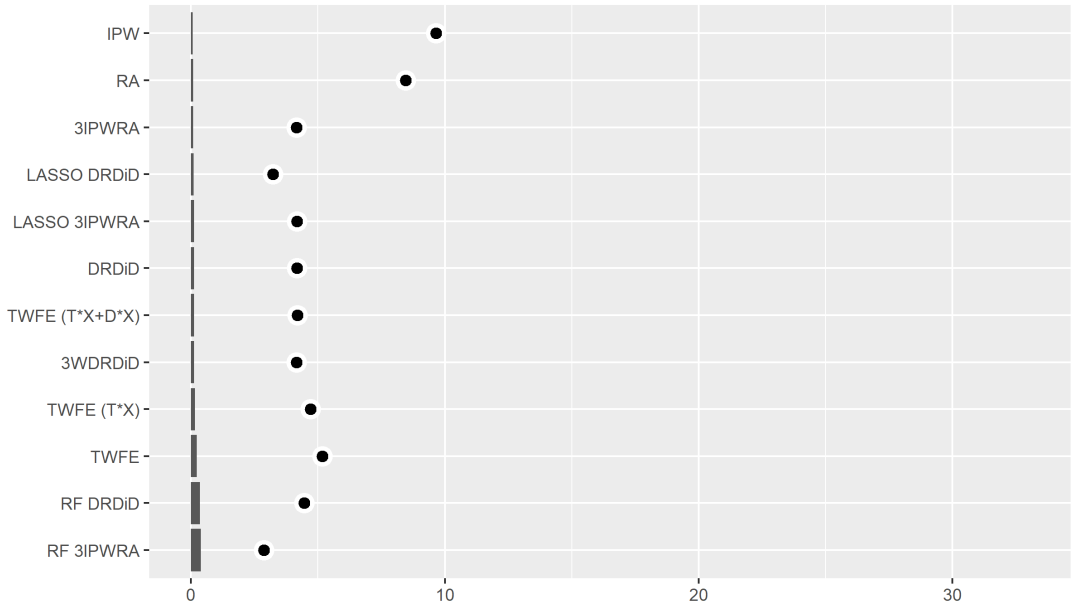


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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 3: Exp.0CD Outcome regression model

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	0.228	5.167	26.647	0.001
Regression, eq. (14)	TWFE (T·X)	0.157	4.702	22.083	0.001
Regression, eq. (15)	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. (32)	3IPWRA	0.089	4.159	17.292	0.053
author's work, eq. (32)	LASSO 3IPWRA	0.112	4.164	17.327	0.445
author's work, eq. (32)	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'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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.

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 doubly robust DRDiD estimator of Sant'Anna and Zhao (2020) ($RMSE = 0.184$), the lasso and standard version of 3IPWRA ($RMSE = 0.184$ and $RMSE = 0.185$ 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 among all estimates. However, overall all estimators perform relatively well, and methods that employs machine-learning first stage estimates display a better efficiency in terms of variance since they better capture the non-linearities needed to reproduce the true DPG. The lowest bias (-0.074) is displayed by the IPW estimator of Abadie (2005) , but probably the DRDiD and 3IPWRA have an overall better performance. The lowest RMSE is the one of the random forest specification of the 3IPWRA ($RMSE = 2.875$).

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

Experiment 1 closely replicates the simulation present in Sant'Anna and Zhao (2020). Besides X-specific trends, here the selection to the treatment is not randomized, causing additional obstacles to the identification of the causal parameter. The covariates are assumed to be time-invariant, therefore disallowing the possibility of compositional changes in the independent variables, 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 DPGs are specified as follows:

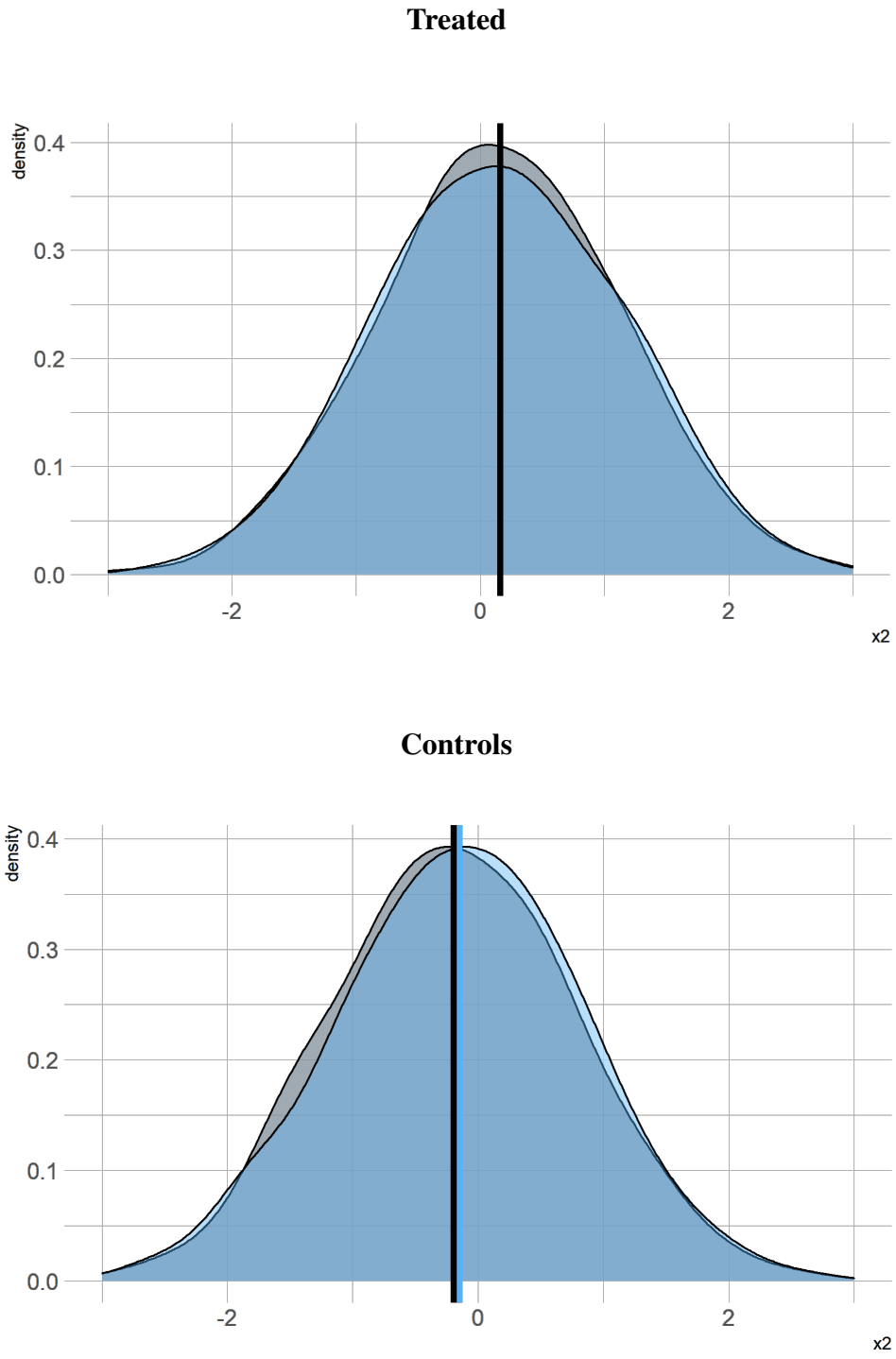
DPG.A (PS and OR models correct)	**DPG.B (PS model incorrect, OR correct)**
$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$	$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0(d)$
$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(d)$	$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \epsilon_1(d)$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

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

where the notation closely follows the one in Experiment 0 (refer to Section 3.1). 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 we can retrieve a correctly specified model or not. When the treatment selection is driven by the propensity score, treatment and control groups are generally heterogeneous in the terms of covariate characteristics. Figure 2 demonstrates that distribution of X_2 is indeed significantly different between the two groups. This can be shown for the other three covariates as well. As explained by Kang and Schafer

(2007), the DPG is studied to cause a difference in means of one-quarter of the population standard deviation, creating therefore a realistic but meaningful selection bias that is enough to cause some estimators to fail. The results of the simulation can be found in Table 4, Table 5, Table 6 and Table 7.

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

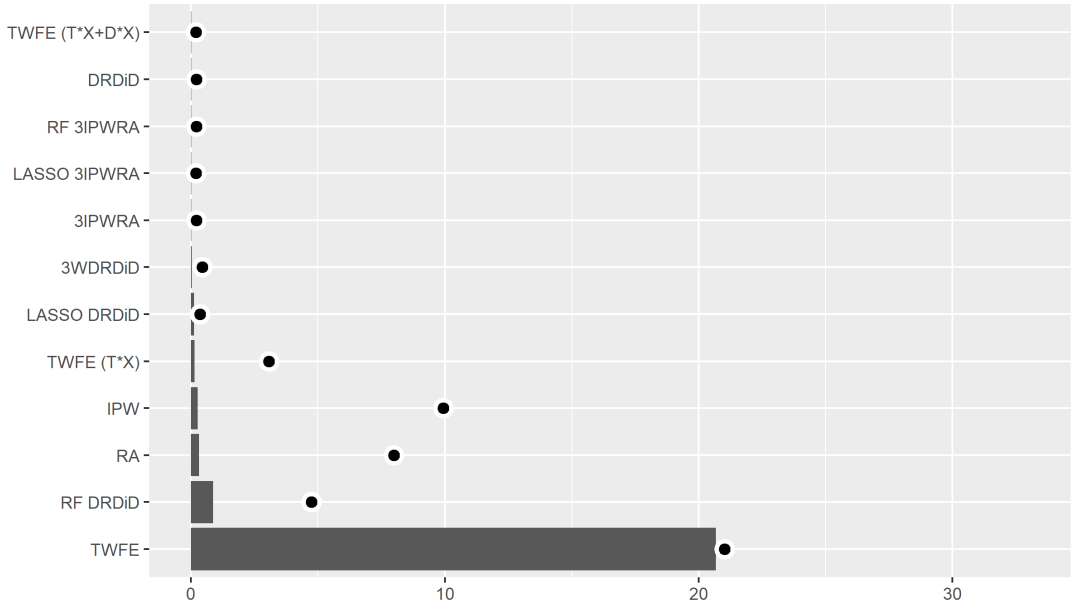


Notes: The graph considers a representative random sample from Exp.1 with DPG D. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the 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.

Table 4: Exp.1A Propensity score model correct, outcome regression correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-20.686	21.021	13.991	0.001
Regression, eq. (14)	TWFE (T·X)	-0.131	3.066	9.386	0.001
Regression, eq. (15)	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. (32)	3IPWRA	0.007	0.203	0.041	0.045
author's work, eq. (32)	LASSO 3IPWRA	0.007	0.190	0.036	0.492
author's work, eq. (32)	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

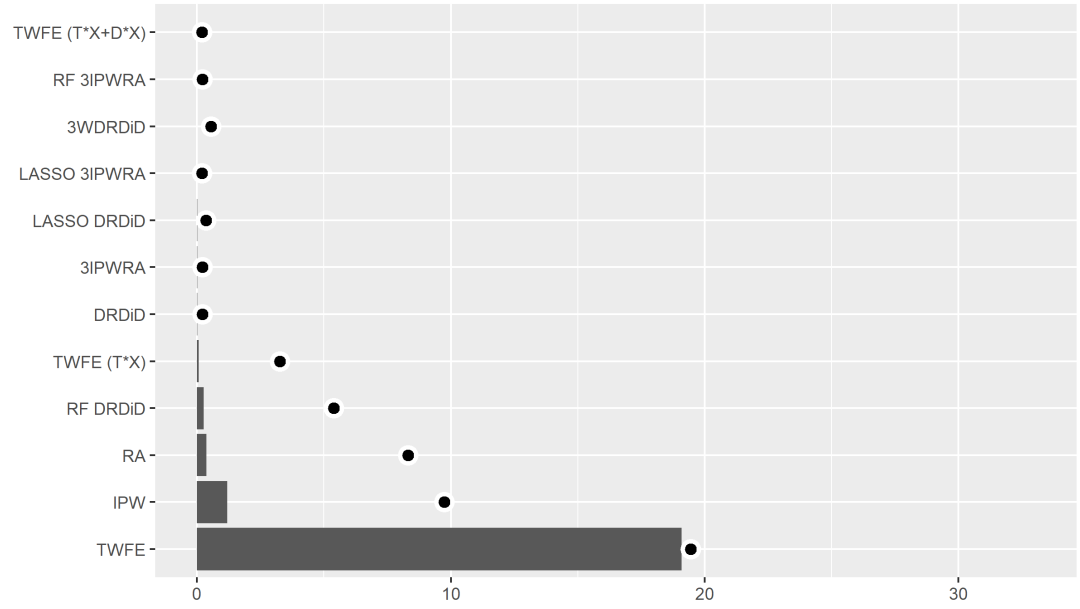


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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.1B Propensity score model incorrect, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-19.095	19.449	13.670	0.001
Regression, eq. (14)	TWFE (T·X)	-0.066	3.254	10.586	0.001
Regression, eq. (15)	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. (32)	3IPWRA	0.006	0.207	0.043	0.039
author's work, eq. (32)	LASSO 3IPWRA	0.003	0.197	0.039	0.522
author's work, eq. (32)	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

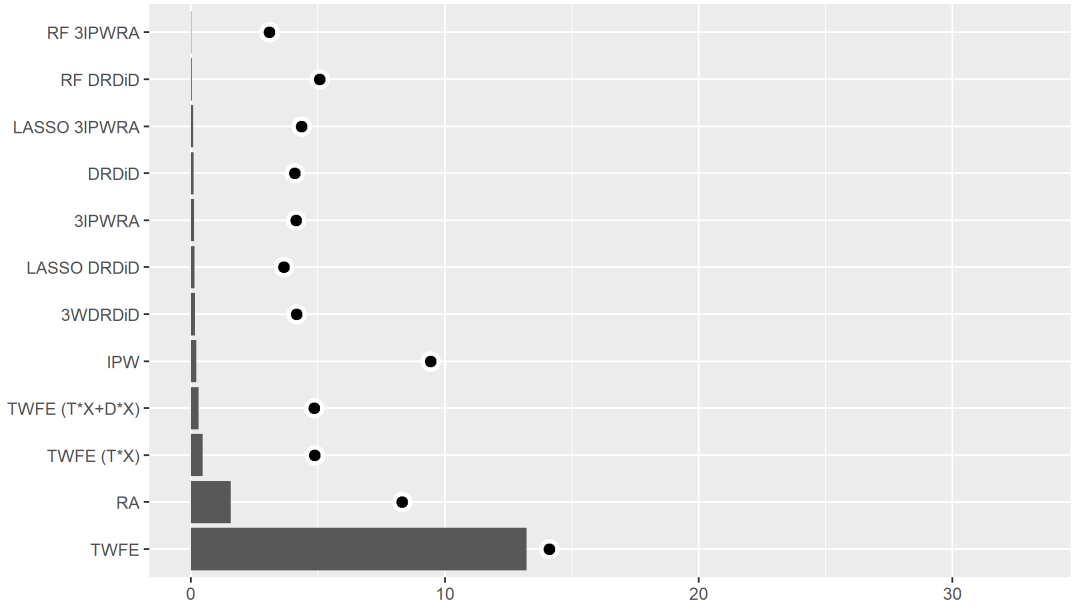


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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: Exp.1C Propensity score model correct, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-13.220	14.120	24.609	0.001
Regression, eq. (14)	TWFE (T·X)	-0.459	4.875	23.559	0.002
Regression, eq. (15)	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. (32)	3IPWRA	-0.113	4.134	17.081	0.020
author's work, eq. (32)	LASSO 3IPWRA	-0.078	4.347	18.894	0.518
author's work, eq. (32)	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

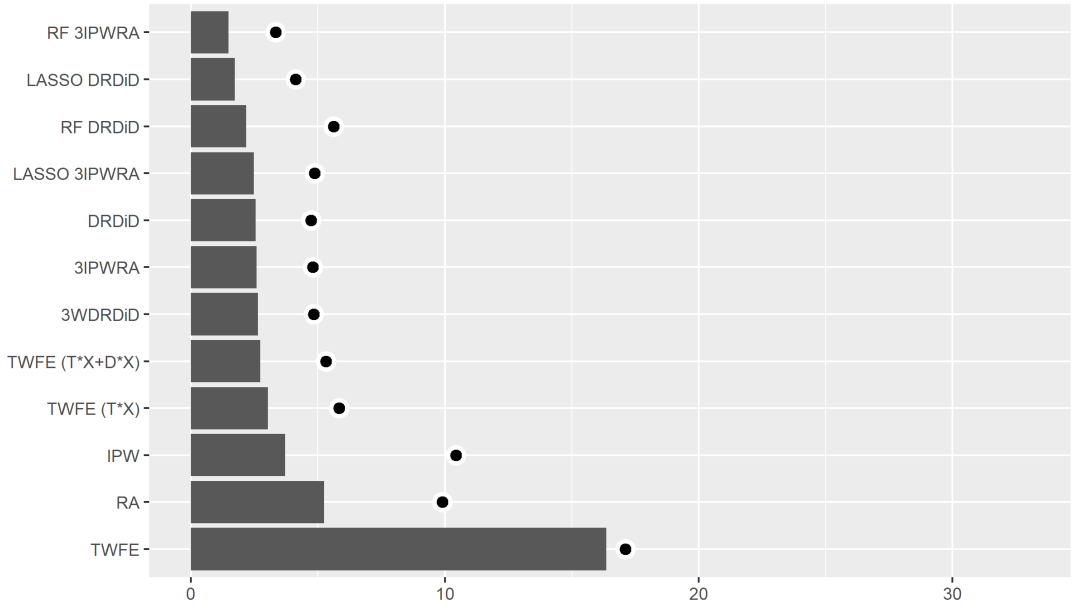


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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 7: Exp.1D Propensity score model incorrect, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-16.355	17.106	25.127	0.001
Regression, eq. (14)	TWFE (T·X)	-3.030	5.838	24.904	0.001
Regression, eq. (15)	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. (32)	3IPWRA	-2.578	4.787	16.267	0.026
author's work, eq. (32)	LASSO 3IPWRA	-2.471	4.860	17.514	0.514
author's work, eq. (32)	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'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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.

In Experiment 1, the weaknesses of the TWFE specification with covariates are apparent. Independently of the correct or incorrect specification of the propensity score and outcome regression models, TWFE is severely biased, reaching of -20.686 even in the most favourable scenario embodied by DPG A. However, as outlined by Zeldow and Hatfield (2019), in the case the conditional independence assumption is satisfied by observing time-invariant covariates, TWFE fixed effect can be corrected by including the interactions between these stationary covariates and the time dummy. In experiment 1A, where the propensity score and outcome regression are correctly specified, the correction works properly, but its performance, despite offering a great improvement, gradually worsens when for the researcher is to possible to correctly specify the models. In such a scenario, different 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 outperform other estimators, in particular the IPW and RA which are less efficient, especially in terms of variance.

3.3 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 period. In addition, in this design the treatment effects are allowed to vary for different values of X. Section 2.4 highlighted that such setting is a real threat to identification. Indeed, including time-varying covariates causes bias in the TWFE if either the variation in X between time periods is not the same between and treated and control or the effect of the covariates varies over time. As a consequence, time-varying covariates cannot be used to satisfy the conditional parallel trend assumption, since when there are X-specific trend, the effect of the covariate that determines the trend is time-varying, causing bias. Likewise, allowing for heterogeneous effects may invalidate its estimates as well, as discussed in section 2.4. In the experiment the four DPGs are denoted in the following way:

DPG.A (PS and OR models correct)

$$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0$$

$$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \delta(Z, D) + \epsilon_1(d)$$

$$p(Z) = \frac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$$

$$\lambda(Z) = \frac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$$

$$D = 1\{p(Z) \geq U_d\}$$

$$T = 1\{\lambda(Z) \geq U_t\}$$

DPG.B (PS model incorrect, OR correct)

$$Y_0^0 = f_{reg}(Z) + \upsilon(Z, D) + \epsilon_0(d)$$

$$Y_1^d = 2 \cdot f_{reg}(Z) + \upsilon(Z, D) + \delta(Z, D) + \epsilon_1(d)$$

$$p(X) = \frac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$$

$$\lambda(X) = \frac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$$

$$D = 1\{p(X) \geq U_d\}$$

$$T = 1\{\lambda(X) \geq U_t\}$$

DPG.C (PS model correct, OR incorrect)

$$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$$

$$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \delta(Z, D) + \epsilon_1(d)$$

$$p(Z) = \frac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$$

$$\lambda(Z) = \frac{\exp\left(f_{ps}(Z)\right)}{\left(1 + \exp\left(f_{ps}(Z)\right)\right)}$$

$$D = 1\{p(Z) \geq U_d\}$$

$$T = 1\{\lambda(Z) \geq U_t\}$$

DPG.D (PS and OR models incorrect)

$$Y_0^0 = f_{reg}(X) + \upsilon(X, D) + \epsilon_0$$

$$Y_1^d = 2 \cdot f_{reg}(X) + \upsilon(X, D) + \delta(Z, D) + \epsilon_1(d)$$

$$p(X) = \frac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$$

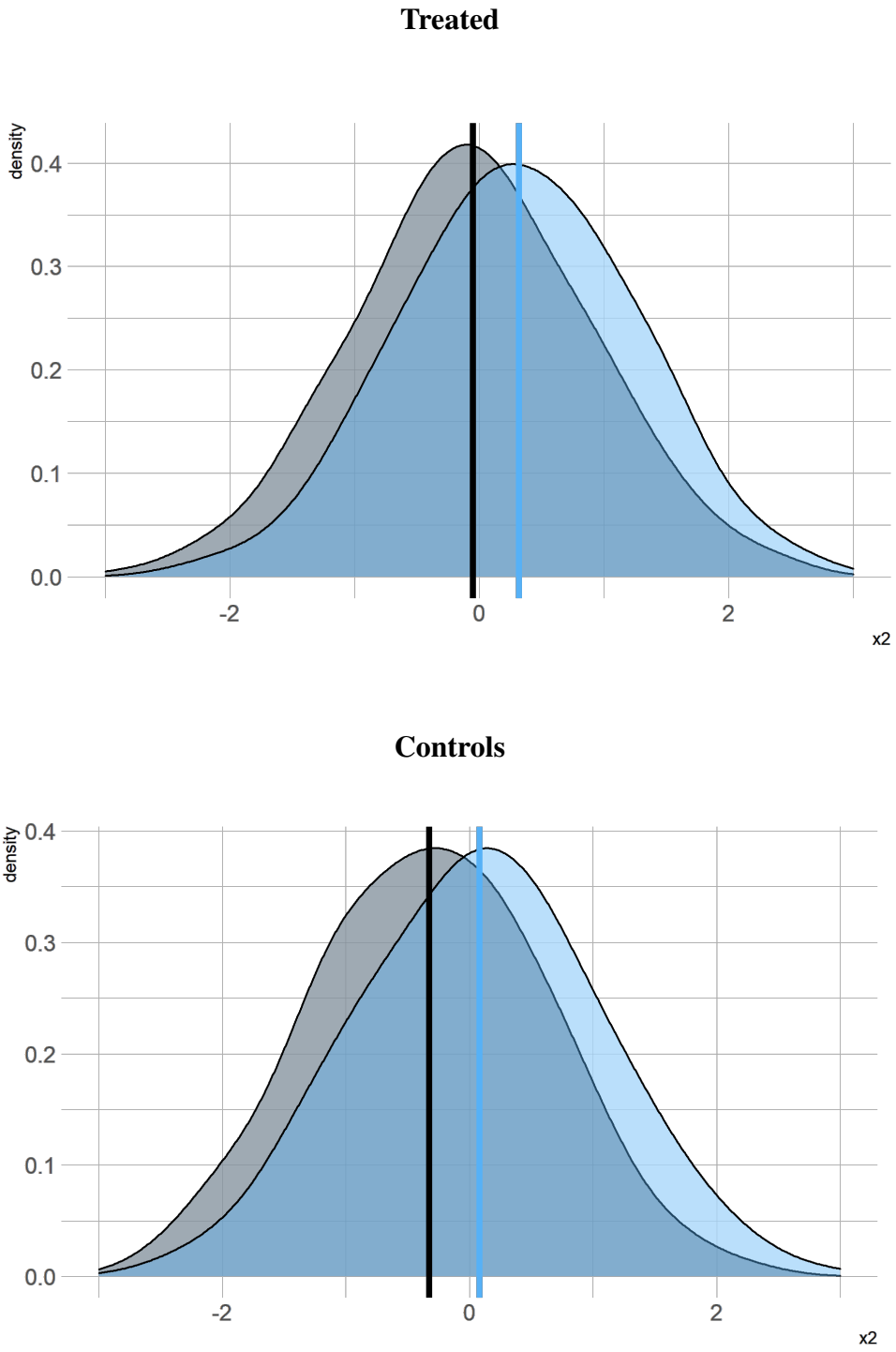
$$\lambda(X) = \frac{\exp\left(f_{ps}(X)\right)}{\left(1 + \exp\left(f_{ps}(X)\right)\right)}$$

$$D = 1\{p(X) \geq U_d\}$$

$$T = 1\{\lambda(X) \geq U_t\}$$

In this design, the DPGs were subject to two main changes. The first is the creation of a selection parameter through time, which replicates the mechanism of the propensity score in designating treated individuals among the population, but it acts in a different dimension, namely time. This is achieved, instead of defining λ as a constant parameter, by denoting $\lambda(W)$ as a function of the generic variable W. Because of the desirable properties of $f_{ps}(W)$ in terms of causing a realistic but sizeable differences in the distribution of covariates (Kang and Schafer, 2007), the same function and logistic transformation is applied for selecting which variables are observed in the pre and post-treatment periods. Figure 3 shows the resulting distribution in the representative covariates X_2: treated and controls group are heterogeneous, but also within each of both groups there is heterogeneity between $T = 0$ and $T = 1$. The second change is the definition of $\overline{\delta}(W, D) = -10W_1 + 10W_2 - 10W_3 - 10W_4$, which is the function that defines the heterogenous effects. In the DPG, I used the demeaned quantity $\delta(W, D) = \overline{\delta}(W, D) - E_{i|D=1}[\overline{\delta}(W, D)]$, where $E_{i|D=1}[\overline{\delta}(W, D)]$ denoted the average effect among the treated units. The ATT is therefore approximately zero, but the results accounts also for the small deviations from that value in each repetition of the Monte Carlo simulation. The results of the simulations are displayed in Table 8, Table 9, Table 10 and Table 11.

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

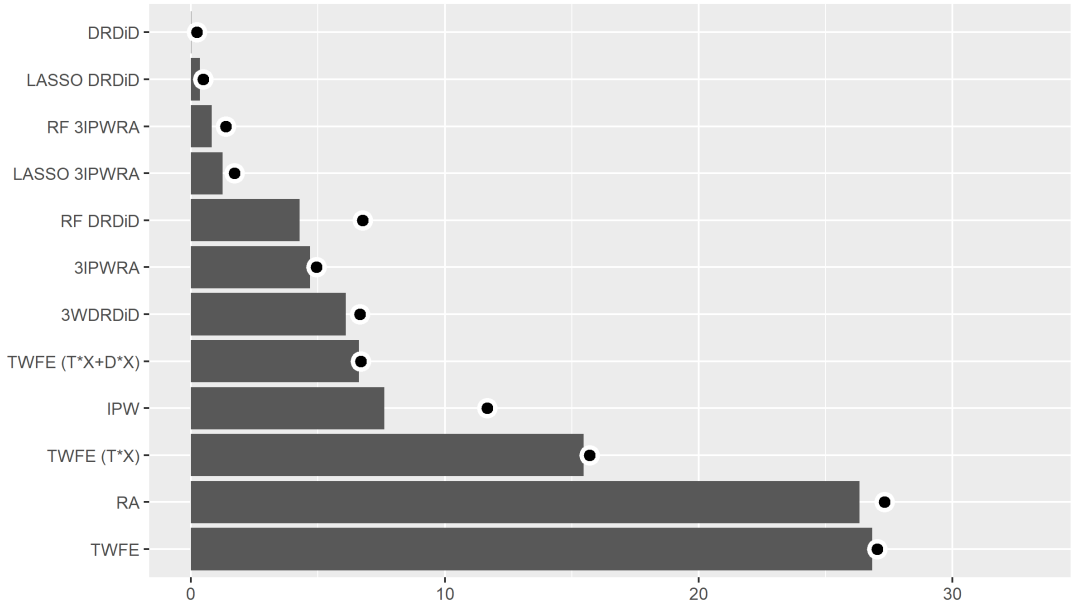


Notes: The graph considers a representative random sample from Exp.2 with DPG D. The upper plot compares the distribution of covariate X_2 among the treated, while the lower one among controls. The black vertical line represents the mean of the distribution of X_2 in the pre-treatment period among the 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.

Table 8: 2A Propensity score model correct, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-26.834	27.033	10.691	0.001
Regression, eq. (14)	TWFE (T·X)	-15.475	15.694	6.836	0.001
Regression, eq. (15)	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. (32)	3IPWRA	4.681	4.937	2.456	0.024
author's work, eq. (32)	LASSO 3IPWRA	-1.244	1.706	1.363	0.566
author's work, eq. (32)	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

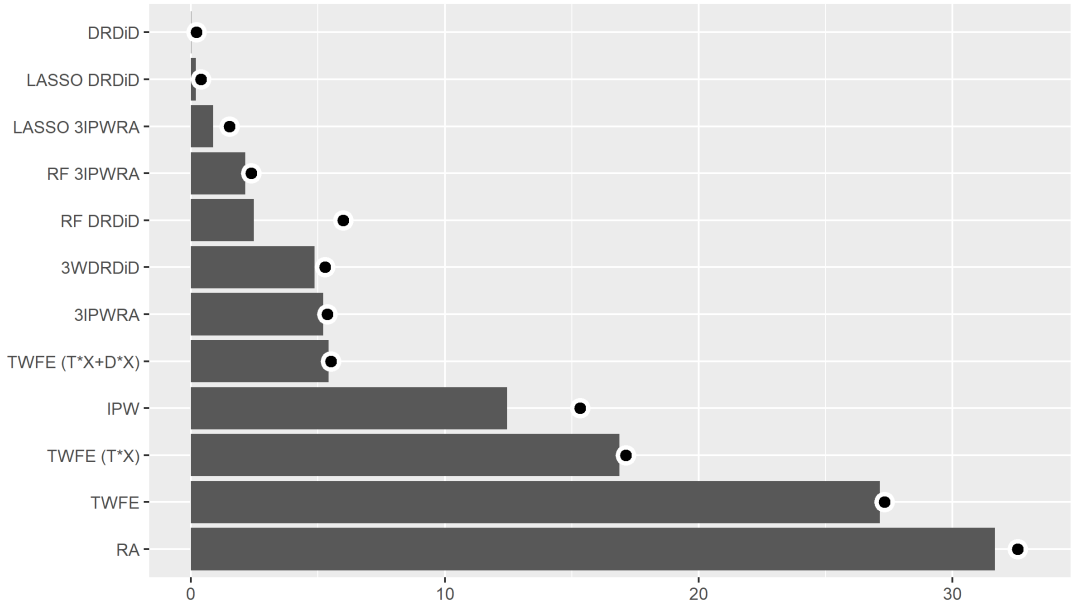


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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: 2B Propensity score model incorrect, outcome regression model correct

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-27.136	27.329	10.517	0.001
Regression, eq. (14)	TWFE (T·X)	-16.878	17.125	8.400	0.001
Regression, eq. (15)	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. (32)	3IPWRA	5.198	5.368	1.801	0.025
author's work, eq. (32)	LASSO 3IPWRA	0.876	1.507	1.503	0.582
author's work, eq. (32)	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

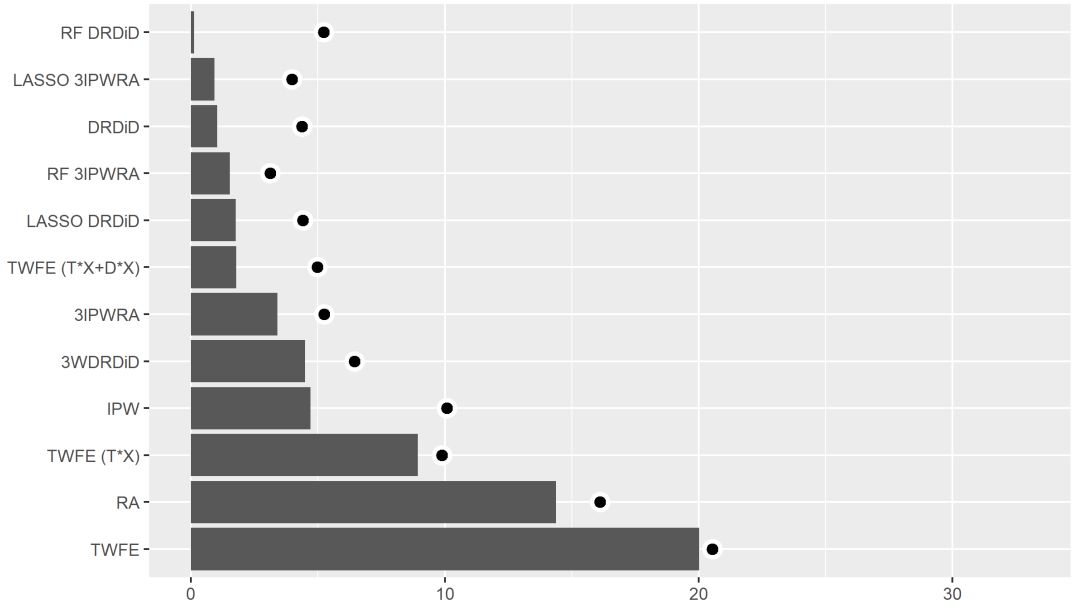


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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: 2C Propensity score model correct, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-20.013	20.545	21.604	0.001
Regression, eq. (14)	TWFE (T·X)	-8.938	9.891	17.948	0.001
Regression, eq. (15)	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. (32)	3IPWRA	3.404	5.244	15.914	0.016
author's work, eq. (32)	LASSO 3IPWRA	0.923	3.980	14.985	0.574
author's work, eq. (32)	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

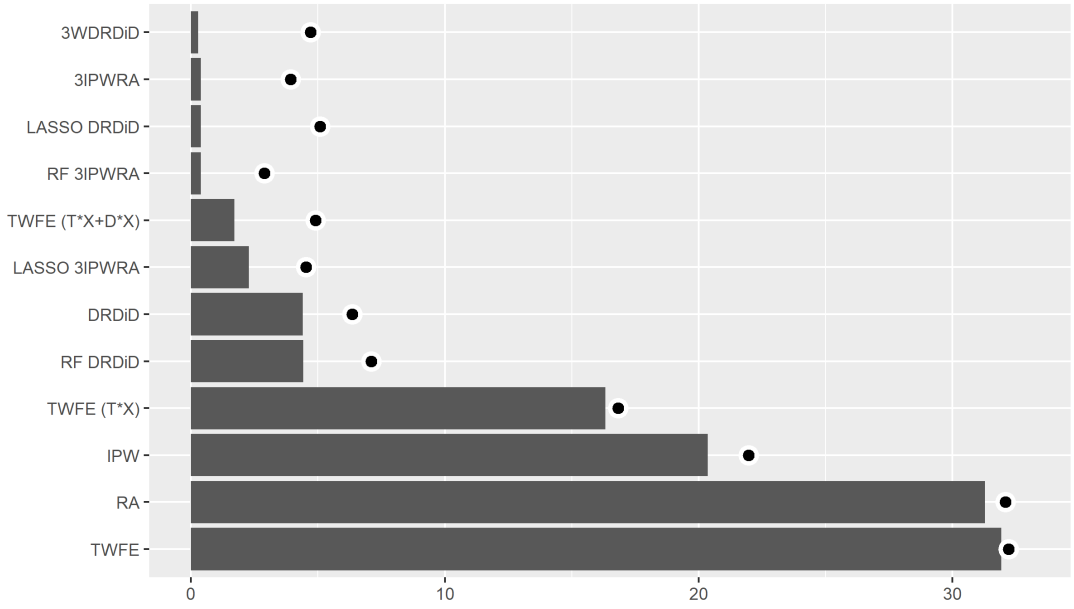


Notes: Simulations based on sample size $n = 1000$ and 500 Monte Carlo repetitions. The sign '*' stands for 'modified'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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: 2D Propensity score model incorrect, outcome regression model incorrect

Reference	Estimator	Bias	RMSE	Variance	Time
Regression, eq. (11)	TWFE	-31.917	32.212	18.890	0.001
Regression, eq. (14)	TWFE (T·X)	-16.331	16.832	16.641	0.001
Regression, eq. (15)	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. (32)	3IPWRA	-0.385	3.913	15.167	0.025
author's work, eq. (32)	3LASSO 3IPWRA	-2.278	4.521	15.252	0.569
author's work, eq. (32)	3RF 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'. TWFE is the standard regression specification with naively adding a set of covariates (eq. (11)), TWFE (T·X) is the regression specification that adds also the interaction terms between the covariates and the time dummy (eq. (14)). IPW is the inverse probability weighting (eq. (21)), RA is the regression adjustment approach (eq. (18)), DRDiD is the doubly robust estimator and it is proposed in three versions: the original from the paper (eq. (30)) 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. (32)). 3WDRDiD is the modified doubly robust estimator that uses triple matching weights (eq. (33)). 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.

The results of Experiment 2, since they are the most realistic setting, are particularly informative. In all DPGs, the traditional TWFE specification is severely biased, as well as the version including the time dummy and covariates interactions. However, controlling also for the interaction between covariates and treatment group dummy provides a substantial correction, even if still subject from the bias deriving from the heterogeneity of treatment effects. For example, in Exp.2D, when both propensity score and outcome regression models are incorrectly specified, this last version has a limited bias (-1.707) and RMSE (4.897). The IPW and RA, since they do not handle time-varying covariates, show relevant bias and variance, especially in the case of misspecified models. Contrarily, the doubly robust estimator DRDiD, despite relying on the same assumption of the two previous estimators, demonstrates to be particular robust to compositional changes when the underlying models are correctly computed. In Exp.2A, despite being an optimistic scenario, 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 sizeable (-4.402). Here indeed, other estimators work better. In addition to the lasso version of the DRDiD, which reduced bias close to zero, the 3WDRDiD, namely the modified version of the original DRDiD Sant'Anna and Zhao (2020) with the triple matching propensity score weights,

achieves the lowest bias (-0.285). However, similar or even better performances are showed by 3IPWRA and its alternative random forest version, which have an almost identical degree of bias (-0.385 and 0.390 respectively) but are more efficient and show the lowest overall RMSE (3.913 and 2.893 respectively). Overall, the 3IPWRA estimator yielded evidence of being the more robust estimator in terms of bias and RMSE in case of misspecified propensity score and outcome regression models, which is the more likely scenario in which researchers operates. However, when using methods that do only rely exclusively on pre-treatment levels of the covariates, such as IPW, RA and DRDiD, researchers must pay attention of not including bad controls, which are likely cause additional bias.

A useful empirical strategy may therefore be to combine the estimation methods that showed the lowest bias in the Monte Carlo simulation, such as the different versions of DRDiD and 3IPWRA, since they rely on different assumptions. Instead, when utilizing TWFE, the Monte Carlo simulation has given substantial evidence of the need of controlling for the interactions between treatment group and covariates, and, only when possible, of the time dummy and covariates interactions.

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

In this example, I reproduce the paper of Sequeira (2016) who analyzed the effect of tariff reduction on corruption behaviors by using the bribe payment data between South Africa and Mozambique. This contribution enters in a vivid 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 between the ports in Mozambique and South Africa from 2007 to 2013. As previously documented in Sequeira and Djankov (2014), it was widespread that cargo owners, in exchange for tariff evasion, or simply to avoid the threat

of being cited for real or fictitious irregularities, bribed the border officials who were 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, they constitute a credible control group for those that did and enable the use of a Difference-in-Difference design to isolate the causal relationship between tariffs and corruption. Indeed, the author, after pooling together the cross-section data between 2007 and 2013 for a total of 1084 observations, uses the canonical TWFE estimator in the following specification:

$$\begin{aligned} yit = \gamma_1(TariffChangeCategory_i \times POST) + \mu POST \\ + \beta_1 TariffChangeCategory_i + \beta_2 BaselineTariff_i \\ + \Gamma_i + p_i + +\omega_i + \delta_i + \epsilon_{it} \end{aligned} \tag{54}$$

where y_{it} represents the natural log of the amount of bribe paid for shipment i in period t, conditional on paying a bribe, $TariffChangeCategory \in \{0, 1\}$ takes value one if the com-

modity was subject to tariff reduction, $POST \in \{0, 1\}$ denotes the years following 2008, and $BaselineTariff$ is the tariff rate before the tariff reduction. 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 12. Industry, year, and clearing agent fixed effects are included controlling for p_i, ω_t, and δ_i respectively. The parameter of interest is the coefficient of the interaction between the time and treament dummy, namely γ_1.

Table 12: Variables included in Γ_i

	Description
diff	If the product have differentiated prices among countries
agri	If the product is am 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	Product 4-digits HS code

The main result of Sequeira (2016) is that the tariff reduction led to a drop in the amount of bribe paid. However, we have previously shown in Section 2.4 that the standard TWFE is likely to be biased under a non-randomized experiment because of the possible presence of X-specific trends, compositional changes, heterogenous effects and non-linearities. 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 13 summarizes the results obtained in the two papers:

Table 13: The effect of tariff reduction on bribes

	TWFE Sequeira (2016)	TWFE ($\Gamma_i \times POST$) 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 our-digit HS code.

where TWFE is the standard specification in Sequeira (2016), which is found in Equation 1 of Table 9 of the paper, while TWFE ($\Gamma_i \times$ POST) is the specification that adds also 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 that employs a kernel estimation or a lasso. The semiparametric estimates therefore give evidence

that the effect of the reduction was of higher magnitude than originally thought. However, such estimates have weaknesses that question their validity. First, they suffer from very high standard error which really blurs the interpretation of its results. For example, the 95 percent confidence interval lies approximately in between 0.318 and -14.306 for the kernel DMLDiD, and the same applies to its lasso version, even if in a smaller degree. This is consistent with the preliminary results encountered by including DMLDiD in the Monte Carlo simulations, since the estimator seemed to show a significantly higher variance than the others. In addition, simulations in Section 3 showed that the IPW estimator can be severely biased under realistic setting and was outperformed by other estimators. Finally, DMLDiD requires a substantial amount of computational time, and may not feasible with very large datasets. Motivated by these reasons, I employ the estimators proposed in Section 2 to the current setting. Final estimates are shown in Table 14.

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 has conspicuous observations for the group of the control in the post-treatment period, but limited observations for the other three groups, namely the treated

Table 14: The effect of tariff reduction on bribes

	TWFE Sequeira (2016)	TWFE($\Gamma_i \times POST$) Sequeira (2016)	TWFE($\Gamma_i \times POST$+$\Gamma_i \times D$)
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. (32)) 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.

in the pre and post-treatment period (120 and 56 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. 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 created for computational tractability. The ATT is estimated using both lasso and random forest 3IPWRA, and the lasso version of DRDiD, since all of them showed good performance in the Monte Carlo simulations in Section 3. 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 associated at a cluster and not an individual level.

Among different methods and specifications, our results corroborate the hypothesis that the tariff reduction led to a drop in the amount of bribe paid, but gives 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 that showed the best results in our simulation converge to lower values: TWFE($\Gamma_i \times POST + \Gamma_i \times D$) estimates -3.667, lasso 3IPWRA -3.023, random forest 3IPWRA -3.216, and lasso DRDiD -2.764. The standard errors are significally lower than those in Chang (2020), yielding stronger evidence in favour of our estimates. Therefore, our results reveal that the tariff reduction had a significant but lower effect on bribing behavior than originally estimated by the standard TWFE specification and by DMLDiD as in Chang (2020). The average of our estimates is -3.167, closer to TWFE with the correction in Sequeira (2016).

5 Conclusion

Through analytical derivations and empirical simulations, the thesis showed that the commonly-used standard TWFE is severely biased under recurrent settings. In the Monte Carlo simulations, we assessed the performance of TWFE corrections and other semi-parametric estimators. Despite including both time and treatment group interactions with the covariates provides a substantial correction, TWFE is outperformed by other semi-parametric methods, such as DRDiD and 3IPWRA. The first has better performances in terms of bias and mean square error when both the propensity score and outcome models are correctly specified. However, in the more realistic case of when they are not, 3IPWRA outperforms DRDiD in the case of compositional changes, even if modified versions of the latter, such as allowing for lasso first-stage estimates or using its modified specification 3WDRDiD, yielded interesting results as well. Therefore, a useful strategy in empirical settings may be to compare the different versions of 3IPWRA, DRDiD, and corrected TWFE since they rely on different assumptions. Indeed if on the one hand DRDiD is originally built to handle time-invariant controls, 3IPWRA and the proposed versions of TWFE may be subject to bad controls if the covariates are not accurately handled. In addition, TWFE may be less precise in case of particularly severe heterogeneity in treatment

effects. Thus, in case those methods converge on similar results, this can yield strong evidence in favor of a hypothesis. Having this in mind, the strategy is replicated to estimate the effect of tariff reduction on bribes, as in Sequeira (2016). As in Chang (2020), the estimates in the thesis found that tariff reduction led to a decrease in bribes paid, but on the other hand they assess that the effect is close and even lower in magnitude than the one of the original paper. Further research may encompass and analyze other possible DiD estimators, such as the ones of Nie et al. (2021) and Zimmert (2018), which offer interesting alternatives and follow 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.

Orley Ashenfelter. Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, pages 47–57, 1978.

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, 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 "Missing Imports" 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.