TOMMASO MANFÈ

University of Chicago

LUCA NUNZIATA

University of Padova and IZA

DIFFERENCE-IN-DIFFERENCE DESIGN WITH REPEATED CROSS-SECTIONS UNDER COMPOSITIONAL CHANGES: A MONTE-CARLO EVALUATION OF ALTERNATIVE APPROACHES

November 2023 (First version: May 2023)

Marco Fanno Working Papers – 305

dSEA Dipartimento di Scienze Economiche e Aziendali 'Marco Fanno'

Università degli Studi di Padova

Difference-In-Difference Design With Repeated Cross-Sections Under Compositional Changes: a Monte-Carlo Evaluation of Alternative Approaches*

Tommaso Manfè†[1] and Luca Nunziata‡[2,3]

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

First draft: December 22, 2022
This draft: June 6, 2023

Abstract

We discuss the potentially severe bias in the commonly-used Difference-in-Difference estimators under compositional changes and propose a Double Inverse Probability Weighting estimator for repeated cross-sections based on both the probability of being treated and of belonging to the post-treatment period, deriving also its doubly-robust version. Through Monte Carlo simulations, we compare its performance with several methods suggested by the literature. Results show that our proposed estimator outperforms all alternatives in the most realistic scenarios. We provide an empirical application estimating the effect of tariff reduction on bribing behavior on trades data between South Africa and Mozambique during the period 2006–2014.

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

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

*This paper is a revised version of the work originally presented in Tommaso Manfè's Master Thesis in Economics and Finance at the University of Padua entitled "Beyond regression: evaluating different semi-parametric approaches and machine learning tools in the difference-in-difference design" submitted in February 2022 and available here. The coding material in its latest and intermediate versions can be downloaded from https://github.com/tommaso-manfe. We thank Enrico Rettore, the Master in Economics and Finance examination committee at the University of Padua, Scott Cunningham, Pedro H.C. Sant'Anna, Michael Weber, Francesco Ruggieri, Pietro Ramella, and other faculty members at the University of Chicago for their precious comments and advice. The usual disclaimer applies.

†tommaso.manfe@chicagobooth.edu

‡luca.nunziata@unipd.it

1 Introduction

Difference-in-Difference (DiD) is a widespread research design that estimates the causal effects of a policy treatment that affects a specific group of subjects, called the treated group, while leaving unaffected another typically comparable group, referred to as the control group. The rationale of this empirical strategy is that if treated and control groups are subject to the same time trend, the control group can be used to estimate the counterfactual potential outcome for the treatment group in the absence of treatment.

However, the parallel trend assumption is implausible if selection into treatment depends on individual characteristics that correlate with the outcome variable. A weaker assumption consists in assuming that the parallel trend hypothesis holds after conditioning on individual observable characteristics, the so-called "conditional trend assumption".

Despite being often applied in empirical studies, the Two-Way-Fixed-Effects (TWFE) regression is potentially biased when adding covariates to its specification. Even when assuming time-invariant covariates, Zeldow and Hatfield (2019) show that, when the effect of the covariate on the outcome varies over time, naively adding the covariates in the regression does not eliminate the bias. In addition, the TWFE model implicitly assumes homogeneous treatment effects in the covariates (Sant'Anna and Zhao, 2020), and it is subject to violations of the linear functional form.

To overcome these limitations, novel semi-parametric alternatives were proposed by the literature, but all still assume that the covariates are time-invariant between the two periods in which treatment takes place. Among these methods, Outcome Regression (OR) (Heckman et al., 1997) and Inverse Probability Weighting (IPW) (Abadie, 2005) are complementary methodologies where the first assumes a model for the outcome evolution and the second for the probability of receiving the treatment. More recently, Doubly-Robust DiD (DRDiD) (Sant'Anna and Zhao, 2020) combined the two methods into a doubly-robust estimand, which has the property of identifying the causal effect also when just one of the two models is correctly specified.

However, practitioners are often exposed to settings where the distribution of the covariates changes from the pre to the post-treatment period. Under this scenario, also referred to as compositional changes, all the mentioned estimators may deliver biased estimates of the causal effect. In this paper, similarly as in Caetano et al. (2022), we evaluate a set of novel estimators that allow the distribution of the covariates to change over time but, differently from their analysis, we derive methods suited for repeated cross-sections. The main difference from the panel data case is that the researcher does not directly observe the first difference neither in the observed outcome nor in the covariates for a given individual. As a consequence, in this setting the researcher needs to account for the potential heterogeneity between the pre- and post-treatment cross-sections. To this aim, we develop a Double Inverse Probability Weighting (DIPW) scheme based on both the probability of being treated and that of belonging to the post-treatment period and derive its doubly-robust version (DR-DIPW).

In addition, in order to relax the parametric assumptions of the proposed estimators, we propose and test alternative versions of DR-DIPW that employ machine learning algorithms for the first-stage estimates, following Chernozhukov et al. (2018). In particular, since doubly-robust estimands satisfy the Neyman orthogonality condition needed for debiased machine learning, we construct estimators based either on lasso or random forests for the first-stage estimates. Our analysis evaluates also the debiased orthogonal extension of Abadie's semiparametric DiD estimator (DMLDiD) proposed by Chang (2020). In general, relaxing the parametric assumptions for the models under study is beneficial since the researcher is typically unaware of their correct functional form, which may result in model misspecification.

To corroborate our findings, we conduct a series of Monte Carlo simulations in order to investigate the finite sample properties of all semi-parametric estimators presented above. The simulations allow for trends that depend on the individual level of the covariates, for heterogeneous treatment effects, and, most importantly, for different evolutions of the covariates between treated and control groups. Therefore, they provide practical guidance for settings common to practitioners.

Our analysis confirms that the commonly-used DiD methods may be severely biased under recurrent settings. Overall, our proposed DR-DIPW method is approximately unbiased in most of the specifications and outperforms the alternative estimators in terms of bias. When we assume that the researcher cannot correctly specify both the propensity scores and the outcome models, the machine learning versions of DR-DIPW have even lower bias, since they do not assume an *a prori* parametric form for the DGP. Finally, the simulations show that TWFE corrections drastically reduce the bias, but are outperformed by DR-DIPW and in general by doubly-robust estimators.

We apply our novel estimators to the analysis in Sequeira (2016), which investigates the effect of tariff reduction on corruption behaviors by using bribe payment data on the cargo shipments transiting from South Africa into the ports in Mozambique. Overall, we provide strong evidence against the results of the replication produced by Chang (2020), since our findings show that the effect is close and even lower in magnitude than the traditional TWFE estimation presented in the original paper.

The paper is organized as follows: Section 2 presents the baseline features of the DiD and it illustrates TWFE and alternative semi-parametric estimators, including DIPW and DR-DIPW; Section 3 implements 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 a discussion of our most relevant findings.

2 Identification

2.1 Notation and Setup

We study the baseline case where the researcher has access to two time periods of repeated cross-sections. Define the treatment variable D, where d $\in \{0, 1\}$,[1] as the binary indicator for whether the individual i belongs to the treated group, where the i subscript is dropped for ease of notation. Similarly, define T, where t $\in \{0, 1\}$, as the binary indicator that takes value zero in the pre-treatment period and one in the post-treatment period. Since the treatment is assumed to take place in between the two periods, every member of the population is untreated in the pre-treatment period. We define the potential levels of the outcome variable by using indexes that refer to the potential states of the treatment, so that $Y_{d,t}$ denotes the outcome that would be realized for a specific value of d in period t. However, for each individual only one potential outcome is observed at each time period. At $t = 0$, the treatment has no effect on the potential outcomes so that $Y_{1,0} = Y_{0,0} = Y_0$, where we refer to the realized outcome as Y_t (i.e., not indexed by d). At $t = 1$, instead, $Y_1 = dY_{1,1} + (1 - d)Y_{0,1}$, so we just observe the potential outcome of treatment for the treated and the potential outcome of in case of no treatment for the controls. Likewise, we denote as $X_{d,t}$ the potential level of the covariate for the treatment group d and time t, noting again that in the pre-treatment period $X_{1,0}, = X_{0,0}$. The object we are interested in estimating is the average effect on the treated (ATT),[2] which is defined as follows:

$$ATT = E(Y_{1,1} - Y_{0,1}|D = 1)$$

i.e. it is the average difference between treated and untreated potential outcomes among the treated population. The fundamental problem of causal inference is that $Y_{0,1}$ is not

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

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

observed and thus it must be imputed.

2.2 DiD With Time-Invariant Covariates

We initially review the existing literature on DiD with covariates, which has primarily focused on covariates that do not vary in distribution between pre- and post-treatment periods. Throughout the paper, we make the following set of assumptions.

Assumption 1.a. *(Sampling scheme) The pooled repeated cross-section data* $\{Y_i, D_i, X_i, T_i\}_{i=1}^n$ *consist of iid draws from the mixture distribution*

$$P(Y \leq y, D = d, X \leq x, T = t) = t \cdot \lambda \cdot P\left(Y_1 \leq y, D = d, X \leq x \mid T = 1\right) + (1-t) \cdot (1-\lambda) P\left(Y_0 \leq y, D = d, X \leq x \mid T = 0\right)$$

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

Assumption 1.a is the standard assumption among DiD that rules out compositional changes, namely a time-varying distribution of observables. Note that $T \perp\!\!\!\perp (D, X)$ is equivalent to (i) $X \perp\!\!\!\perp T | D$ and (ii) $D \perp\!\!\!\perp T$, i.e. (i) the observed covariates of individuals within a treatment group do not change over time, and (ii) the proportion of individuals belonging to the treatment group does not vary over time. The sampling scheme allows for each observation to be randomly chosen from either (Y_0, D, X) or (Y_1, D, X) with fixed probability λ. Note also that since the covariates are time-invariant, they are exogenous to the treatment. In Section 2.3, we relax Assumption 1.a to allow for compositional changes.

Assumption 2. *(Conditional Parallel Trend)*

$$E(Y_{0,1}|X, D = 1, T = 1) - E(Y_{0,0}|X, D = 1, T = 0) = E(Y_{0,1}|X, D = 0, T = 1) - E(Y_{0,0}|X, D = 0, T = 0)$$

Assumption 2, usually referred as "conditional parallel trend assumption", is key for the identification of causal effects in the DiD design. It states that, conditional on X, the average difference in untreated potential outcomes in the pre and post-treatment periods is the same between treated and control groups. Therefore, the inclusion of the covariates X as controls is aimed at capturing all variables that may cause different time trends. We emphasize that this is a more robust extension of the unconditional parallel trend assumption, which claims that the parallel trend holds even when not conditioning on the covariates X. However, this latter assumption seems unlikely to hold in practice.

Assumption 3. *(Common Support Treatment Score)* $P[D = 1|X] < 1 - \epsilon \quad a.s.$

for some $\epsilon > 0$, where we define $p(X) \equiv P[D = 1|X]$ as the treatment score. Assumption 3 implies that it is possible to observe individuals with characteristics X among both treated and controls. In other words, the conditional probability of belonging to the treatment group given X is uniformly bounded away from one, imposing that for every value of the covariates X there is at least a small chance that the unit is not treated, and in addition, the proportion of treated units is bounded away from zero, meaning that at least a small fraction of the population is treated.

2.2.1 Two-Way-Fixed Effect With Covariates

Frequently, practitioners use the following regression, usually referred to as Two-Way-Fixed Effect (TWFE) with covariates, to estimate the ATT in a DiD setting:

$$Y = \alpha + \gamma T + \beta D + \delta(T \cdot D) + X'\theta + \epsilon \tag{1}$$

where $X = (X_1, X_2, ..., X_p)'$ is the set of covariates with coefficients $\theta = (\theta_1, \theta_2, ..., \theta_p)'$, γ is the constant time effect between T=0 and T=1, β represents the treatment-group fixed effect, namely the differential in the potential outcome between treated and controls, and δ represents the effect of the treatment. This specification implicitly imposes, even when the covariates are time-invariant, three additional restrictive assumptions: (i) the coefficients of the covariates do not vary over time if the treatment is not randomized (X-specific trends), (ii) homogeneous treatment effects in X, and (iii) additive linear form of how the covariates affect the outcome. Each of the three cases is discussed in detail in Appendix A.1.

The standard TWFE specification can be improved by allowing some corrections. Zeldow and Hatfield (2019) argue that by adding an interaction between the time dummy T and the time-invariant covariates X, the confounder effect of the covariates in presence of homogeneous treatment effects in X can be eliminated. A natural extension that allows for both X-specific trends and heterogenous effects can be written as:

$$Y = \alpha + \gamma T + \beta D + \delta(TD) + X_i'\theta + (TX')\omega + (DX')\nu + (TDX')\rho + \epsilon \qquad (2)$$

where the interaction term $T \cdot D \cdot X'$ explicitly allows for the effect of the treatment to change depending on the level of X.

2.2.2 Outcome Regression

The outcome regression (OR henceforth) approach relies on the specification of a model for the evolution of the outcome of interest given X. In Assumption 2, two conditional expectations refer to the untreated group and must be predicted for the treated sample. To do so, an outcome model is estimated on untreated units for both $T = 0$ and $T = 1$, and then fitted values are predicted using the empirical distribution of X among treated units. Usually, the outcome model is estimated through regression, but other more flexible non-parametric methods can be employed as well. More formally, following Heckman et al. (1997), starting from the definition of the ATT under conditional parallel trends and using the law of iterated expectations, we obtain:

$$ATT = E[E(Y_1 - Y_0|X, D = 1) - E(Y_1 - Y_0|X, D = 0)|D = 1]$$
$$= E(Y_1 - Y_0|D = 1) - E[E(Y_1 - Y_0|X, D = 0)|D = 1] \quad (3)$$

where the first term in Eq. (3) can be computed by taking sample averages, while the second expected value must be estimated. One way of estimating it is by fitting a regression on the controls group data and taking predictions based on the empirical distribution of X among treated units. More formally:

$$\delta^{OR} = \bar{Y}_{1,1} - \bar{Y}_{1,0} - \left[\frac{1}{n_{treat}} \sum_{i|D_i=1} (\hat{\mu}_{0,1}(X_i) - \hat{\mu}_{0,0}(X_i))\right] \quad (4)$$

where $\bar{Y}_{d,t} = \sum_{i|D_i=1} Y_{it}/n_{d,t}$ is the sample average outcome among treated units in treatment group d at time t, and $\hat{\mu}_{d,t}(X)$ is an estimator of the true, unknown $m_{d,t}(x) \equiv E[Y_t|D = d, X = x]$. Intuitively, when using a linear specification for $\hat{\mu}_{d,t}(X)$, the model would be close to the version of TWFE with covariates that includes also all the interactions between X_i and both treatment group and time dummies, as in Eq. (2). The two models differ because the outcome regression approach adopts a re-weighting scheme based on the distribution of X among units with $D = 1$ (Roth et al., 2022). The condition for the consistency of the ATT of the outcome regression is the correct specification of $\hat{\mu}_{d,t}(X)$.

2.2.3 Inverse Probability Weighting

The Inverse Probability Weighting (IPW) approach proposed by Abadie (2005) avoids the direct modeling of the outcome evolution. Its focus is on the treatment model for $p(X) \equiv P(D = 1|X)$, which is the conditional probability of treatment given X. The idea of the IPW estimator is to adjust for confounding factors using the propensity score to balance individual characteristics in the treated and untreated groups. When dealing with repeated cross-sections, we can retrieve the $ATT = E(Y_{1,1} - Y_{0,1}|D = 1)$ by the following estimand:

$$\delta^{IPW} = \frac{1}{E(D)\cdot\lambda}\cdot E\left[\frac{D-p(X)}{1-p(X)}\cdot\frac{T-\lambda}{1-\lambda}\cdot Y\right] \tag{5}$$

which can be estimated using the following sample analog:

$$\delta^{IPW} = \frac{1}{\lambda\cdot\frac{1}{n}\sum_{j=1}^{n}(D_j)}\cdot\sum_{i=1}^{n}\left[\frac{D_i-\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\cdot\frac{T_i-\lambda}{1-\lambda}\cdot Y_i\right] \tag{6}$$

Intuitively, IPW produces a weighting scheme that weights-down the observed outcome $Y_{d,t}$ for the individuals with covariate values over-represented among their time and treatment group category, and weights-up the observed outcome for the individuals with covariate values under-represented among their group. Consequently, the adjustment balances the distribution of covariates between treated and untreated groups. The unknown propensity score $p(X) = P(D = 1|X)$ is usually estimated by means of logistic regression or a linear probability model, even if non-parametric models can be employed as well. The IPW approach will generally be consistent when the propensity score model is correctly specified.

2.2.4 Doubly Robust Difference-in-Difference

Sant'Anna and Zhao (2020) combine the OR and the IPW approaches into a doubly robust estimand for the ATT. The double robustness property means that if either the propensity score model or the outcome regression models are misspecified (but not both), the resulting estimand still identifies the ATT. Denote $\mu_{d,t}(X)$ as the arbitrary model for the true, unknown CEF $m_{d,t}(x) \equiv E[Y|D = d, T = t, X = x]$, and for ease of notation define $\mu_{d,Y}(T, X) \equiv T\cdot\mu_{d,1}(X) + (1-T)\cdot\mu_{d,0}(X)$, where recall $d, t \in \{0, 1\}$. Intuitively, $\mu_{d,Y}(T, X)$ represents the outcome model for a given treatment group $D = d$. Then the estimand is defined as:

$$\delta_1^{dr} = E\left[\Big(\omega_1(D,T) - \omega_0(D,T,X;p)\Big)\Big(Y - \mu_{0,Y}(T,X)\Big)\right] \tag{7}$$

where:

$$\omega_1(D,T) = \omega_{1,1}(D,T) - \omega_{1,0}(D,T)$$

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

and for t $\in \{0,1\}$:

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

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

The relative sample analog is obtained by replacing $p(x)$ with $\hat{\pi}$ and the expectation with sample means. The first term of δ_1^{dr} represents the IPW weighting scheme based on the propensity score, while the second term represents the outcome regression part of the estimand. Sant'Anna and Zhao (2020) present also a locally semi-parametrically efficient version of the above estimator, which is characterized by an asymptotic variance that achieves the semi-parametric efficiency bound when the propensity score and outcome regression are correctly specified:

$$\delta_2^{dr} = \delta_1^{dr} + (E[\mu_{1,1}(X) - \mu_{0,1}(X)|D=1] - E[\mu_{1,1}(X) - \mu_{0,1}(X)|D=1, T=1])$$
$$- (E[\mu_{1,0}(X) - \mu_{0,0}(X)|D=1] - E[\mu_{1,0}(X) - \mu_{0,0}(X)|D=1, T=0]) \quad (8)$$

In the Monte Carlo simulations in Section 3, we consider only the estimand δ_2^{dr}. The outcome equation and the propensity score can be modeled either parametrically, for instance with a linear and logistic regression respectively, or non-parametrically. In the first case, the authors name the estimator DRDiD. The authors also use the inverse probability tilting estimator (Graham et al., 2012) for the treatment model and weighted least-squares for the outcome model. In this case, they name the estimator Improved DRDiD (IMP DRDiD). We will stick to this definition in the remaining sections. The two estimators will generally be consistent if either the propensity score or the outcome model is correctly specified.

2.3 DiD Under Compositional Changes

Despite researchers more often observe time-varying covariates in empirical settings, just a few papers consider this setup. This is problematic since time-varying covariates can act as a confounder in DiD settings. For example, Zeldow and Hatfield (2019) shows that in this context TWFE retrieves the ATT only on implausible assumptions and just adding the covariate as a control is ineffective since it imposes the coefficient on the covariate to be constant over time (see Appendix A.1 for more details). Moreover, all methods considered above, i.e. OR, IPW, and DRDiD, assume time-invariant covariates and therefore might be biased from confounding effects when instead the distribution of the covariates varies over time.

Among the few papers focussing on compositional changes, Caetano et al. (2022) propose a doubly robust estimand for panel data which controls both for the pre- and post-treatment levels of the covariate and outline specific assumption under which the ATT is identified. Instead, Hong (2013) proposes a two-variate matching estimator for repeated cross-sections based on the probability of being treated in the pre- and post-treatment periods, which are defined separately. We adopt a similar approach by specifying two propensity scores: one for the probability of being treated (treatment score), and the other for the probability of belonging to the post-treatment period (time score). We then construct an inverse probability weighting scheme based on both scores and derive its doubly-robust estimand. Independent and parallel work in Sant'Anna and Xu (2023) complements some of the theoretical findings of our proposed estimators. From now on, we rely on the following additional set of assumptions.

Assumption 1.b. *(Sampling scheme) The pooled repeated cross-section data*

$\{Y_i, D_i, X_i, T_i\}_{i=1}^n$ *consist of iid draws from the mixture distribution*

$$P(Y \leq y, D = d, X \leq x, T = t) = t \cdot \lambda \cdot P\left(Y_1 \leq y, D = d, X \leq x \mid T = 1\right) + (1-t) \cdot (1-\lambda) P\left(Y_0 \leq y, D = d, X \leq x \mid T = 0\right)$$

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 time-varying with respect to*** T.

Assumption 1.b replaces Assumption 1.a allowing covariates to be time-varying. Since now the covariates evolve over time, we cannot rule out the possibility that they are affected by the treatment.

Assumption 4. *(Common Support Time Score)* $P[T = 1|D, X] < 1 - \epsilon \quad a.s.$

for some $\epsilon > 0$. The time score $t(D, X) \equiv P[T = 1|D, X]$ is the probability of belonging to the post-treatment period conditional on the covariate X and the treatment group D. Intuitively, it allows for heterogeneous time-trends in the covariates among treated and untreated individuals. As for the treatment score, Assumption 4 ensures that, for any values of X, there will be some units in the post-treatment period for both treated and untreated units.

Assumption 5. *(Covariates exogeneity)* $X_{1,1} = X_{0,1} \quad \forall i : D_i = 1$

Assumption 5 states that in the post-treatment period the potential covariate level in case of treatment is equal to the potential covariate level in case of no treatment. This condition rules out bad controls, namely covariates affected by the treatment. Despite this may be a too restrictive assumption, in repeated cross-sections it is not possible to control for the pre-treatment levels of the covariates since each individual is observed in only one time period. Therefore, there is an inherent trade-off between correcting for the heterogeneity in the evolution of the covariates and allowing bad controls.

2.3.1 Double inverse-probability weighting (DIPW)

Building on Abadie (2005), we propose a weighting scheme that corrects for the heterogeneous trends in X between treated and controls. By replacing $\lambda = E[T]$ with the time score $t(D, X)$, which is the probability of being observed at $T = 1$ conditional on covariates X and treatment status D, it is possible to identify the ATT under compositional changes.[3] Indeed, the DIPW estimand of the ATT can be written as:

$$\delta^{dipw} = E\left[\left(\omega_1^{mod}(D,T) - \omega_0^{mod}(D,T,X;p)\right)Y\right] \tag{9}$$

where:

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

$$\omega_0^{mod}(D,T,X;p,t) = \omega_{0,1}^{mod}(D,T,X;p) - \omega_{0,0}^{mod}(D,T,X;p)$$

and:

$$\begin{aligned}
\omega_{1,1}^{mod}(D,T;t) &= \frac{D\cdot T}{t(D,X)}\bigg/ E\left[\frac{D\cdot T}{t(D,X)}\right]\\
\omega_{1,0}^{mod}(D,T;t) &= \frac{D\cdot(1-T)}{(1-t(D,X))}\bigg/ E\left[\frac{D\cdot(1-T)}{(1-t(D,X))}\right]\\
\omega_{0,1}^{mod}(D,T,X;p,t) &= \frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot t(D,X)}\bigg/ E\left[\frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot t(D,X)}\right]\\
\omega_{0,0}^{mod}(D,T,X;p,t) &= \frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot(1-t(D,X))}\bigg/ E\left[\frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot(1-t(D,X))}\right]
\end{aligned}$$

where $\omega_{d,t}^{mod}(D,T,X;p,t)$ are the standardized Hayek weights after some rearrangements. Note that we used the $\omega_{d,t}^{mod}$ notation to allow for easy comparison to the weights in Sant'Anna and Zhao (2020). The terms $t(D,X)$ and $p(X)$ must be estimated in the sample, usually by logit or probit regression. If the models are correctly specified, the estimator retrieves the ATT.

2.3.2 Doubly Robust Inverse Probability Weighting (DR-DIPW)

Similarly to Sant'Anna and Zhao (2020), we combine the OR and the DIPW approaches into a doubly robust estimand for the ATT. As in Section 2.2.4, denote $\mu_{d,t}(X)$ as the

[3] For proofs and details, see Appendix A.3.

arbitrary model for the true, unknown CEF $m_{d,t}(x) \equiv E[Y|D = d, T = t, X = x]$, and $\mu_{d,Y}(T, X) \equiv T \cdot \mu_{d,1}(X) + (1 - T) \cdot \mu_{d,0}(X)$, where $d, t \in \{0, 1\}$. Then the new estimand is defined as:

$$\delta_1^{drdipw} = E\left[\left(\omega_1^{mod}(D,T) - \omega_0^{mod}(D,T,X;p)\right)\left(Y - \mu_{0,Y}(T,X)\right)\right] \tag{10}$$

where:

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

$$\omega_0^{mod}(D,T,X;p,t) = \omega_{0,1}^{mod}(D,T,X;p) - \omega_{0,0}^{mod}(D,T,X;p)$$

and:

$$\begin{aligned}
\omega_{1,1}^{mod}(D,T) &= \frac{D \cdot T}{t(D,X)} \Big/ E\left[\frac{D \cdot T}{t(D,X)}\right] \\
\omega_{1,0}^{mod}(D,T) &= \frac{D \cdot (1-T)}{(1-t(D,X))} \Big/ E\left[\frac{D \cdot (1-T)}{(1-t(D,X))}\right] \\
\omega_{0,1}^{mod}(D,T,X;p,t) &= \frac{(1-D) \cdot T \cdot p(X)}{(1-p(X)) \cdot t(D,X)} \Big/ E\left[\frac{(1-D) \cdot T \cdot p(X)}{(1-p(X)) \cdot t(D,X)}\right] \\
\omega_{0,0}^{mod}(D,T,X;p,t) &= \frac{(1-D) \cdot T \cdot p(X)}{(1-p(X)) \cdot (1-t(D,X))} \Big/ E\left[\frac{(1-D) \cdot T \cdot p(X)}{(1-p(X)) \cdot (1-t(D,X))}\right]
\end{aligned}$$

The first term in δ_1^{drdipw} represents the DIPW weighting scheme, while the second is the outcome regression part of the estimand. The latter can be adjusted for compositional changes by noting that the locally semi-parametrically efficient version in Sant'Anna and Zhao (2020) already specifies an outcome model for all four $m_{d,t}(x)$. Therefore we proposed its adapted version:

$$\begin{aligned}
\delta_2^{drdipw} &= \delta_1^{drdipw} + (E[\mu_{1,1}(X) - \mu_{0,1}(X)|D = 1] - E[\mu_{1,1}(X) - \mu_{0,1}(X)|D = 1, T = 1]) \\
&\quad - (E[\mu_{1,0}(X) - \mu_{0,0}(X)|D = 1] - E[\mu_{1,0}(X) - \mu_{0,0}(X)|D = 1, T = 0])
\end{aligned} \tag{11}$$

The relative sample analog is obtained by replacing $p(X)$, $t(D, X)$, and $\mu_{d,t}(X)$ with their in-sample estimations. In the Monte Carlo simulations in Section 3, we consider only the estimand δ_2^{drdipw}. To maintain consistency and comparability with the DRDiD and IMP DRDiD estimators, we name, respectively, DR-DIPW the version that employs linear and logistic regression for the outcome equation and the propensity scores, and Improved DR-DIPW (IMP DR-DIPW) the version using the inverse probability tilting estimator, as in Graham et al. (2012), for the treatment and time scores models and weighted least-squares for the outcome model. The two estimators will generally be consistent if either the propensity score or the outcome model is correctly specified.

2.3.3 Machine Learning DR-DIPW

We propose two alternative versions of the DR-DIPW estimator that employ lasso and random forest algorithms for first-stage estimates, named LASSO DR-DIPW and RF DR-DIPW respectively. The advantage of machine learning over traditional estimation methods is that they do not need a parametric assumption of the functional form of the model under study, avoiding the risk of misspecification. Chernozhukov et al. (2018) and the related literature that followed identified three main conditions that enable the use of first-stage machine learning without creating bias in the estimates of the causal parameter.

The first is the so-called Neyman orthogonality condition which guarantees that the estimand must be insensitive to small perturbations of the nuisance functions. This property is typically satisfied for estimands that are doubly robust (Chernozhukov et al., 2018; Sant'Anna and Zhao, 2020; Farrell et al., 2021), like the DR-DIPW estimator in Eq. (10) and (11). Under this condition, machine learning estimates of the nuisance functions (in our case the propensity scores and outcome model functions) are allowed even if they are generally biased due to regularization. Indeed, using a Neyman-orthogonal score eliminates the biases arising from the first-stage estimates.

The second condition refers to the rate of convergence of the machine learning estimators used for the nuisance parameters. In particular, they have to converge to the true parameter using the $L^2(P)$ norm at a rate faster than $o(N^{-1/4})$. 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 using 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. We therefore adapt the DML1 cross-fitting algorithm in Bach et al. (2021) to our lasso and random forest DR-DIPW estimators. In our simulation setting, we find that applying cross-fitting to our lasso version does not improve in-sample bias, while some benefits are found in the case of random forest.[4] Therefore, in Section 3 we apply sample splitting just to the random forest version of the DR-DIPW estimator.

3 Monte Carlo Simulations

In this section, we conduct a series of Monte Carlo simulations in order to investigate the finite sample properties of the proposed estimators in a repeated cross-sections setup. The different methodologies are tested across two different experimental settings. Each design is characterized by two repeated cross-sections, one observed at $T = 0$ and another at $T = 1$, with a total sample size of $n = 1000$ observations. The Monte Carlo simulation consists of 10000 randomly generated datasets and estimation results are stored at each repetition.

Our two simulations focus on cases in which the treatment is not randomized, since in that instance all methods presented in the paper yield unbiased estimates of the ATT. Conversely, Experiment 1 allows instead for non-randomized selection into treatment and X-specific trends, while still assuming time-invariant covariates and homogeneous treatment effects. In Experiment 2, though, the distribution of covariates is allowed to vary between the pre- and post-treatment periods and the treatment effects are heterogeneous

[4]For details on the simulation and cross-fitting algorithm, see Appendix A.4

in X. For this reason, Experiment 2 reproduces the most realistic and indicative version of the data generating process (DGP, henceforth). The choice of the functional forms of our DGPs, presented below, is aimed at preserving the comparability with the work of Sant'Anna and Zhao (2020) and Kang and Schafer (2007), which employed the same functional specifications. Indeed, Experiment 1 closely reproduces the Monte Carlo simulations in Sant'Anna and Zhao (2020), while Experiment 2 extends the study into more realistic conditions.

First of all, for a generic variable $W = (W_1, W_2, W_3, W_4)'$, we define the underlying true outcome and propensity score model as:

$$f_{reg}(W) = 210 + 25.4 \cdot W_1 + 13.7 \cdot (W_2 + W_3 + W_4) \tag{12}$$

$$f_{ps}(W) = 0.75 \cdot (-W_1 + 0.5 \cdot W_2 - 0.25 \cdot -0.1 \cdot W_4) \tag{13}$$

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

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

For each of our two experiments, we replicate both scenarios when the researcher correctly specifies or mispecifies the parametric form of the model. The aim is to assess the estimators' performance in terms of bias when the researcher cannot correctly specify the functional form of the model under study. Overall, each of the two experiments considers four different DGPs.

We define the generic vector W, which can either represent vector Z, which is the set of variables observed by the researcher, or vector X, which is not observable. The idea is that the unique generic DGP (expressed in terms of W) leads, for each experiment, to four cases depending on whether W is replaced by the observed vector Z or by the unobservable vector X. The misspecfication of the models derive from the fact that Z is a highly non-linear transformation of X and its interactions. When the modelling functions are defined as $f_{ps}(Z)$ and $f_{reg}(Z)$, i.e. they are functions of the observed Z, then both the propensity score and outcome regression models will be correctly specified since the variables we observe coincide with those affecting the outcome. We call this scenario DGP A. When the data are generated by $f_{ps}(X)$ and $f_{reg}(X)$, then the researcher, who has only access to Z, will misspecify both models. We call this scenario DGP D. Typically, such a scenario is the most realistic since researchers do not have an *a priori* knowledge of the phenomenon under analysis. We also consider the two cases in which just one of the two models is correctly specified. We call this scenarios DGP B (when the outcome model is correctly specified) and DGP C (when only the propensity score model is correctly specified) and we include the relative tables in the Appendix.

More formally, assume $X = (X_1, X_2, X_3, X_4)'$ is distributed as $N(0, I_4)$ with I_4 representing the 4×4 identity matrix. For $j = 1, 2, 3, 4$ define the standardized variable $Z_j = (\tilde{Z}_j - E[\tilde{Z}_j])/\sqrt{Var(\tilde{Z}_j)}$ where $\tilde{Z}_1 = \exp(0.5X_1)$, $\tilde{Z}_2 = 10 + X_2/(1 + \exp(X_1))$, $\tilde{Z}_3 = (0.6 + X_1X_2/25)^3$, and $\tilde{Z}_4 = (20 + X_2 + X_4)^2$. Note that the non-linear transformations that generate the relationship between the individual variables of Z and X include a wide range of functional forms and interaction terms. As a result, when the propensity score and outcome regression models are misspecified, i.e. when the DGPs are built from X whereas we observe Z only, this is likely to cause bias using conventional first-stage methods that do not capture these non-linearities. On the contrary, using non-parametric methodologies, such as lasso or random forest, may better capture the non-linearities between Z and X and minimize the bias.

All the different estimation methods are evaluated in terms of average bias, root mean square error (RMSE), variance, and computational time required for the estimation.[5] When not otherwise specified, all estimators employ a logit model for the propensity score and a linear regression model for the outcome. Note that, by construction, when the DGPs are built from $f_{ps}(Z)$ and $f_{reg}(Z)$, the estimated models match those of the true DGP.

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

Experiment 1 closely replicates the simulation presented by Sant'Anna and Zhao (2020). Indeed, selection into treatment is not randomized, there are X-specific trends, but at the same time there are homogeneous treatment effects in X and the covariates are assumed to be time-invariant, ruling out compositional changes.

The four different DGPs are specified as in Table 1, where $\epsilon_0(d)$, $\epsilon_1(d)$, $d = 0, 1$, are independent standard normal random variables representing the stochastic error term of the potential outcomes, the propensity score $p(W)$ is a logistic transformation of the generic function $f_{ps}(W)$, λ is the proportion of observations in $T = 1$ and U_d and U_t are independent standard uniform stochastic variables used to randomly select individuals into treatment and post-treatment period, respectively. For a generic variable W, $\upsilon(W, D)$ is an independent normal random variable with mean $D \cdot f_{reg}(W)$ and unit variance which represents the time-invariant unobserved group heterogeneity between treated and untreated populations. The trend is specified as $\tau(W) = f_{reg}(W)$, and therefore in the

[5]For a summary list of all estimators included in the analysis, see Table A.1 in the Appendix.

post-treatment period $T = 1$ it sums to the standard function of the outcome model f_{reg}. This explains the presence of the factor 2 that multiplies the term $f_{reg}(W)$ in the formula of the potential outcome $Y_{1,1}$. The observed outcome is $Y = DTY_{1,1} + D(1-T)Y_{1,0} + (1-D)TY_{0,1} + (1-D)(1-T)Y_{0,0}$. In the aforementioned DGPs, the true ATT is zero. Overall, the DGP is built so that for each treatment group category there are no changes in the distribution of X between $T = 0$ and $T = 1$, but there is heterogeneity in the distribution of the covariates among treated and controls, as showed in Figure A.1. The results of the Experiment 1A and 1D are displayed in Tables 2 to 3, while Experiments 1B and 1C are shown in the Appendix (see Table A.2 and Table A.3).

In Experiment 1, the bias of the TWFE estimator with covariates is evident. Independently of whether the model is correctly specified, TWFE is typically severely biased, with a bias of 20.762 even in the most favorable scenario embodied by DGP A. The TWFE correction is characterized by a significantly lower bias: in Experiments 1A and 1B, where the outcome model is correctly specified, it is approximately unbiased, while in Experiments 1C and 1D, it is outperformed by most of the other estimators. The most efficient class of estimators is represented by the doubly robust methods, which are approximately unbiased when either the propensity score or the outcome model are correctly specified. They also have better in-sample properties when both models are misspecified. In the latter case indeed, IMP DRDiD and IMP DR-DIPW have approximately half of the bias (2.550 and 2.563 respectively) compared to the TWFE correction (5.108). In Experiment 1D, the lasso version of the DR-DIPW has the lowest bias (1.894), thanks to its more flexible parametric assumptions. Conversely, the other machine learning method DMLDiD is severely biased in all four scenarios and is characterized by a very high variance in its estimates, even when IPW and DIPW have a low bias, not far from the best performing-estimators.

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

Experiment 2 tests the proposed estimators in presence of X-specific trends, non-randomized selection into treatment, heterogeneous treatment effects in X, and compositional changes in the distribution of X between the pre and post-treatment periods. As discussed in Section 2.3, the inclusion of time-varying covariates in the TWFE is likely to yield biased estimates, and the other aforementioned alternative semi-parametric estimators may perform poorly as well since they assume time-invariant X.

Table 4 describes the four DGPs in Experiment 2. In this design, the DGPs are subject to two main changes. The first is that we model differently the time trends in the covariates between treated and controls by specifying a distinct function $t(D, X)$ for the two groups. In particular, the probability is calculated as the logistic transformation of $f_{ps}(-W)$ for treated and $f_{ps}(W)$ for controls. This way, the observations are assigned to a specific time period so that different evolutions of the distribution of the covariates among treated and controls are produced (see Figure A.2). The second change consists in allowing the treatment effect to vary with X. This is achieved by denoting the treatment effect as $\tilde{\delta}(W) = -10W_1 + 10W_2 - 10W_3 - 10W_4$. In addition, to guarantee that the ATT is zero as in the two previous experiments, we use the demeaned transformation of $\overline{\delta}(W)$, e.g. $\delta(W) = \tilde{\delta}(W) - E_{i|D=1}[\tilde{\delta}(W)]$, where $E_{i|D=1}[\tilde{\delta}(W)]$ denotes the ATT before demeaning. The results of the Experiment 2A and 2D are displayed in Tables 5 to 6, while Experiments 2B and 2C are shown in the Appendix (see Table A.4 and Table A.5).

In most DGPs of Experiment 2, the traditional TWFE specification is severely biased. An exception is Experiment 2D, where TWFE is characterized by a relatively low bias (3.437), but this is likely caused by different sources of bias offsetting each other since in all other scenarios standard regression works poorly. However, the proposed TWFE correction substantially improves the estimates. Indeed, in Experiments 2A and 2B, TWFE CORR is approximately unbiased as the doubly-robust estimators.

In the same two experiments, also DRDiD and IMP DR-DiD are approximately unbiased, even if they are originally built for time-invariant covariates. This indicates that their flexible specification of the outcome model can be naturally extended to time-varying covariates under our assumptions. This is not always the case, since the OR approach and the DRDiD versions that are not locally-efficient (not reported, available upon request) are substantially biased also in the case of correctly specified models.

However, in all four simulations, the traditional weighting scheme of the IPW, which is also embedded in Sant'Anna and Zhao (2020), is markedly biased. On the contrary, the DIPW correctly accounts for compositional changes and shows very limited bias when the treatment and time scores models are correctly specified. Indeed, it is characterized by one of the best in-sample performances in terms of bias also in Experiment 2D.

Overall, the doubly robust versions of the DIPW are the models with better properties in all four simulations. They are approximately unbiased in all settings where at least either the propensity scores or the outcome models are correctly specified. In particular, they strictly outperform their estimators in Sant'Anna and Zhao (2020) that do not account for compositional effects in Experiment 2C and 2D, where the weighting scheme plays a more important role. In particular, DR-DIPW and DRDiD have a bias of 0.277 and 4.379, respectively, in Experiment 2C, while IMP DR-DIPW and IMP DRDiD have 0.519 and 1.092. Similarly, in Experiment 2D the bias for DR-DIPW and DRDiD is 12.793 and 17.736, respectively, and for IMP DR-DIPW and IMP DRDiD is 10.429 and 12.793. Noticeably, the RF DR-DIPW and LASSO DR-DIPW have even lower bias (0.831 and 6.631 respectively) in Experiment 2D, which replicates the most realistic scenario where the researcher cannot know the functional form of the phenomenon under study. For this reason, we suggest using this version in empirical studies.

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

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

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

$$\begin{aligned} y_{it} = \gamma_1 (TariffChangeCategory_i \times POST) + \mu POST \\ + \beta_1 TariffChangeCategory_i + \beta_2 BaselineTariff_i \\ + \Gamma_i + p_i + + \omega_i + \delta_i + \epsilon_{it} \end{aligned} \tag{14}$$

where y_{it} represents the natural log of the amount of bribe paid for shipment i in period t, conditional on paying a bribe, $TariffChangeCategory_i \in \{0,1\}$ takes value one if the commodity was subject to tariff reduction, $POST \in \{0,1\}$ denotes the years following 2008, and $BaselineTariff_i$ is a control for the pre-treatment tariff for product i. The specification also accounts for a vector of product, shipment, clearing agent, and firm-level characteristics Γ_i which includes the elements summarised in Table A.6. Industry, year, and clearing agent fixed effects are included, denoted by p_i, ω_t, and δ_i respectively. The parameter of interest is the coefficient of the interaction between the time and treatment dummies, namely γ_1.

The main finding of Sequeira (2016) is that the tariff reduction led to a significant drop in the amount of bribe paid. Chang (2020) replicates the estimation by using the DMLDiD estimator. In Table 7, we report the results obtained by the two authors, where TWFE refers to the standard specification in Sequeira (2016) (Equation 1 of Table 9 in their paper), TWFE ($\Gamma_i \times$ POST) is the specification that also includes the interactions between the covariates Γ_i and $POST$ (which differs from Eq. (2) where all the interactions between the covariates and the time and treatment group dummies are added), while DMLDiD is estimated by either using kernel or lasso in the first-stage estimates. Overall, the DMLDiD estimates claim that the effect of the reduction was larger than originally thought.

However, both classes of estimators used in this analysis are likely to be characterized by a substantial degree of bias. Figure A.3 shows the standardized mean difference in trend for each of the 33 covariates between treated and controls. The latter is defined as $\frac{(\overline{X}_{11}-\overline{X}_{10})-(\overline{X}_{01}-\overline{X}_{00})}{std(X)}$, where the overbar indicates the mean. The graph suggests that the distribution of the covariates is time-varying because, in the case of time-invariant X, the metrics should be 0, and the presence of heterogeneous covariate trends between treated and controls. This is the condition tested in Experiment 2 in Section 3, where we assume a non-randomized treatment scenario with X-specific trends, compositional changes, heterogeneous effects, and potential non-linearities in the DGP. Because in our simulation the two estimators were strongly biased, computing the effect of tariff reduction on bribing patterns with TWFE and DMLDiD may be misleading. In addition, DMLDiD estimates in Table 7 suffer from very high standard errors which blur the interpretation of the empirical findings. For example, the 95 percent confidence interval lies approximately between 0.318 and -14.306 for the kernel DMLDiD, and the same applies to its lasso version, even if to a smaller degree.

Motivated by these considerations, we employ the lasso and random forest DR-DIPW estimators, which proved to be the least biased estimator in the most realistic setting of the Monte Carlo simulations (Experiment 2D), to the current study of the effect of tariff reduction on bribing behaviour. The lasso specification captures non-linearities by allowing for a richer set of covariates: Γ_i is expanded to include all second order terms and interactions, leading to a set of 112 controls. Contrarily to Chang (2020), we stick to the specification in Sequeira (2016) by including all industry, time and clearing agent fixed-effects, even if their interactions are not generated due to computational tractability. The DR-DIPW standard errors are clustered at the level of product's four-digit HS code and are computed through weighted bootstrap, similarly to Sant'Anna and Zhao (2020).

Our final estimates are displayed in Table 7. Our results, across different methods and specifications, corroborate the hypothesis that the tariff reduction led to a drop in the amount of bribe paid, but give compelling evidence against the assumption that the effect was higher in magnitude. In fact, the standard TWFE seems to overestimate the ATT (-3.748), while the lasso and random forest DR-DIPW estimates are -2.387 and -2.301 respectively. The standard errors are typically lower than those in Chang (2020), producing more precise confidence intervals estimations.

In summary, our findings reveal that the tariff reduction had a significant effect on bribing behavior, but the impact is smaller than originally estimated by the standard TWFE specification and by DMLDiD as in Chang (2020).

5 Conclusions

Our analysis shows that the commonly-used DiD estimators, including TWFE, may be severely biased when invoking the conditional parallel trend assumption in presence of a distribution of the covariates that varies over time. We specify a set of inverse probability weights (DIPW) that use both a treatment and a time score to retrieve the ATT under compositional changes and, building on Sant'Anna and Zhao (2020), we propose its doubly-robust version (DR-DIPW). When comparing the performance of the various estimators proposed by the literature through a set of Monte Carlo simulations, we show that DR-DIPW is robust to compositional changes and tends to outperform all other available methods. In particular, when the researcher cannot correctly specify the functional form of the outcome and propensity scores models under study, the lasso and random forest versions of DR-DIPW have even relatively less bias than the one employing linear and logistic regressions for first-stage estimates.

We apply both the lasso and random forest DR-DIPW estimators in the empirical setting of Sequeira (2016), which investigates the effect of tariff reduction on corruption behaviors by using bribe payment data on the cargo shipments transiting from South Africa into the ports in Mozambique. Our estimates show that tariff reduction led to a decrease in bribes paid but the effect is significantly lower in magnitude than the one estimated in the original paper using a TWFE specification with covariates and in the replication by Chang (2020) adopting a DMLDiD specification.

References

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

[2] M. G. Allingham and A. 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.

[3] P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. Doubleml–an object-oriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603*, 2021.

[4] C. Caetano, B. Callaway, S. Payne, and H. S. Rodrigues. Difference in differences with time-varying covariates. *arXiv preprint arXiv:2202.02903v2*, 2022.

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

[6] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018. URL https://doi.org/10.1111/ectj.12097.

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

[8] M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[9] J. 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.

[10] R. Fisman and S.-J. Wei. Tax Rates and Tax Evasion: Evidence from "Missing Imports" in China. NBER Working Papers 8551, National Bureau of Economic Research, Inc, Oct. 2001. URL https://ideas.repec.org/p/nbr/nberwo/8551.html.

[11] J. Friedman, T. Hastie, and R. 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/.

[12] B. S. Graham, C. C. De Xavier Pinto, and D. 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.

[13] J. J. Heckman, H. Ichimura, and P. 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.

[14] S.-H. 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.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[16] J. D. Kang and J. 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.

[17] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

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

[19] J. 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.

[20] J. Roth, P. H. Sant'Anna, A. Bilinski, and J. Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*, 2022.

[21] P. H. C. Sant'Anna and Q. Xu. Difference-in-differences with compositional changes. *arXiv preprint arXiv:2304.13925*, 2023.

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

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

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

[25] J. Slemrod and S. Yitzhaki. Tax avoidance, evasion, and administration. 3:1423–1470, 2002.

[26] B. Zeldow and L. A. Hatfield. Confounding and regression adjustment in difference-in-differences. *arXiv preprint arXiv:1911.12185*, 2019.

6 Main Tables

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

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

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

Table 2: Exp.1A Propensity score model correct, outcome regression correct

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	20.762	21.071	12.952	0.002
TWFE CORR	Regression, Eq. (2)	0.002	0.201	0.040	0.002
IPW					
IPW	Abadie (2005)	0.158	9.587	91.892	0.005
DMLDiD	Chang (2020)	34.356	71.469	3,927.540	43.660
DIPW	Author's work	0.002	4.705	22.141	0.009
OR					
OR	Heckman et al. (1997)	0.102	7.585	57.516	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	0.001	0.210	0.044	0.011
DR-DIPW	Author's work	0.001	0.211	0.044	0.015
IMP DRDiD	Sant'Anna and Zhao (2020)	0.001	0.210	0.044	0.015
IMP DR-DIPW	Author's work	0.001	0.210	0.044	0.025
Debiased ML					
LASSO DR-DIPW	Author's work	0.123	0.343	0.102	3.026
RF DR-DIPW	Author's work	4.156	6.037	19.176	2.314

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X, and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute 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.1D Propensity score model incorrect, outcome regression model incorrect

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	16.269	17.067	26.619	0.002
TWFE CORR	Regression, Eq. (2)	5.108	6.919	21.778	0.002
IPW					
IPW	Abadie (2005)	4.037	10.569	95.401	0.005
DMLDiD	Chang (2020)	115.748	145.629	7,810.146	43.449
DIPW	Author's work	3.915	6.940	32.833	0.009
OR					
OR	Heckman et al. (1997)	5.248	10.001	72.473	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	3.166	6.048	26.558	0.011
DR-DIPW	Author's work	3.164	5.960	25.508	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	2.550	4.874	17.258	0.015
IMP DR-DIPW	Author's work	2.563	4.886	17.309	0.025
Debiased ML					
LASSO DR-DIPW	Author's work	1.894	3.953	12.043	3.437
RF DR-DIPW	Author's work	4.665	6.797	24.431	2.318

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X, and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table 4: DPGs in Experiment 2 (PS=propensity score, OR=outcome regression)

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

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

Table 5: 2A Propensity score model correct, outcome regression model correct

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	8.928	9.695	14.276	0.002
TWFE CORR	Regression, Eq. (2)	0.002	0.219	0.048	0.002
IPW					
IPW	Abadie (2005)	45.102	46.124	93.208	0.005
DMLDiD	Chang (2020)	297.085	316.542	$11,939.030$	43.685
DIPW	Author's work	0.481	6.689	44.510	0.010
OR					
OR	Heckman et al. (1997)	26.066	27.144	57.363	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	0.002	0.226	0.051	0.011
DR-DIPW	Author's work	0.001	0.266	0.071	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	0.001	0.237	0.056	0.015
IMP DR-DIPW	Author's work	0.001	0.259	0.067	0.026
Debiased ML					
LASSO DR-DIPW	Author's work	0.241	0.380	0.086	3.208
RF DR-DIPW	Author's work	2.696	7.622	50.827	2.177

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute 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: 2D Propensity score model incorrect, outcome regression model incorrect

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	3.437	6.240	27.123	0.002
TWFE CORR	Regression, Eq. (2)	16.414	17.066	21.834	0.002
IPW					
IPW	Abadie (2005)	54.838	55.659	90.716	0.005
DMLDiD	Chang (2020)	342.449	367.046	$17,451.260$	43.129
DIPW	Author's work	7.553	14.246	145.889	0.009
OR					
OR	Heckman et al. (1997)	44.137	44.880	66.163	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	17.736	18.416	24.575	0.011
DR-DIPW	Author's work	13.740	16.231	74.672	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	12.793	13.643	22.459	0.015
IMP DR-DIPW	Author's work	10.429	11.551	24.664	0.026
Debiased ML					
LASSO DR-DIPW	Author's work	6.631	7.968	19.513	3.555
RF DR-DIPW	Author's work	0.894	7.952	62.438	2.166

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute 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: The effect of tariff reduction on bribes

	TWFE Sequeira (2016)	TWFE ($\Gamma_i \cdot POST$) Sequeira (2016)	DMLDiD (Kernel) Chang (2020)	DMLDiD (lasso) Chang (2020)
ATT	-3.748^{***}	-2.928^{***}	-6.998^{*}	-5.222^{**}
St.Err.	1.075	0.944	3.752	2.647
	RF DR-DIPW	LASSO DR-DIPW		
ATT	-2.301^{**}	-2.387^{**}		
St.Err.	0.935	1.052		

Notes: TWFE and TWFE($\Gamma_i \times POST$) are Equation 1 and 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), where the parenthesis indicates the method utilized for the first-stage estimates. LASSO DR-DIPW and RF DR-DIPW are Eq. (11) with lasso and random forest first stage estimates respectively. The coefficients capture the difference in the log of bribes paid for products that changed tariff level, before and after the tariff change took place. Standard errors are clustered at the level of product's four-digit HS code. Regarding the p-value, *** stands for $p < 0.01$, ** for $p < 0.05$, and * for $p < 0.1$.

Appendix

A.1 TWFE Limitations

In this section, we review the three cases specified in Section 2.2.1 where TWFE with covariates may deliver biased estimates of the ATT.

Case (i). (X-specific trends) Even when the covariates X are time-invariant, their coefficients to the outcome may vary over time. We can write the expected value of the time-invariant covariate as $\overline{X}_{1,1} = \overline{X}_{1,0} \equiv \overline{X}_1$ and $\overline{X}_{0,1} = \overline{X}_{0,0} \equiv \overline{X}_0$, where $\overline{X}_{d,t}$ is the expected value among the group $D = d$ and $T = t$. Consider for simplicity just one covariate and denote θ_t the time-varying coefficient of X, then TWFE without controls implicitly assumes:

$$E(Y_{0,0}|D=0) = \alpha_0 + \theta_0 \overline{X}_0$$

$$E(Y_{0,1}|D=0) = \alpha + \gamma + \theta_1 \overline{X}_0$$

$$E(Y_{0,0}|D=1) = \alpha + \beta + \theta_0 \overline{X}_1$$

$$E(Y_{0,1}|D=1) = \alpha + \gamma + \beta + \theta_1 \overline{X}_1$$

Assuming that the parallel trend assumption holds, we can write:

$$E(Y_{0,1} - Y_{0,0}|D=1) = E(Y_{0,1} - Y_{0,0}|D=0)$$

$$\alpha + \gamma + \beta + \theta_1 \overline{X}_1 - (\alpha + \beta + \theta_0 \overline{X}_1) = \alpha + \gamma + \theta_1 \overline{X}_0 - (\alpha + \theta_0 \overline{X}_0)$$

$$(\theta_1 - \theta_0) \cdot (\overline{X}_1 - \overline{X}_0) = 0 \tag{15}$$

where the last line rearranges the terms. This implies that for covariates that do not vary over time, TWFE without covariates identifies the ATT if either: (i) the means of the covariates are the same across groups or (ii) the effects of the covariates on the outcome variable are equal in the pre and post-treatment periods (Zeldow and Hatfield, 2019). From this derivation, it is possible to see that adding the covariate in the TWFE to claim the conditional parallel trend assumption imposes the condition of a constant estimated

$\hat{\theta}$ while it is perfectly possible that $\theta_1 \neq \theta_0$, i.e. the effect of the covariate to the outcome varies over time.

When allowing time-varying covariates in the TWFE regression, Eq. (15) can be rewritten:

$$E(Y_{0,1} - Y_{0,0}|X, D = 1) = E(Y_{0,1} - Y_{0,0}|X, D = 0)$$
$$\theta_1(\overline{X}_{1,1} - \overline{X}_{0,1}) - \theta_0(\overline{X}_{1,0} - \overline{X}_{0,0}) = 0$$

Therefore, in the case of time-varying covariates, TWFE without covariates retrieves the ATT only if (i) the relationship between the covariates and the outcome is constant in time, and (ii) the difference in the mean of the covariates between treated and controls is the same in pre and post-treatment periods (i.e. $\overline{X}_{1,1} - \overline{X}_{0,1} = \overline{X}_{1,0} - \overline{X}_{0,0}$) (Zeldow and Hatfield, 2019). Also in this instance, just adding the covariate as a control is ineffective since it imposes the estimated $\hat{\theta}$ to be constant over time.

Case (ii). (Heterogeneous effects) In most realistic settings, the effect of the treatment is likely to vary for different values of the covariates X. However, TWFE implicitly assume homogeneous treatment effects in X (Meyer, 1995; Abadie, 2005; Sant'Anna and Zhao, 2020; Roth et al., 2022). For instance, let the treatment effect be heterogeneous in X, as in Cunningham (2021), namely redefining the potential outcomes for the treated in the post period as $E(Y_{1,1}|X, D = 1) = \alpha + \gamma + \beta + (\delta + \rho X) + \theta X$. Then, even assuming time-invariant coefficients of the covariates $\theta_1 = \theta_0$ we have:

$$E(Y_{1,1} - Y_{1,0}|X, D = 1) = \delta + [(\theta + \rho)X - \theta X]$$
$$= \delta + \rho X$$

while the TWFE estimate of the ATT is just δ. As a consequence, whenever $\rho \neq 0$ and thus the treatment is heterogeneous in X, the regression estimate does not identify the true ATT, even when the covariates are restricted to be time-invariant.

Case (iii). (Non-additive linear functional form) Since in most settings it is not possible to use a fully saturated model in X, TWFE assumes a Conditional Expectation Function (CEF) that is linear in X and which may not be true in the data. For example, if the vector X does not affect the potential outcome linearly, then the potential outcome is:

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

and the TWFE estimate is biased since the estimated conditional expectations do not capture non-linearities.

A.2 Inverse Probability Weighting

Consider the following weighting scheme:

$$\begin{aligned} E\left[E\left[Y_{1,1}|D=1,T=1,X\right]\right] &= E\left[\frac{YDT}{\lambda p(X)}\right] \\ E\left[E\left[Y_{1,0}|D=1,T=0,X\right]\right] &= E\left[\frac{YD(1-T)}{(1-\lambda)p(X)}\right] \\ E\left[E\left[Y_{0,1}|D=0,T=1,X\right]\right] &= E\left[\frac{Y(1-D)T}{\lambda(1-p(X))}\right] \\ E\left[E\left[Y_{0,0}|D=0,T=0,X\right]\right] &= E\left[\frac{Y(1-D)(1-T)}{(1-\lambda)(1-p(X))}\right] \end{aligned}$$

where $\lambda = E[T]$, i.e. the proportion of individuals observed in $T = 1$. The first equation is obtained by the following passages:

$$\begin{aligned} E\left[\frac{YDT}{\lambda p(X)}\right] &= E\left[E\left[\frac{YDT}{\lambda p(X)}|X\right]\right] \\ &= E\left[E\left[\frac{YT}{\lambda p(X)}|X\right]p(D=1|X)\right] \\ &= E\left[E\left[\frac{Y}{\lambda p(X)}|D=1,T=1,X\right]\lambda p(D=1|X)\right] \\ &= E\left[E\left[Y|D=1,T=1,X\right]\right] \\ &= E\left[E\left[Y_{1,1}|D=1,T=1,X\right]\right] \end{aligned}$$

where in the first line we exploited the law of iterated expectations, in the second and third lines the definition of an expectation function for a dummy variable, in the fourth the definition of the treatment score, and the fifth uses the potential outcome notation. A similar reasoning applies to the other conditional expectations. By taking the difference in differences of the potential outcomes, we obtain:

$$E\big[E\left[Y_{1,1}|D=1,T=1,X\right]-E\left[Y_{1,0}|D=1,T=0,X\right]-$$

$$E\left[Y_{0,1}|D=0,T=1,X\right]+E\left[Y_{0,0}|D=0,T=0,X\right]\big]$$

$$=E\left[Y\cdot\frac{(T-\lambda)}{\lambda(1-\lambda)}\frac{(D-p(X))}{(1-p(X))p(X)}\right]=E\left[Y\omega\right]$$

where we defined the general weighting scheme as $\omega\equiv\frac{(T-\lambda)}{\lambda(1-\lambda)}\frac{(D-p(X))}{(1-p(X))p(X)}$. By using the conditional parallel trend and readjusting the weights to $\omega_{att}=\omega\frac{p(X)}{E(D)}$ to account for the distribution of X in $D=1$ (see appendix A.3), we can retrieve the $ATT=E(Y_{1,1}-Y_{0,1}|D=1)$ by the following estimand:

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

which proves Eq. (5) of the main text.

A.3 Double inverse-probability weighting (DIPW)

Consider the following weighting scheme:

$$E\left[E\left[Y_{1,1}|D=1,T=1,X\right]\right]=E\left[\frac{YDT}{t(D,X)p(X)}\right]$$

$$E\left[E\left[Y_{1,0}|D=1,T=0,X\right]\right]=E\left[\frac{YD(1-T)}{(1-t(D,X))p(X)}\right]$$

$$E\left[E\left[Y_{0,1}|D=0,T=1,X\right]\right]=E\left[\frac{Y(1-D)T}{t(D,X)(1-p(X))}\right]$$

$$E\left[E\left[Y_{0,0}|D=0,T=0,X\right]\right]=E\left[\frac{Y(1-D)(1-T)}{(1-t(D,X))(1-p(X))}\right]$$

Indeed, the following passages show that the equivalence for the first equation:

$$
\begin{aligned}
E\left[\frac{YDT}{t(D,X)p(X)}\right] &= E\left[E\left[\frac{YDT}{t(D,X)p(X)}|X\right]\right] \\
&= E\left[E\left[\frac{YT}{t(D,X)p(X)}|D=1,X\right]p(D=1|X)\right] \\
&= E\left[E\left[\frac{Y}{t(D,X)p(X)}|D=1,T=1,X\right]p(T=1|D=1,X)p(D=1|X)\right] \\
&= E\left[E\left[\frac{Y}{t(D,X)p(X)}|D=1,T=1,X\right]t(D,X)p(X)\right] \\
&= E\left[E\left[Y|D=1,T=1,X\right]\right] \\
&= E\left[E\left[Y_{1,1}|D=1,T=1,X\right]\right]
\end{aligned}
$$

where in the first line we exploited the law of iterated expectations, in the second and third lines the definition of an expectation function for a dummy variable, in the fourth the definition of the treatment and time scores, the fifth line simplifies and last uses the potential outcome notation. Similar passages applies for the other conditional expectations. By taking the difference in differences of these conditional expectations we define the general weighting scheme ω^{mod}:

$$
\begin{aligned}
&E\big[E\left[Y_{1,1}|D=1,T=1,X\right] - E\left[Y_{1,0}|D=1,T=0,X\right] - \\
&\quad E\left[Y_{0,1}|D=0,T=1,X\right] + E\left[Y_{0,0}|D=0,T=0,X\right]\big] = \\
&= E\left[Y \cdot \frac{(T-t(D,X))}{t(D,X)(1-t(D,X))}\frac{(D-p(X))}{(1-p(X))p(X)}\right] \\
&= E\left[Y\omega^{mod}\right]
\end{aligned}
$$

where we defined $\omega^{mod} = \frac{(T-t(D,X))}{t(D,X)(1-t(D,X))}\frac{(D-p(X))}{(1-p(X))p(X)}$ using a similar notation as in Abadie (2005). Since we are mainly focused on the ATT, the weights must be adjusted to take account of the distribution of X in $D = 1$. This can be retrieved by:

$$\begin{aligned} ATT &= E(Y_{1,1} - Y_{0,1}|D = 1) \\ &= \int E\left[Y_{1,1} - Y_{0,1}|D = 1, X\right] dp(X|D = 1) \\ &= \int E\left[Y\omega^{mod}|X\right] dp(X|D = 1) \\ &= E\left[Y\omega^{mod}\frac{P(D = 1|X)}{P(D = 1)}\right] \\ &= E\left[Y\omega^{mod}\frac{p(X)}{E(D)}\right] \\ &= E\left[Y\omega^{mod}_{att}\right] \end{aligned}$$

where we defined $\omega^{mod}_{att} = \omega^{mod}\frac{p(X)}{E(D)}$ and in the first line we used the definition of the ATT, in the second the law of iterated expectations, in the third the conditional parallel trend assumption, in the fourth the Bayes law for the conditional probability, and the last line rearranges. The DIPW weights $\omega^{mod}_{d,t}(D, T, X; p, t)$ are finally obtained from ω^{mod}_{att} using standardization to ensure they sum up to 1. These standardized weights are of the Hayek type and they reduce the variance of the estimator by eliminating extreme weights at the bounds of $[0, 1]$. Therefore the DIPW estimand of the ATT can be written as:

$$\delta^{dipw} = E\left[\left(\omega^{mod}_1(D, T) - \omega^{mod}_0(D, T, X; p)\right)Y\right]$$

where:

$$\omega^{mod}_1(D, T) = \omega^{mod}_{1,1}(D, T) - \omega^{mod}_{1,0}(D, T)$$

$$\omega^{mod}_0(D, T, X; p, t) = \omega^{mod}_{0,1}(D, T, X; p) - \omega^{mod}_{0,0}(D, T, X; p)$$

and:

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

where $\omega_{d,t}^{mod}(D,T,X;p,t)$ are the standardized Hayek weights after some rearrangements. This proves Eq. (10) of the main text.

A.4 Cross-Fitted Machine Learning DR-DIPW

In this section, we provide additional evidence on whether cross-fitting helps reduce the bias in our proposed machine-learning versions of the DR-DIPW estimator. As explained in Section 2.3.3, sample splitting is one of the conditions often invoked to guarantee consistent estimates of the causal parameter in the debiased machine learning literature. In what follows, we explain in detail the implementation of our machine learning methods, our cross-fitting algorithm, and we finally compare the estimates of the lasso and random forest DR-DIPW estimators with or without cross-fitting by running the same Monte Carlo simulations presented in Section 3.

When implementing LASSO DR-DIPW, the outcome and the treatment model are designed as a penalized linear and a penalized logistic regression, respectively. Lasso is performed in R using the 'glmnet' package (Friedman et al., 2010) and the shrinkage parameter λ is selected through 10-fold cross-validation and represents the largest value of λ whose cross-validation error is within 1 standard deviation from the minimum. This allows to select the sparsest model with performances approximately equal to the optimum. In the simulations we allow lasso to employ an expanded set of covariates that include all third-order terms and interactions of the original variables and to perform variable selection.

When RF DR-DIPW is used, the estimation is implemented using the 'randomForest' R package (Liaw and Wiener, 2002). The number of trees is set to 500 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).

For the implementation of the cross-fitting technique, we follow the DML1 algorithm as in Bach et al. (2021). Following their notation, define the data as $(W_i)_{i=1}^{N}$, where N=1000 is the sample size in our simulations. The DR-DIPW in Eq. (11) is a doubly-robust estimand which is a Neyman-orthogonal score function in the form of $\psi(W;\delta,\eta)$, where δ is the parameter of interest, in our case the ATT, and the nuisance parameter $\eta = (\mu_{d,t}, p, t)$, namely the outcome models and the treatment and time scores. The population values are defined as δ_0 and η_0 respectively. The cross-fitting algorithm we employed can be divided in 3 steps:

1. Train machine learning estimators: take a 2-fold random partition $(I_k)_{k=1}^{2}$ of observation indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/2$. For each $k \in [K] = \{1, 2\}$, construct machine learning estimators

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i \notin I_k}\right)$$

of η_0, where $x \mapsto \hat{\eta}_{0,k}(x)$ depends only on the subset of data $(W_i)_{i \notin I_k}$.

2. Compute ATT in each fold: for each $k \in [K]$, compute the estimator $\check{\delta}_{0,k}$ as the solution to the equation

$$\frac{1}{n}\sum_{i \in I_k} \psi\left(W_i; \check{\delta}_{0,k}, \hat{\eta}_{0,k}\right) = 0.$$

3. Take the average among folds: the final estimate of the ATT is obtained by aggregating the estimates in each fold

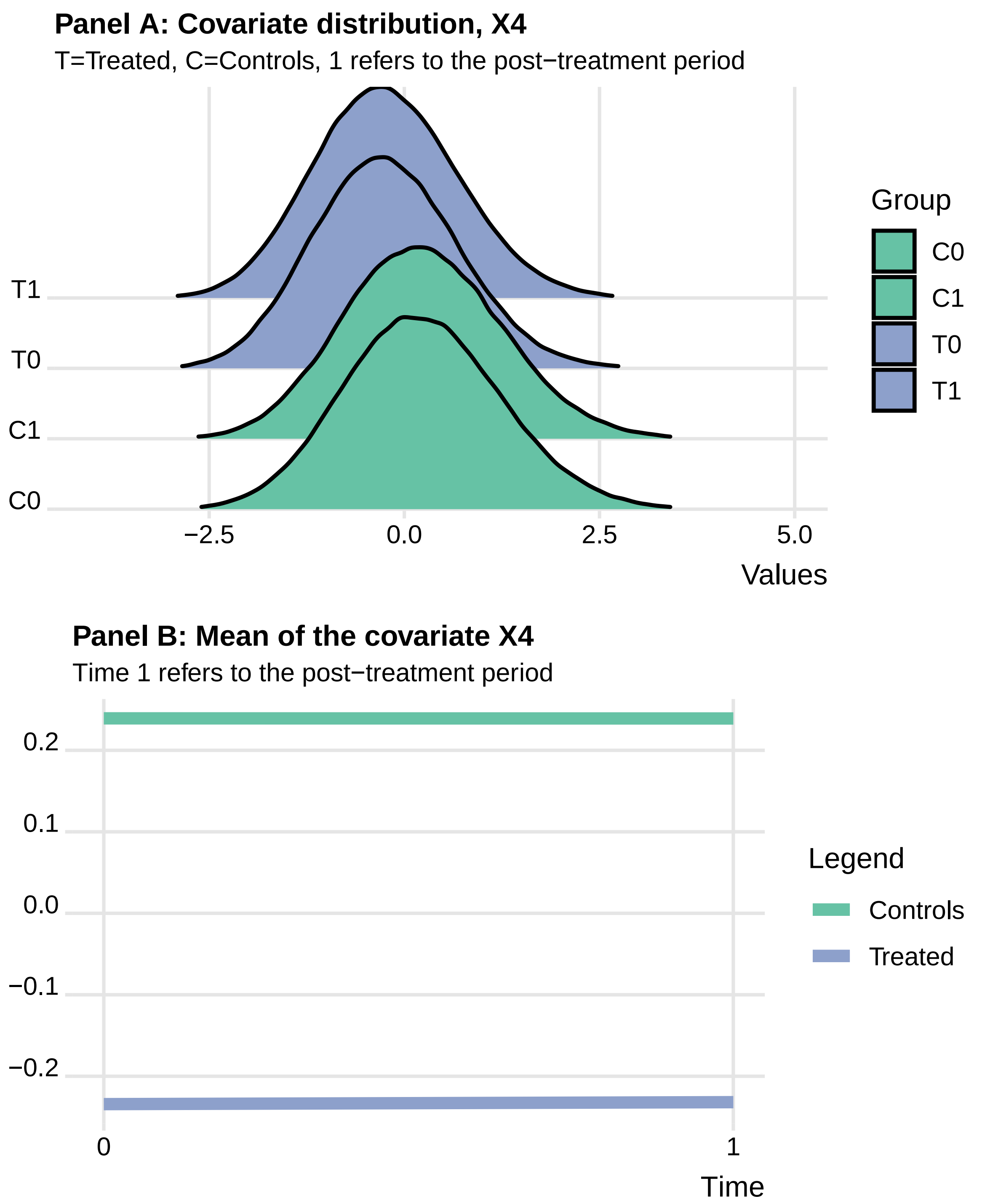
$$\tilde{\delta}_0 = \frac{1}{2}\sum_{k=1}^{2}\check{\delta}_{0,k}.$$

where $\tilde{\delta}_0$ is the resulting estimate of the ATT.

The estimators are then evaluated in Experiments 1 and 2 presented in Section 3, leaving all the conditions of the Monte Carlo simulations identical to the ones explained in the main paper. The results of the simulations are presented in Table A.7 and Table A.8. In our setting, the cross-fitted version of the DR-DIPW with lasso first stage estimates (LASSO DR-DIPW SPLIT) is outperformed by its version that does not perform sample split, especially in Experiment 2C and 2D. Contrarily, the cross-fitted estimator RF DR-DIPW SPLIT tends to have a significant lower bias in most of the Experiment 2 DGPs, even if these gains in performance are not present in the simulations of Experiment 1. Since sample splitting in general allows for less restrictive assumptions, we use the cross-fitted version of the random forest DR-DIPW estimator in Section 3.

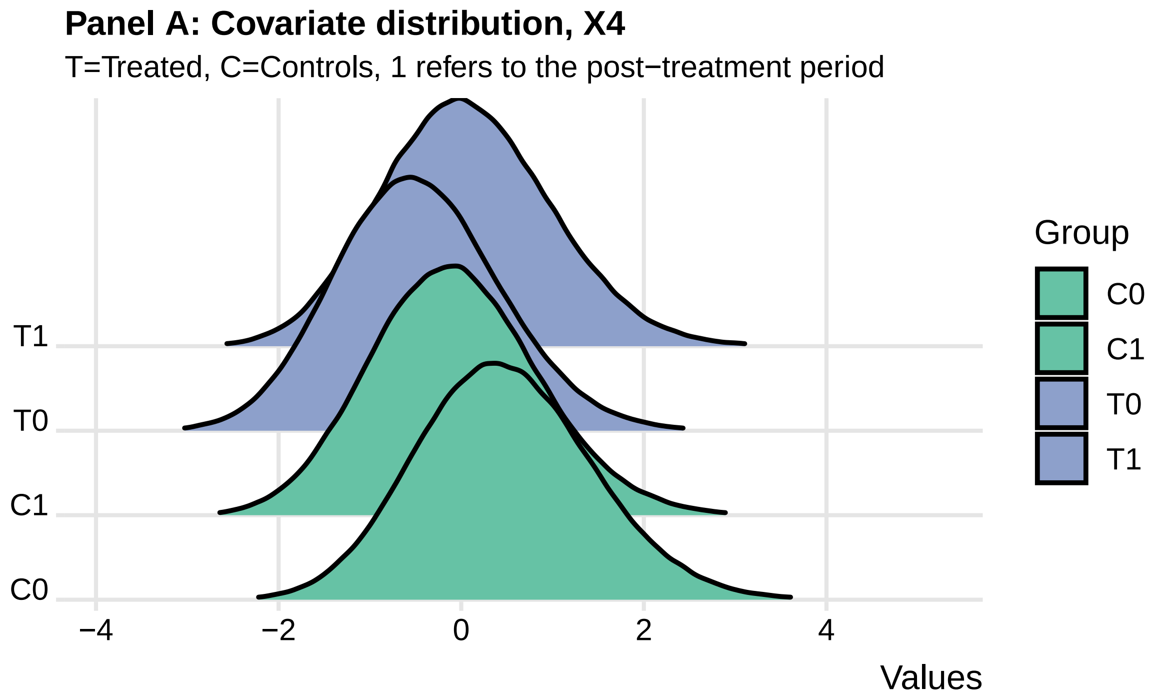
A.5 Appendix Figures

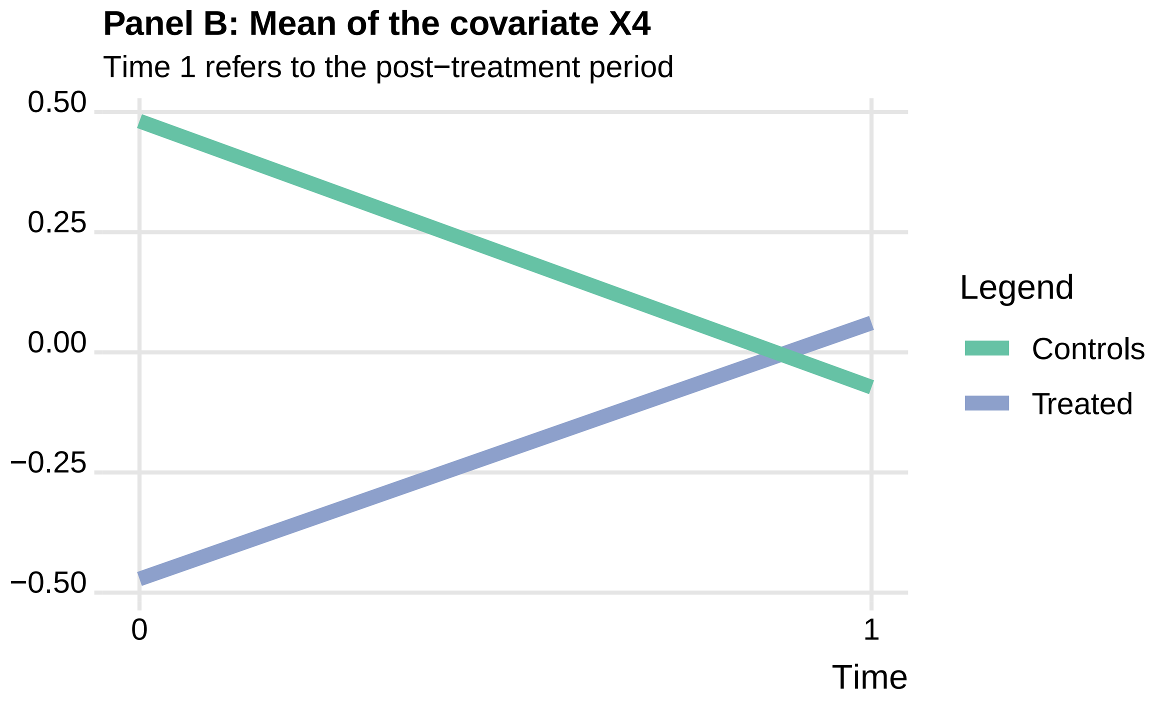
Figure A.1: Distribution of X_4 in Experiment 1



Notes: The graph considers a representative random sample from Exp.1 with DGP D. The upper plot compares the distribution of covariate X_4 among the treated and controls in the pre- and post-treatment periods. The lower plot, instead, compares the respective means among treated and controls in the two time periods. Note that the distribution of X_4 is time-invariant but there is heterogeneity between treated and controls populations, as captured by their difference in means.

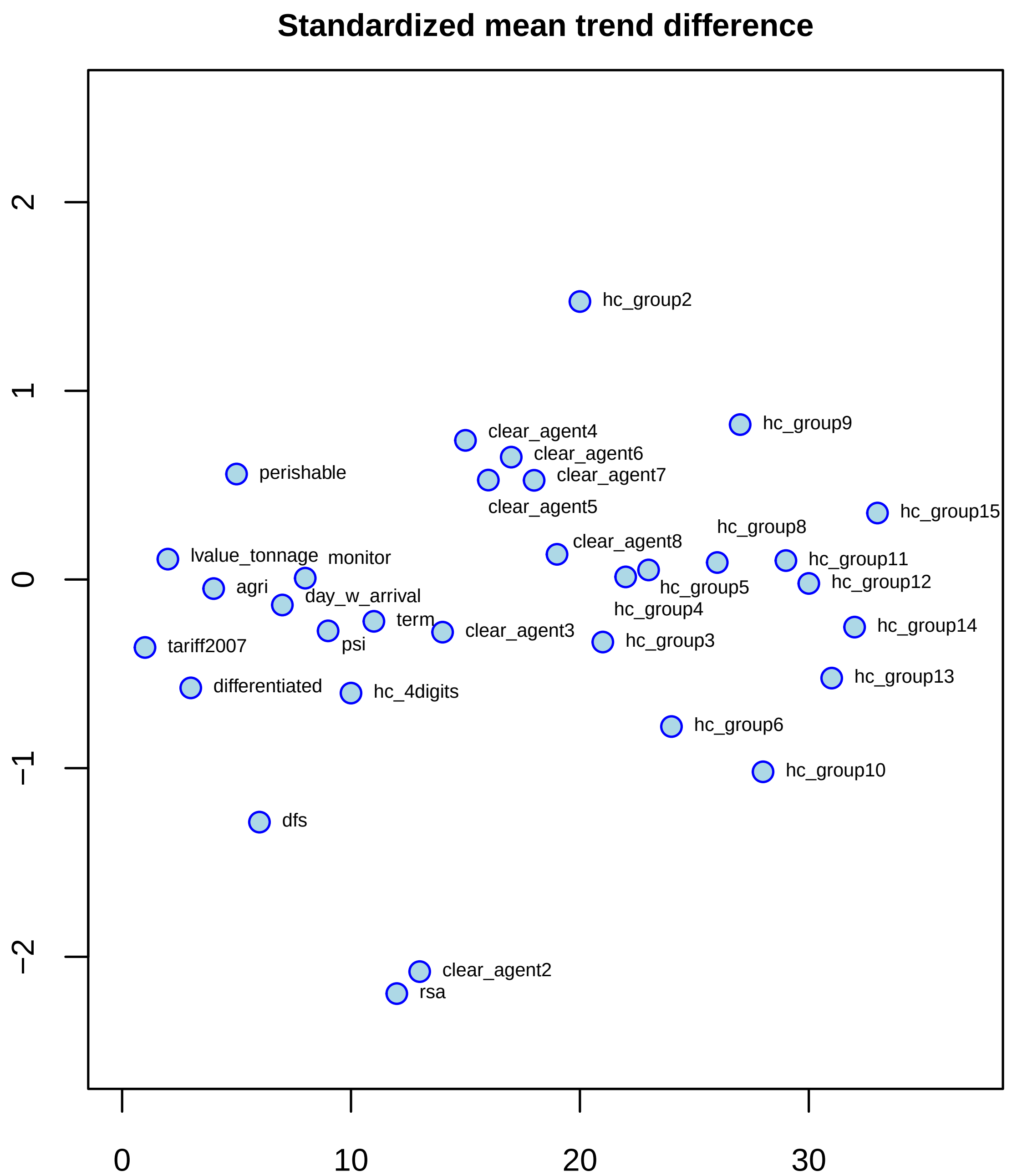
Figure A.2: Distribution of X_4 in Experiment 2





Notes: The graph considers a representative random sample from Exp.2 with DGP D. The upper plot compares the distribution of covariate X_4 among the treated and controls in the pre- and post-treatment periods. The lower plot, instead, compares the respective means among treated and controls in the two time periods. Note the heterogeneity in this case is also in the trend of the covariate between treated and controls.

Figure A.3: Standardized mean trend difference among treated and controls for each covariate



Notes: The graph plots the 33 covariates used as controls. The covariates are enumerated and ordered arbitrarily from left to right for graphical purposes. The y-axis diplay the standardized mean difference in trend among treated and controls, namely $(\overline{X}_{11} - \overline{X}_{10}) - (\overline{X}_{01} - \overline{X}_{00})$ divided by the standard deviation of X, where the overbar indicates the mean. In case of time-invariant control this measure should be 0. The figure therefore indicates the presence of a strong heterogeneity in the evolution of the covariates among treated and controls.

A.6 Appendix Tables

Table A.1: Summary table of the estimators analyzed in the Monte Carlo simulations

Estimator	Description
TWFE	Two-Way-Fixed-Effects regression with covariates as in Eq. (1)
TWFE CORR	Two-Way-Fixed-Effects correction as in Eq. (12)
IPW	Inverse probability weighting (Abadie, 2005)
DMLDiD	Debiased machine learning IPW using lasso first-stage estimates (Chang, 2020)
DIPW	Double inverse proability weighting
OR	Outcome regression (Heckman et al., 1997)
DRDiD	Locally efficient doubly robust estimator, original version (Sant'Anna and Zhao, 2020)
DR-DIPW	Locally efficient doubly robust estimator with DIPW weights
IMP DRDiD	Improved locally efficient doubly robust estimator, original version (Sant'Anna and Zhao, 2020)
IMP DR-DIPW	Improved locally efficient doubly robust estimator with DIPW weights
LASSO DR-DIPW	Locally efficient doubly robust estimator with DIPW weights and lasso first-stage estimates
RF DR-DIPW	Locally efficient doubly robust estimator with DIPW weights, random forest first-stage estimates, and sample-splitting

Table A.2: Exp.1B Propensity score model incorrect, outcome regression model correct

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	19.139	19.490	13.569	0.002
TWFE CORR	Regression, Eq. (2)	0.004	0.203	0.041	0.002
IPW					
IPW	Abadie (2005)	0.854	9.793	95.167	0.005
DMLDiD	Chang (2020)	81.060	109.279	5,371.175	43.882
DIPW	Author's work	0.833	3.980	15.146	0.009
OR					
OR	Heckman et al. (1997)	0.031	8.196	67.179	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	0.005	0.210	0.044	0.011
DR-DIPW	Author's work	0.005	0.210	0.044	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	0.005	0.211	0.045	0.015
IMP DR-DIPW	Author's work	0.005	0.211	0.045	0.025
Debiased ML					
LASSO DR-DIPW	Author's work	0.004	0.330	0.109	3.239
RF DR-DIPW	Author's work	3.773	6.073	22.648	2.296

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X, and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table A.3: Exp.1C Propensity score model correct, outcome regression model incorrect

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	13.117	14.056	25.524	0.002
TWFE CORR	Regression, Eq. (2)	1.291	4.715	20.565	0.002
IPW					
IPW	Abadie (2005)	0.030	9.204	84.717	0.005
DMLDiD	Chang (2020)	62.061	99.318	6,012.513	44.992
DIPW	Author's work	0.054	5.575	31.082	0.009
OR					
OR	Heckman et al. (1997)	1.444	8.079	63.188	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	0.004	4.748	22.543	0.011
DR-DIPW	Author's work	0.004	4.665	21.761	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	0.061	4.079	16.637	0.015
IMP DR-DIPW	Author's work	0.068	4.086	16.690	0.025
Debiased ML					
LASSO DR-DIPW	Author's work	0.226	3.173	10.018	3.151
RF DR-DIPW	Author's work	1.441	4.726	20.261	2.292

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X, and time-invariant covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table A.4: 2B Propensity score model incorrect, outcome regression model correct

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	9.165	9.875	13.518	0.002
TWFE CORR	Regression, Eq. (2)	0.003	0.217	0.047	0.002
IPW					
IPW	Abadie (2005)	51.096	52.055	98.854	0.005
DMLDiD	Chang (2020)	308.232	332.434	$15,505.580$	43.200
DIPW	Author's work	4.567	19.656	365.507	0.010
OR					
OR	Heckman et al. (1997)	32.182	33.184	65.478	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	0.003	0.221	0.049	0.011
DR-DIPW	Author's work	0.002	0.256	0.066	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	0.003	0.232	0.054	0.015
IMP DR-DIPW	Author's work	0.004	0.249	0.062	0.027
Debiased ML					
LASSO DR-DIPW	Author's work	0.241	0.387	0.092	3.590
RF DR-DIPW	Author's work	5.425	9.657	63.818	2.164

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table A.5: 2C Propensity score model correct, outcome regression model incorrect

Estimator	Reference	Bias	RMSE	Variance	Time
TWFE					
TWFE	Regression, Eq. (1)	5.816	7.863	28.010	0.002
TWFE CORR	Regression, Eq. (2)	4.966	6.720	20.499	0.002
IPW					
IPW	Abadie (2005)	31.208	32.585	87.807	0.005
DMLDiD	Chang (2020)	338.258	358.140	13,845.390	44.272
DIPW	Author's work	0.296	6.699	44.785	0.010
OR					
OR	Heckman et al. (1997)	21.740	23.115	61.659	0.004
Doubly-Robust					
DRDiD	Sant'Anna and Zhao (2020)	4.379	6.437	22.267	0.011
DR-DIPW	Author's work	0.277	5.743	32.910	0.016
IMP DRDiD	Sant'Anna and Zhao (2020)	1.092	4.562	19.616	0.015
IMP DR-DIPW	Author's work	0.519	4.569	20.604	0.027
Debiased ML					
LASSO DR-DIPW	Author's work	0.055	3.661	13.398	3.374
RF DR-DIPW	Author's work	3.909	7.512	41.154	2.177

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. TWFE is the standard regression specification with naively adding a set of covariates (Eq. (1)); TWFE CORR is the regression correction that adds also all possible interaction terms between D, T, and X (Eq. (2)); IPW is the inverse probability weighting (Eq. (5)); DMLDiD is the debiased machine learning version of the IPW estimator using lasso; DIPW ia the double inverse probability weighting estimator (Eq. (9)); DRDiD is the locally-efficient doubly robust estimator as in (Eq. (8)) and it is proposed in its "improved" version IMP DRDiD; likewise, DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is also proposed in its "improved" (IMP DR-DIPW), lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. If not otherwise specified, the propensity score is estimated with logit and the outcome model through linear regression. Finally, 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table A.6: Variables included in Γ_i

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

Table A.7: Experiment 1 with cross-fitting in machine learning algorithms

Estimator	Bias	RMSE	Variance	Time
Experiment 1A				
LASSO DR-DIPW	0.123	0.343	0.102	3.026
LASSO DR-DIPW SPLIT	0.140	0.515	0.245	10.189
RF DR-DIPW	1.535	3.403	9.228	3.695
RF DR-DIPW SPLIT	4.156	6.037	19.176	2.314
Experiment 1B				
LASSO DR-DIPW	0.004	0.330	0.109	3.239
LASSO DR-DIPW SPLIT	0.001	0.571	0.327	10.865
RF DR-DIPW	1.193	3.506	10.867	3.671
RF DR-DIPW SPLIT	3.773	6.073	22.648	2.296
Experiment 1C				
LASSO DR-DIPW	0.226	3.173	10.018	3.151
LASSO DR-DIPW SPLIT	0.460	13.622	185.340	10.954
RF DR-DIPW	0.045	3.462	11.987	3.648
RF DR-DIPW SPLIT	1.441	4.726	20.261	2.292
Experiment 1D				
LASSO DR-DIPW	1.894	3.953	12.043	3.437
LASSO DR-DIPW SPLIT	1.849	24.300	587.056	11.913
RF DR-DIPW	2.661	4.661	14.647	3.652
RF DR-DIPW SPLIT	4.665	6.797	24.431	2.318

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.1 assumes a non-randomized experiment, homogeneous effects in X, and time-invariant covariates. DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is proposed in its lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. These two estimators are compared with their respective cross-fitted versions, i.e. LASSO DR-DIPW SPLIT and RF DR-DIPW SPLIT. 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.

Table A.8: Experiment 2 with cross-fitting in machine learning algorithms

Estimator	Bias	RMSE	Variance	Time
Experiment 2A				
LASSO DR-DIPW	0.241	0.380	0.086	3.208
LASSO DR-DIPW SPLIT	0.363	0.891	0.663	11.262
RF DR-DIPW	6.155	7.197	13.904	3.338
RF DR-DIPW SPLIT	2.696	7.622	50.827	2.177
Experiment 2B				
LASSO DR-DIPW	0.022	0.303	0.091	3.384
LASSO DR-DIPW SPLIT	0.038	2.282	5.207	11.689
RF DR-DIPW	6.399	7.632	17.298	3.330
RF DR-DIPW SPLIT	5.425	9.657	63.818	2.164
Experiment 2C				
LASSO DR-DIPW	0.055	3.661	13.398	3.374
LASSO DR-DIPW SPLIT	2.337	34.211	1,164.952	12.189
RF DR-DIPW	0.093	3.885	15.083	3.331
RF DR-DIPW SPLIT	3.909	7.512	41.154	2.177
Experiment 2D				
LASSO DR-DIPW	6.631	7.968	19.513	3.555
LASSO DR-DIPW SPLIT	10.100	56.758	3,119.514	12.687
RF DR-DIPW	11.311	12.137	19.358	3.294
RF DR-DIPW SPLIT	0.894	7.952	62.438	2.166

Notes: Simulations based on sample size $n = 1000$ and 10000 Monte Carlo repetitions. EXP.2 assumes a non-randomized experiment, heterogeneous effects in X and time-varying covariates. DR-DIPW is the locally-efficient doubly robust estimator with DIPW weights (Eq. (11)), which is proposed in its lasso (LASSO DR-DIPW) and random forest (RF DR-DIPW) versions. These two estimators are compared with their respective cross-fitted versions, i.e. LASSO DR-DIPW SPLIT and RF DR-DIPW SPLIT. 'Bias', 'RMSE', 'Variance', and 'Time', stand for the average simulated absolute bias, simulated root mean-squared errors, average estimator variance, and average required computational time respectively. Refer to the main text for further details.