Marginal Treatment Effects in Difference-in-Differences

Pedro Picchetti* Cristine Pinto†

August 1, 2022

Abstract

Difference-in-Differences (DiD) is a popular method used to evaluate the effect of a treatment. In its most simple version a control group remains untreated at two periods, whereas the treatment group becomes fully treated at the second period. However, it is not uncommon in applications of the method that the treatment rate only increases more in the treatment group. This article presents identification results for the marginal treatment effect (MTE) in such fuzzy designs. We show that we can modify the standard identifying assumptions in DiD designs with covariates to identify the MTE in models with essential heterogeneity. We propose two different procedures for the estimation of the MTE that rely on different assumptions regarding the potential outcomes model and prove their asymptotical normality. Furthermore, we derive a doubly-robust estimator for the local average treatment effect (LATE) which augments the two-way fixed effects regression model with a control function and unit-specific weights that rise from the propensity score. We assert the desirable finite-sample properties through simulation studies of a linear MTE model. Finally, we use our results to investigate heterogeneity on the returns to primary school attendance in Indonesia.

Keywords: Marginal treatment effects; Heterogeneous effects; Difference-in-differences.

JEL Codes: C01, C13, C21.

*Insper. E-mail: pedrop3@al.insper.edu.br
†Insper. E-mail: cristinecxp@insper.edu.br

1 Introduction

Difference-in-Differences (DiD) is a popular method to estimate the effect of a treatment using observational data. In its canonical form, referred to as the "sharp" 2×2 DiD design, a control group remains untreated through two periods of time, whereas a treatment group becomes fully treated in the second period. Under the parallel trends assumption, the DiD estimand identifies the average effect of treatment on the treated (ATT).

However, it is not uncommon in applications of the method that the treatment rate increases more in one group than the other, but no group becomes fully treated and no group remains fully untreated, characterizing what we call "fuzzy" DiD settings, or DiD settings with imperfect compliance.

An emblematic example of a "fuzzy" DiD design in the applied economics literature is Duflo (2001) on the returns of education in Indonesia following a nationwide primary school construction program. Years of schooling increased considerably in the control group considered, while remaining unaltered for a share of the treatment group after the treatment took place.

Another example is Field (2007), in which the author investigates the effects of property rights on labour supply using a national urban titling program in Peru as a natural experiment in a DiD design. Illegal settlers are defined as the treatment group and the period after the program is defined as the post-treatment group. Imperfect compliance in this setting arises from the fact that despite being eligible for the program, not all urban squatters acquired property rights.

Fuzzy designs could also arise from settings in which there are no control and treatment groups readily available in the data that satisfy the parallel trends assumption. In such cases, identification can be achieved by creating groups for which the assumption holds. Imperfect compliance comes from the fact that there might be untreated units in the estimated treatment groups and treated units in the control group. An example is Gentzkow et al. (2011), which investigates the effect of newspaper on electoral participation in the US.

De Chaisemartin and D'Haultefoeuille (2018) were the first to explore treatment effect identification and estimation in such "fuzzy" DiD designs. The authors find that the standard methods used to estimate treatment effects in DiD settings with imperfect compliance fail to recover relevant parameters under the usual assumptions and propose new estimators to recover the local average treatment effect (LATE), which is the average treatment effect for the individuals in the treatment group that become treated in the second period, hereafter referred to as the control group "switchers". The authors show how to obtain consistent estimates for the LATE as long as the exposure to treatment in the control group remains

constant over time.

The identification of LATE in "fuzzy" DiD settings paves the way for the identification of other heterogeneous treatment effects which are relevant for policy evaluation. This paper proposes identification results and a consistent estimator for the marginal treatment effect (MTE) in the 2×2 DiD setting with imperfect compliance. The MTE is defined as the average treatment effect for individuals which are indifferent between participating in the treatment or not. Thus, it carries a meaningful economic interpretation as the individual's willingness to pay for the treatment. Moreover, the MTE is a natural building block for all the other aggregate average treatment parameters, which can be obtained as weighted averages of the MTE (Heckman and Vytlacil, 2005).

We show how we can make slight modifications in the standard identification assumptions of DiD settings with covariates to propose an estimand that identifies the MTE.

We proposed two different estimators for the MTE that rely on different assumptions regarding the functional form of potential outcomes. Both estimators follow the control function approach. We impose a functional form that relates the unobserved heterogeneity with the propensity score, and use this function to control for the endogeneity as an omitted variable.

Furthermore, we derive a doubly-robust estimator for the LATE that is consistent whether either the model for the control function is correctly specified or the model for the selection into treatment is correctly specified.

We illustrate the desirable finite-sample properties of the proposed estimators through Monte Carlo simulations of a linear MTE model[1]. The simulations show that the parametric control function estimator consistently recovers the MTE curve and has a better performance on the estimation of the LATE when compared to other estimators commonly used in the applied literature. When only one of the working nuisance models is correctly specified, the doubly-robust estimator consistently estimates the LATE and outperforms all the other DiD estimators.

We establish $\sqrt{n}$-consistency and asymptotic normality of the parametric control function estimator and the doubly robust estimator[2]. The control function estimator for the LATE and MTE is consistent when the control function and propensity score are correctly specified, and the doubly-robust estimator for the LATE is consistent when either one of the nuisance parameters is correct.

Related Literature: Our paper relates to several strands of the causal inference literature. First of all, the results are intimately related to other papers in the MTE literature. The

[1]Simulations of the semiparametric method are still in progress.

[2]Asymptotic theory for the semiparametric estimator is still in progress.

concept of marginal treatment effect was first introduced by Björklund and Moffitt (1987) as the gain from treatment for individuals who are shifted into treatment by a marginal change in its cost. Heckman and Vytlacil (1999, 2001, 2005, 2007) further defined the MTE as the gain from treatment for individuals shifted into treatment by a marginal change in the propensity score, the predicted probability of treatment (see Cornelissen et al. (2016) for a comprehensive review of the literature).

Identification and estimation of the MTE has been widely discussed in instrumental variables (IV) settings (Heckman and Vytlacil, 2001; Carneiro et al., 2011; Brinch et al., 2017). We are the first to define, identify and estimate marginal treatment effects in DiD settings.

Second, our results are directly related to the recent advances in the DiD literature. Bonhomme and Sauder (2011) consider a DiD model which allows for heterogeneous effects of time, but De Chaisemartin and D'Haultefoeuille (2018) were the first to explore the identification heterogeneous treatment effects in DiD settings with imperfect compliance. We build on the framework proposed by the authors and show how the MTE relates to the LATE parameter. Moreover, our work is also connected to Sant'Anna and Zhao (2020), as we draw insights from their results in order to adequately account for covariates in our parametric outcome regression.

Third, our work relates to the control function literature. In particular, to the work of Brinch et al. (2017), in which the authors show how to estimate the MTE using a control function approach in IV settings. We show how we can use a conditional parallel trends assumption to build a control function and consistently estimate the MTE curve in DiD designs. In order to do so, we build on the theoretical results presented by Olsen (1980), Heckman and Robb (1985) and Heckman et al. (2006).

Finally, our work is directly related to the literature on doubly-robust estimators (see Robins et al. (1994), Rothe and Firpo (2018), Sant'Anna and Zhao (2020) and Graham and Pinto (2021)). In particular, our doubly-robust estimator is an adaptation of the reshaped inverse probability weighting (RIPW) estimator from Arkhangelsky et al. (2021). We modify it to account for essential heterogeneity by including the control function and for the case of repeated cross-sections.

Organization of the paper: In the next section we define the DiD setting in a potential outcomes model with essential heterogeneity and provide an estimand for the MTE. In section 3 we propose two estimators for the MTE that rely on different assumptions and derive their large sample properties. In section 4 we present tests for the conditional parallel trends assumption and for the external validity of the estimates for LATE. In section 5 we present the doubly-robust RIPW estimator for the LATE and derive its large sample properties. We

examine the properties of the estimators by the means of a Monte Carlo simulation study in Section 6 and provide an empirical illustration in Section 7. Section 8 concludes.

2 Differences-in-Differences with heterogeneous treatment effects

2.1 Set-up and Notation

We present the framework through a model that is best suited for repeated cross sections or single cross sections in which the cohort of birth plays the role of time, but the model can be easily adapted for panel data. Data can be divided into time periods, represented a random variable T, and into groups represented by a random variable G. We focus on the 2×2 DiD setting, in which there are only two groups and two periods of time. Hence, G is a dummy for units in the treatment group, and T is a dummy for the post-treatment period. Time-invariant covariates X are also observed.

In sharp DiD designs, treatment assignment is given by the interaction between the group and period dummies, that is, $D = G \times T$. However, it is not uncommon that there are units in the control group that go into treatment, units in the treatment group that are being treated in the pre-treatment period or even units in the treatment group remain untreated through both periods. Such settings in which $D \neq G \times T$ are called "fuzzy" DiD settings.

"Fuzzy" DiD settings are somewhat analogous to Instrumental Variables (IV) settings with a single binary instrument. In this settings, there is a binary exogenous variable often denoted by Z which takes value 1 for units assigned into treatment and 0 otherwise. However, actual treatment status can differ from treatment assignment. In 2×2 DiD settings, we can interpret the interaction between group and time variables as the treatment assignment. Sharp designs are characterized by full compliance of treatment status towards the treatment assignment, whereas the "fuzzy" design can be interpreted as as DiD setting with imperfect compliance.

"Fuzzy" designs could also arise from the fact that there are no groups readily available for which the assumptions in the setting are valid. In such case, these treatment and control groups require estimation (see De Chaisemartin and D'Haultefoeuille (2018)) and the inequality between treatment status and the interaction between group and time comes precisely from such estimation procedure.

Throughout the article we use the notation introduced by De Chaisemartin and D'Haultefoeuille (2018). For any random variable R, let $\mathbf{S}(R)$ denote its support. Moreover, define the random variables R_{gt} and R_{dgt} such that $R_{gt} = R|G = g, T = t$ and $R_{dgt} = R|D = d, G =$

$g, T = t$. For example, the notation implies that $E(R_{10}) = E(R|G = 1, T = 0)$ and $E(R_{110}) = E(R|D = 1, G = 1, T = 0)$.

2.2 Framework and treatment effects

We are interested in the effect of a binary treatment D on some outcome Y. The potential outcomes for a unit with and without treatment are respectively denoted by $Y(1)$ and $Y(0)$. The realized outcome is $Y = DY(1) + (1 - D)Y(0)$.

Our general framework is a model based on potential outcomes and a latent variable discrete choice model for selection into treatment, as it has been established in the MTE literature since Heckman and Vytlacil (1999).

We specify potential outcomes as

$$Y(0) = \mu_0(G, T, X) + U_0 \tag{1}$$

$$Y(1) = \mu_1(G, T, X) + U_1 \tag{2}$$

The function $\mu_j(G, T, X)$ is the conditional mean of Y given G,T and X in treatment status $j \in \{0, 1\}$, such that $E(U_j|G, T, X) = 0$. We do not assume that the vectors (G, T, X) and (U_1, U_0) are independent.

We consider the following selection into treatment model, which is the foundation of the MTE approach:

$$D = \mathbb{1}\{P_D(G, T, X) \geq U_D\} \tag{3}$$

Equation (3) imposes a single threshold crossing model for selection into treatment. The function $P_D(G, T, X)$ represents the net benefits from receiving treatment, which are a function of the group and time variables and time-invariant covariates X. The threshold U_D is a random variable representing the unobservable distaste for the treatment, also referred to as the essential heterogeneity among individuals. We assume that both terms in the inequality are bounded within the unit interval, so that $P_D(G, T, X)$ has a propensity score interpretation[3].

The participation equation implies that units from the same group switch treatment status in the same direction. This monotonicity assumption will be key for our identification results.

[3]This assumption is made for simplicity of the exposition only. If the terms are not bounded within the unit interval, we can obtain the propensity score interpretation by applying the CDF of U_D on both sides of the inequality.

We can define the selection into treatment as a function of time:$D(t) = \mathbb{1}\{P_D(G, t, X) \geq U_D\}$. Let $S = \{D(1) > D(0), G = 1\}$ denote the units from the treatment group which go from non-treatment to treatment with the passage of time, defined by De Chaisemartin and D'Haultefoeuille (2018) as the "control group switchers". The control group switchers are analogous to the compliers in IV settings, in the sense that they're treatment status is defined by the treatment assignment.

Our model essentially boils down to the switching regression model (Quandt, 1972; Lee, 1979). The potential outcomes framework can be expressed by the following random coefficients regression model:

$$Y_i = \mu_0(G_i, T_i, X_i) + D_i[\mu_1(G_i, T_i, X_i) - \mu_0(G_i, T_i, X_i) + U_{1i} - U_{0i}] + U_{0i} \tag{4}$$

in which the individual treatment effect is given by

$$\Delta_i = \mu_1(G_i, T_i, X_i) - \mu_0(G_i, T_i, X_i) + U_{1i} - U_{0i} \tag{5}$$

The individual treatment effect can be decomposed in the average gains of treatment for a unit in a given group, period of time and observed characteristics, and an individual-specific gain, $U_{1i} - U_{0i}$.

The existence of individual specific gains, also referred to as essential heterogeneity, implies that aggregate treatment effects will be different from each other. In sharp DiD designs, for example, the treatment effect that can be identified is the average treatment effect on the treated (ATT). Conditional on $X = x$, it is expressed in our framework as

$$\Delta^{ATT}(x) = E(Y_{11}(1) - Y_{11}(0)|X = x) = \mu_1(1, 1, x) - \mu_0(1, 1, x)$$

In "fuzzy" designs, however, De Chaisemartin and D'Haultefoeuille (2018) show that we cannot identify the ATT, but a local average treatment effect (LATE) is identifiable, which is the average treatment effect for the "control group switchers":

$$\Delta^{LATE}(x) = E(Y_{11}(1) - Y_{11}(0)|X = x, S) = \mu_1(1, 1, x) - \mu_0(1, 1, x) + E(U_1 - U_0|G = 1, T = 1, X = x, S)$$

The unobservable gains that are part of the LATE can be expressed as function of the U_D variable that drives selection into treatment:

$$E(U_1 - U_0|G = 1, T = 1, X = x, S)$$
$$= E(U_1 - U_0|G = 1, T = 1, X = x, P_D(1, 0, x) < U_D \leq P_D(1, 1, x))$$

That is, the control group switchers are the units in control group that are not treated in the pre-treatment period, but become treated in the post-treatment period.

While the LATE aggregates treatment effects over a certain range of the U_D distribution, the MTE is defined as the average treatment effect at a particular value of U_D. Before we define the MTE, we introduce the following notation for the conditional expectation of U_1 and U_0:

$$k_j(g, t, x, p) = E(U_j|G = g, T = t, X = x, U_D = p),\ d = \{0, 1\}$$

and

$$k(g, t, x, p) = E(U_1 - U_0|G = g, T = t, X = x, U_D = p)$$

The MTE can then be expressed as follows:

$$\Delta^{MTE}(x, p) = \mu_1(1, 1, x) - \mu_0(1, 1, x) + k(1, 1, x, p)$$

where $\mu_1(1, 1, x) - \mu_0(1, 1, x)$ is the average treatment effect for units in the treatment group and the post-treatment period, and $k(1, 1, x, p)$ the average unobservable gain from treatment for individuals with the distaste for treatment equal to p. Conditioning on $U_D = p$ is equivalent to conditioning on the intercept of $P_D(G, T, X) = p$ and $P_D(G, T, X) = U_D$. Therefore, the MTE can be interpreted as the average effect of treatment for units from the treatment group in the post-treatment period on a margin of indifference between participation in treatment and nonparticipation.

While the MTE is defined for a particular value of the U_D distribution, the other treatment effects that can be identified in DiD settings are defined over different ranges of U_D. Thus, the MTE appears as the building-block for other average treatment effects. Heckman and Vytlacil (1999) show that the aggregate average treatment parameters can be expressed as weighted averages of the MTE. In DiD setting, this can be expressed as

$$\Delta^{ATT}(x) = \int_0^1 \Delta^{MTE}(x, u_D) du_D$$

$$\Delta^{LATE}(x, S) = \frac{1}{\|S\|} \int_S \Delta^{MTE}(x, u_D) du_D$$

Note that in the sharp design essential heterogeneity plays no role in selection into treatment, from which it follows that for $d = \{0, 1\}$, $E(U_d|G, T, X, U_D = p) = E(U_d|G, T, X) = 0$. Thus, treatment effects are constant and equal to the ATT.

2.3 Identification

We now invoke the main assumptions that are used for identification of the MTE in this setting. The first concerns the evolution of the share of treated units in the treatment and control group:

Assumption 1. (Fuzzy Design). Almost surely, $E\left(D_{11} \mid X\right) > E\left(D_{10} \mid X\right)$ and $E\left(D_{11} \mid X\right) - E\left(D_{10} \mid X\right) > E\left(D_{01} \mid X\right) - E\left(D_{00} \mid X\right)$.

Assumption 1 is used to represent treatment and control group in the setting. The treatment group is the one which experiences the greater increase in its treatment rate. The sharp design can be viewed as an extreme case of the fuzzy setting, in which $E\left(D_{11} \mid X\right) - E\left(D_{10} \mid X\right) = 1$ and $E\left(D_{01} \mid X\right) = E\left(D_{00} \mid X\right)$. It is not uncommon the DiD literature that there is imperfect compliance among the treatment group, but there is a pure treatment group with treatment rate equal to zero (Field, 2007), this case is also contemplated by our framework. Assumption 1 only rules out the case in which the two groups experience the same evolution in their treatment rates.

The next assumptions are standard in DiD settings, but we modify them to account for essential heterogeneity among units:

Assumption 2. (Conditional Parallel Trends). Almost surely, $E(Y_{G1}(0) - Y_{G0}(0)|X, P_D = p)$ does not depend on G.

Assumption 3. (Common Support). For some $\varepsilon > 0$, $Pr(D_{gt} = 1) > \varepsilon$ and $Pr(D_{gt} = 1|X) \leq 1 - \varepsilon$, almost surely.

Assumption 2 is the conditional parallel trends assumption, which is standard in DiD settings with covariates. We modify it to account for the heterogeneity that arises from the unobservables driving selection into treatment. The conditional parallel trends assumption implies that the mean evolution of the outcome through time across groups would be the same in the absence of treatment, conditional on covariates and the propensity score. Assumption

3 is the common support assumption, it asserts that at least a small fraction of the population is treated and that for every value of the covariates X, there is at least a small probability that the unit is not treated, which is also a standard assumption in conditional DiD methods.

3 Estimation of the MTE Curve

In this section we provide two different estimators for the MTE in the fuzzy 2×2 setting. Both methods are control function estimators that are consistent under different assumptions regarding i) separability of the potential outcomes in terms of X and $U_D = p$ and ii) the functional form of $\mu_j(.)$. Our control function approach conditions on D and $P_D = p$.

From our potential outcomes model and the common support assumption, we have that

$$\begin{aligned} E(Y_{1GT}|X) = E(Y_{GT}(1)|X, D=1) &= \mu_1(G,T,X) + E(U_1|G,T,X,D=1) \\ &= \mu_1(G,T,X) + E(U_1|G,T,X,U_D \leq P_D(G,T,X)) \\ &= \mu_1(G,T,X) + K_1(G,T,X,P_D(G,T,X)) \end{aligned}$$

and

$$\begin{aligned} E(Y_{0GT}|X) = E(Y_{GT}(0)|X, D=0) &= \mu_0(G,T,X) + E(U_0|G,T,X,D=0) \\ &= \mu_0(G,T,X) + E(U_0|G,T,X,U_D > P_D(G,T,X)) \\ &= \mu_0(G,T,X) + K_0(G,T,X,P_D(G,T,X)) \end{aligned}$$

where $K_j = E(U_j|G,T,P_D(G,T,X))$ are control functions (Olsen, 1980; Heckman and Robb, 1985). We recover the unobservable gains ($k(1,1,x,p)$) by identifying $E(U_1|G,T,X,U_D = u_D)$ and $E(U_0|G,T,X,U_D = u_D)$ separately. Heckman and Vytlacil (2001) show that

$$\begin{aligned} k_1(G,T,X,p) &\equiv E(U_1|G,T,X,U_D = p) \\ = p\frac{\partial K_1(G,T,X,P_D=p)}{\partial p} &+ K_1(G,T,X,P_D=p) \end{aligned}$$

and

$$k_0(G,T,X,p) \equiv E(U_0|G,T,X,U_D=p)$$
$$= -(1-p)\frac{\partial K_0(G,T,X,P_D=p)}{\partial p} + K_0(G,T,X,P_D=p)$$

Therefore, the first step in any estimation method is the estimation of the propensity score trough a $\sqrt{n}$ consistent method (Probit, Logit, Linear Probability Model), which will be used to estimate the control function in a first moment, and later to differentiate it in order to obtain the MTE. Below, we provide two different strategies for the estimation of the control function and the MTE curve, which rely on different assumptions regarding the functional form of the potential outcomes. Thus, applied researchers have freedom to choose the method that seems more suited for the particular application. In both approaches, we follow Brinch et al. (2017) and estimate control functions separately for each treatment status, which allows the identification of richer specifications for the MTE model.

3.1 Semiparametric model

First, we consider a semiparametric estimator, which requires an additional assumption:

Assumption 4. (Additive Separability). $E(Y_{dgt}|X=x,U_D=p) = \mu_d(g,t,x) + k_d(g,t,p)$, for $(d,g,t) \in \{0,1\}^3$.

Assumption 4 implies that the unobservable gains as a function of U_D do not depend on X, so that the MTE is additive separable in X and U_D:

$$\Delta^{MTE}(x,p) = \mu_1(1,1,x) - \mu_0(1,1,x) + k(1,1,p)$$

When justified, the separability assumption allows for the functional form of the control function to remain unespecified, which is fundamental in the proposed semiparametric two-step procedure, which adapts the double residual regression from Robinson (1988) and Heckman et al. (1997).

In the first step, we specify $\mu_j(G,T,X)$ as linear function of the covariates: $\mu_d(G,T,X) = \beta_{dGT}X$. The slope coefficients are allowed to vary with treatment status, period and group.

The outcome model we estimate is

$$Y_{igt}(1) = X_i\beta_{1gt} + K_1(g,t,P_i) + \varepsilon_{i1}, \ \ for\ i \in \{D=1, G=g, T=t\}$$
$$Y_{igt}(0) = X_i\beta_{0gt} + K_0(g,t,P_i) + \varepsilon_{i0}, \ \ for\ i \in \{D=0, G=g, T=t\}$$

In the implementation of the method, we replace P_i with $\widehat{P}_i$ which is estimated using a $\sqrt{n}$-consistent method.

We obtain identification of β_{dgt} by removing the parts of $Y_i(d)$ and X_i that are dependent on P_i, D_i, G_i and T_i:

$$Y_{igt}(d) - E(Y_{dgt}|P_i) = (X - E(X_{dgt}|P_i))\beta_{dgt} + \varepsilon_{id}$$

Let $\tilde{Y}_{gt}(d) = (Y_{gt}(d) - E(Y_{dgt}|\widehat{P}_i)), \tilde{X}_{gt}(d) = (X - E(X_{dgt}|\widehat{P}_i))$

We estimate β_{dgt} by

$$\widehat{\beta}_{dgt} = (\tilde{X}_{dgt}^T \tilde{X}_{dgt})^{-1} \tilde{X}_{dgt}^T \tilde{Y}_{dgt}$$

The first step of the procedure recovers the observable gains from treatment. We estimate $\widehat{\mu}_d(g,t,x) = \widehat{\beta}_{dgt}x$. Note that, by the conditional parallel trends assumption, we have that

$$\widehat{\mu}_0(1,1,x) = \widehat{\mu}_0(1,0,x) + (\widehat{\mu}_0(0,1,x) - \widehat{\mu}_0(0,0,x))$$

Thus, we recover the observable gains through the difference-in-differences of the estimated values from the first stage:

$$\widehat{\mu}_1(1,1,x) - \widehat{\mu}_0(1,1,x) = \widehat{\mu}_1(1,1,x) - \widehat{\mu}_0(1,0,x) - (\widehat{\mu}_0(0,1,x) - \widehat{\mu}_0(0,0,x))$$

After the estimation of β_{dgt}, we estimate the control functions nonparametrically using the residual-adjusted outcome model.

Define $c_i(d) = Y_{idgt} - X\widehat{\beta}_{dgt}$. Let$K(.)$ denote a kernel function and k a bandwidth. The pointwise estimators for $K_d(g,t,p)$ and $\gamma_d(p) = \frac{\partial K_d(g,t,p)}{\partial p}$ in the neighborhood of p are defined as

$$(\widehat{K}_d(p), \widehat{\gamma}_d(p)) = \underset{K_d,\gamma_d}{argmin} \sum_{i \in \{D=d, G=g, T=t\}} (c_i - K_d(p) - \gamma_d(p)(\widehat{P}_i - p))^2 K(\frac{\widehat{P}_i - p}{h})$$

The estimation allows for the recovery of the conditional expectations of U_1 and U_0 on U_D as presented by Heckman and Vytlacil (2001), from which we are able to recover the unobservable gains from treatment. The conditional parallel trends assumption combined with the separability assumption implies that

$$\widehat{k}_0(1,1,p) = \widehat{k}_0(1,0,p) - (\widehat{k}_0(0,1,p) - \widehat{k}_0(0,0,p))$$

Hence, we recover the unobservable gains from treatment through the difference-in-differences of the estimated conditional expectations of the unobservable terms in the potential outcomes model:

$$\widehat{k}_1(1,1,p) - \widehat{k}_0(1,1,p) = \widehat{k}_1(1,1,p) - \widehat{k}_0(1,0,p) - (\widehat{k}_0(0,1,p) - \widehat{k}_0(0,0,p))$$

The MTE is thus estimated by putting together the observable and unobservable gains obtained from the semiparametric procedure. Let $\widehat{E}(Y_{dgt}|X = x, U_D = p) = \widehat{\mu}_d(g,t,x) + \widehat{k}_d(g,t,p)$. We estimate the MTE by

$$\begin{aligned}\widehat{\Delta}^{MTE}(x,p) = \widehat{E}(Y_{111}|X = x, U_d = p) - \widehat{E}(Y_{010}|X = x, U_d = p) - \\ (\widehat{E}(Y_{001}|X = x, U_d = p) - \widehat{E}(Y_{000}|X = x, U_d = p))\end{aligned}$$

In short, the semiparametric estimation of the MTE can be performed in six steps:

1. Estimation of the propensity score using a $\sqrt{n}$-consistent method (e.g. Logit, Probit, LPM).
2. Regress Y onto the estimated $\widehat{P}_D$ and X onto $\widehat{P}_D$.
3. Build the variables $\tilde{Y}_{dgt} = Y_{dgt} - \widehat{E}(Y_{dgt}|\widehat{P}_D)$ and $\tilde{X}_{dgt} = X_{dgt} - \widehat{E}(X_{dgt}|\widehat{P}_D)$.
4. Recover the observable gains from treatment by regressing $\tilde{Y}_{dgt}$ onto $\tilde{X}_{dgt}$ through OLS regression.
5. Build the residual variables $c_{dgt} = Y_{dgt} - X_{dgt}\widehat{\beta}_{dgt}$
6. Recover the unobservable gains by regressing c_{dgt} onto $\widehat{P}_D$ using a local linear regression.

Asymptotic theory for the semiparametric model is not complete yet. At this moment, we suggest the use of boostrap for inference.

3.2 Fully parametric polynomial model

The approach outlined in the previous setting relies on the additive separability assumption, which is not justifiable in all applied settings. In the cases the assumption fails to hold, we propose an alternative fully parametric estimator, which relies on a different assumption regarding the functional form of potential outcomes:

Assumption 5 . (The role of covariates on the potential outcomes model). For $j \in \{0,1\}$, $\mu_j(G,T,X) = \mu_j(G,T)$.

Assumption 5 states that covariates only affect potential outcomes through its unobservable components. The assumption is necessary in order to build a parametric estimator that consistently recovers covariate-specific treatment effects under covariate-specific trends (see Sant'Anna and Zhao (2020) for a detailed discussion on DiD estimation with covariates).

Assumption 5 implies that the potential outcomes model can be expressed by the following two-way fixed effects (TWFE) regression model augmented by the control functions:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 T_i + \tau D_i + D_i K_1(G_i, T_i, X_i, P_i) + (1 - D_i) K_0(G_i, T_i, X_i, P_i) + u_i \quad (6)$$

The parametric estimation of the MTE requires a specification for the functional form of the unobservable heterogeneity in the potential outcomes model. We follow Brinch et al. (2017) and Cornelissen et al. (2018) and specify the functions $K_1(.)$ and $K_0(.)$ as polynomials in the propensity score p.

In the subsection below, we illustrate the parametric estimator for a linear MTE model.

3.2.1 Linear MTE Model

We now define the conditional expectations for U_1 and U_0, characterizing our linear MTE model[4]:

$$k_0(G, T, X, p) = (\delta G + \gamma T + \pi_0 X)(p - \frac{1}{2})$$

and

$$k_1(G, T, X, p) = (\delta G + \gamma T + \pi_1 X)(p - \frac{1}{2})$$

In this case, the MTE is linear and given by

$$\Delta^{MTE}(x, p) = \tau + (\pi_1 - \pi_0) x (p - \frac{1}{2})$$

From the expressions above, we derive

$$\begin{aligned} K_1(G, T, X, p) &= \frac{1}{p} \int_0^p E(U_1 | G, T, X, U_D = u) du \\ &= \frac{1}{2} (\delta G + \gamma T + \pi_1 X)(p - 1) \end{aligned}$$

[4]The constant ensure that the marginal expectations of U_1 and U_0 are zero if we assume that U_D has a standard uniform distribution

and

$$K_0(G,T,X,p) = \frac{1}{1-p}\int_p^1 E(U_0|G,T,X,U_D = u)du = \frac{1}{2}(\delta G + \gamma T + \pi_0 X)p$$

which are the control functions that can be recovered through the parametric estimator. Parametric estimation of the MTE, thus, can be implemented in three steps.

1 . Estimation of the propensity score through a n-consistent method (e.g. Probit, Logit, LPM).

2 . Specification of the control functions as polynomials with respect to the propensity score and OLS estimation of the regression model from equation (6).

3. Plug-in estimation of the MTE using the estimates from item 2.

In the linear model outline above, the MTE estimate amounts to

$$\widehat{\Delta}^{MTE}(x,\widehat{p}) = \widehat{\tau} + (\widehat{\pi}_1 - \widehat{\pi}_0)x(\widehat{p} - \frac{1}{2})$$

where $\widehat{p}$ is recovered in Step 1, and the estimates for the parameters come from Step 2.

We now study the asymptotic properties of the control function estimator for the MTE parameters assuming we have an i.i.d sample for (Y, D, G, T, X).

Assumption 6: $(Y_i, D_i, G_i, T_i, X_i)_{i=1,...,n}$ are i.i.d.

Theorem 1 shows that our parametric estimator is $\sqrt{n}$- consistent and asymptotically normal.

Theorem 1 *Let Assumptions 1-3, 5 and 6 hold, then under standard regularity conditions*

$$\sqrt{n}(\widehat{\Delta^{MTE}}(x,\widehat{p}) - \Delta^{MTE}(x,p)) \sim N(0, V_{CF})$$

where V_{CF} is defined in the section 2 of the Appendix. Moreover, the bootstrap is consistent for $\widehat{\Delta^{MTE}}$.

Our two-step control function can be interpreted as a two-stage GMM estimator (Hansen, 1982). We derive the variance for the MTE estimator in the appendix. The derived expression can be used to manually correct the estimated second-stage variances for the use of the estimated propensity score. However, because the variance can take complicated expressions, using the bootstrap might be convenient for inference. Theorem 1 also shows that bootstraped confidence intervals are asymptotically valid for our estimator.

4 Internal and External Validity Tests

4.1 Testing for Conditional Parallel Trends and Violations of the Functional Form

We now propose a simple test which can be used to test both for the conditional parallel trends assumptions and for violations of the functional form of the model.

We follow Borusyak et al. (2021) and perform the test using untreated observations only. Despite the drawback of losing statistical power by dropping treated observations from the procedure, the choice allows us to overcome traditional challenges in pre-trend testing in DiD designs (see Roth (2022) for a thorough discussion).

We propose a test that comes directly from the functional form for the untreated outcomes that is implied by the conditional parallel trends assumption. Assumptions 3 and 5 imply the following functional form for untreated outcomes:

$$Y(0) = \mu_0(G, T) + K_0(G, T, X, P_D = p)$$

A simple test for the conditional parallel trends assumption is to consider a richer specification for $Y(0)$ than the one imposed by Assumption 3 and test for the statistical significance of the additional regressors.

More specifically, we consider the following regression model,

$$Y_i(0) = \mu_0(G_i, T_i) + K_0(G_i, T_i, X_i, P_D = p) + W_i^T \Theta + \varepsilon_i$$

estimate Θ by $\widehat{\Theta}$ using OLS. The choice of W is therefore fundamental for our validity test. In settings with observations available for multiple periods before treatment, a natural choice is a set of indicator variables for pre-treatment periods (De Chaisemartin and D'Haultefoeuille, 2020; Sant'Anna and Callaway, 2021). In the 2×2 setting, however, the construction of the placebo treatments is not so straightforward. Assume we have data available for period $T = -1$. We can run the regression above for periods $T = 0$, $T = -1$ and define $W = G \times T$. In the absence of pre-treatment observations , we can test for violations of the functional form of potential outcomes by building W using variables that are assumed to be correlated to the covariates in the model, but do not affect the potential outcomes. The choice of the variables should follow context-specific economic knowledge in order to adequately assess the possible violations of conditional parallel trends (Rambachan and Roth, 2022).

4.2 Testing the External Validity of LATE in a Linear Model

Recent work has focused on proposing several tests for the external validity of the LATE in IV settings (Heckman et al., 2010; Angrist and Fernandez-Val, 2013; Brinch et al., 2017).

The control function estimation approach offers a simple test for the external validity of LATE in a linear MTE model. Specifically, we reject the external validity of the treatment parameter if the slope of the MTE model is different from zero, that is, the MTEs are nonconstant.

The MTEs are constant in the absence of unobservable variables driving selection into treatment. Thus, constant MTEs imply that all the average treatment effects from the DiD literature will have the same value, and therefore the LATE estimate will be informative for all population. If the MTEs are nonconstant, however, then the LATE is informative only for the population of treatment group switchers.

Our test is similar to the one presented by Brinch et al. (2017). Define $\Delta_j = E(Y_{j11}) - E(Y_{j10}) - (E(Y_{j01}) - E(Y_{j00}))$. In a linear MTE model, testing the null hypothesis of a constant MTE (i.e., $U_1 - U_0 \perp U_D$) versus the alternative of a nonconstant MTE is equivalent to testing the null hypothesis

$$H_0 : \Delta_1 = \Delta_0$$

versus a two-sided alternative hypothesis.

A straightforward way to implement the test is to run the following regression:

$$Y_i = \alpha_0 + \alpha_1 G_i + \alpha_2 T_i + \alpha_3 D_i + \alpha_4 G_i \cdot T_i + \alpha_5 D_i \cdot G_i \cdot T_i + e_i$$

and perform a two-sided t-test on the estimate for α_5.

The intuition of the test is pretty clear: If we are dealing with a constant MTE model, then it must be the case that treatment effects are independent from the variables that drive selection into treatment.

5 Doubly-Robust estimation of LATE

In the 2×2 setting, our parametric estimator is equivalent to a TWFE estimator augmented by the control function, as illustrated by equation (6). In the regression model, the parameter τ represents the LATE.

We propose a doubly-robust estimator for the LATE, which is the TWFE specification augmented by the control function and unit-specific rates which are a function of the propensity score and a reshaped distribution of selection into treatment.

The estimator is a function of two nuisance parameters: the estimated propensity score $\widehat{P}_D(G,T,X)$ and the outcome model $m_{ij}(G,T,X) = \mu_0(G,T) + K_j(G,T,X,p)$. In that sense, it is an adaptation of the Reshaped Inverse Probability Weighting (RIPW) estimator for panel data proposed by Arkhangelsky et al. (2021). We modify it to account for essential heterogeneity by including the control function and for the case of repeated cross-sections.

Given an estimate $\widehat{P}_D(G,T,X)$ for the propensity score and estimates for $m_j(G,T,X,p)$ we consider the following RIPW estimator:

$$\widehat{\tau}(\pi) = \underset{\mu,\tau}{argmin} \sum_{i=1}^{n}((Y_i - m_{ij}(G_i,T_i,X_i,p_i)) - \mu - D_i\tau)^2 \frac{\pi(D_i;X_i)}{\widehat{P}_D(G_i,T_i,X_i)} \tag{7}$$

where $\pi(D_i;X_i)$ is a density for D_i conditional on the covariates X_i [5].

Theorem 3 defines the form of our doubly-robust estimator for the LATE:

Theorem 2 *Let* $\Theta_i = \frac{\pi(D_i;X_i)}{\widehat{P}_D(G_i,T_i,X_i)}$ *and* $\widetilde{Y}_i = Y_i - m_{ij}(G_i,T_i,X_i,\widehat{P}_D(G_i,T_i,X_i))$. *Furthermore, define* $\Gamma_\Theta = \frac{1}{n}\sum_{i=1}^{n}\Theta_i$, $\Gamma_{DD} = \frac{1}{n}\sum_{i=1}^{n}\Theta_i D_i^2$, $\Gamma_{DY} = \frac{1}{n}\sum_{i=1}^{n}\Theta_i D_i\widetilde{Y}_i$, $\Gamma_D = \frac{1}{n}\sum_{i=1}^{n}\Theta_i D_i$ *and* $\Gamma_Y = \frac{1}{n}\sum_{i=1}^{n}\Theta_i\widetilde{Y}_i$.

Under Assumptions 1-6,

$$\widehat{\tau}(\pi) = \frac{A}{B} = \frac{\Gamma_{DY} - \Gamma_\Theta^{-1}\Gamma_D\Gamma_Y}{\Gamma_{DD} - \Gamma_\Theta^{-1}\Gamma_D^2}$$

In section 3 of the Appendix we derive a linear asymptotic expansion of the RIPW estimator from which we obtain its corresponding influence function:

$$N_* = \frac{1}{2n}\sum_{i=1}^{n} E[\mathbf{V}_i] = E[\Gamma_{DY}]E[\Gamma_\Theta] - E[\Gamma_D]E[\Gamma_Y] - \tau(E[\Gamma_{DD}]E[\Gamma_\Theta] - E[\Gamma_D]^2)$$

where

$$\mathbf{V}_i = \Theta_i\Big\{(E[\Gamma_{DY}] - \tau E[\Gamma_{DD}]) - (E[\Gamma_Y] - \tau E[\Gamma_D])D_i + E[\Gamma_\Theta]D_i(\widetilde{Y}_i - \tau D_i) - E[\Gamma_D](\widetilde{Y}_i - \tau D_i)\Big\}$$

Theorem 3 shows an important property from the RIPW estimator, which is the double-robustness property:

Theorem 3 *Let Assumptions 1-6 hold, then* $\widehat{\tau}(\pi)$ *is a consistent estimator of* τ *if*

[5] Arkhangelsky et al. (2021) show that the density can be arbitrary in the 2×2 DiD setting, as long as positive probabilities are assigned for both treatment status.

1. $\widehat{P}_D(G,T,X) \underset{\rightarrow}{p} P_D(G,T,X)$, *a.s., or*

2. $\widehat{m}_j(G,T,X,p) \underset{\rightarrow}{p} m_j(G,T,X,p)$ *a.s..*

Theorem 3 states that as long as one of the nuisance parameters is correctly specified, we can consistently recover the LATE. Therefore, the RIPW estimator is less demanding when it comes to the researcher's ability to specify the selection into treatment model or the unobserved gains from treatment (see Section 4 of the Appendix for the proof).

We use the plug-in estimates of $\mathbf{V}_i$ to estimate a conservative variance of the term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{V}_i - E[\mathbf{V}_i])$ as

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\mathbf{V}}_i - \overline{\widehat{\mathbf{V}}}_i)^2$$

where $\overline{\widehat{\mathbf{V}}}_i = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{V}}_i$. This yields the following Wald-type confidence interval for τ:

$$\widehat{C}_{1-\alpha} = [\widehat{\tau}(\pi) - z_{1-\alpha/2\widehat{\sigma}}/\sqrt{n}B, \widehat{\tau}(\pi) + z_{1-\alpha/2\widehat{\sigma}}/\sqrt{n}B] \tag{8}$$

where z_η is the η_{-th} quantile of the standard normal distribution (see Section 3 of the appendix for more details).

6 Monte Carlo Simulation Studies

In this section, we conduct a series of Monte Carlo experiments to assess the finite sample properties of our proposed estimators. We compare our proposed parametric control function (CF-DID) and RIPW estimators to the standard unweighted TWFE linear regression model, the doubly-robust DiD estimator (DR-DID) proposed by Sant'Anna and Zhao (2020), the Wald-DID estimator used in Duflo (2001) and the Time-Corrected WALD-DID estimator (TC-WALD) as proposed by De Chaisemartin and D'Haultefoeuille (2018).

In all simulation exercises, we consider a linear probability model for selection into treatment and a linear working model for the evolution of the outcomes.

$$D = \mathbb{1}\{\alpha_1 X + \alpha_2 G + \alpha_3 T + \alpha_4 G \times T \geq U_D\}$$

where U_D follows a standard uniform distribution and the coefficients are specified such that the left hand side of the inequality is also bounded within the unit interval.

Furthermore, we assume that the unobservable terms from the potential outcomes and the treatment participating equation are linearly correlated, characterizing, thus, a linear

model for the MTE as the one presented in Section 3.2.1.[6].[7]

We specify the potential outcomes as

$$\begin{aligned} Y(0) &= \mu_0(G,T) + U_0 \\ &= \beta_0 + \beta_1 G + \beta_2 T + U_0 \end{aligned}$$

and

$$\begin{aligned} Y(1) &= \mu_1(G,T) + U_1 \\ &= \beta_0 + \beta_1 G + \beta_2 T + \tau + U_1 \end{aligned}$$

where $E(U_0|G,T,X) = E(U_1|G,T,X) = 0$. We consider a linear functional form for the observable terms of the potential outcomes where the conditional parallel trends assumptions follows naturally from the linear model with additive separability between the time and group variables.

The parameter τ represents both the ATT and the LATE in this DGP, given that the MTE is linear. We set its value to be equal to zero in the simulations. The parameters that determine the unobservable gains from treatment are defined such that the slope of the MTE curve is equal to 1 ($\pi_1 - \pi_0 = 1$).

We consider a sample size n equal to 1000. We compare the various DiD estimators estimators for the LATE in terms of average bias, median bias, root mean squared error (RMSE), empirical 95% coverage probability and the average length of the 95% confidence interval.

Figure 1 shows the comparison between the true MTE curve and the one estimated by our procedure, for the mean value of the covariates. The blue line represents the true MTE curve, the red line shows the estimated curve and the dashed lines show us the bounds for the 95% bootstrapped confidence interval.

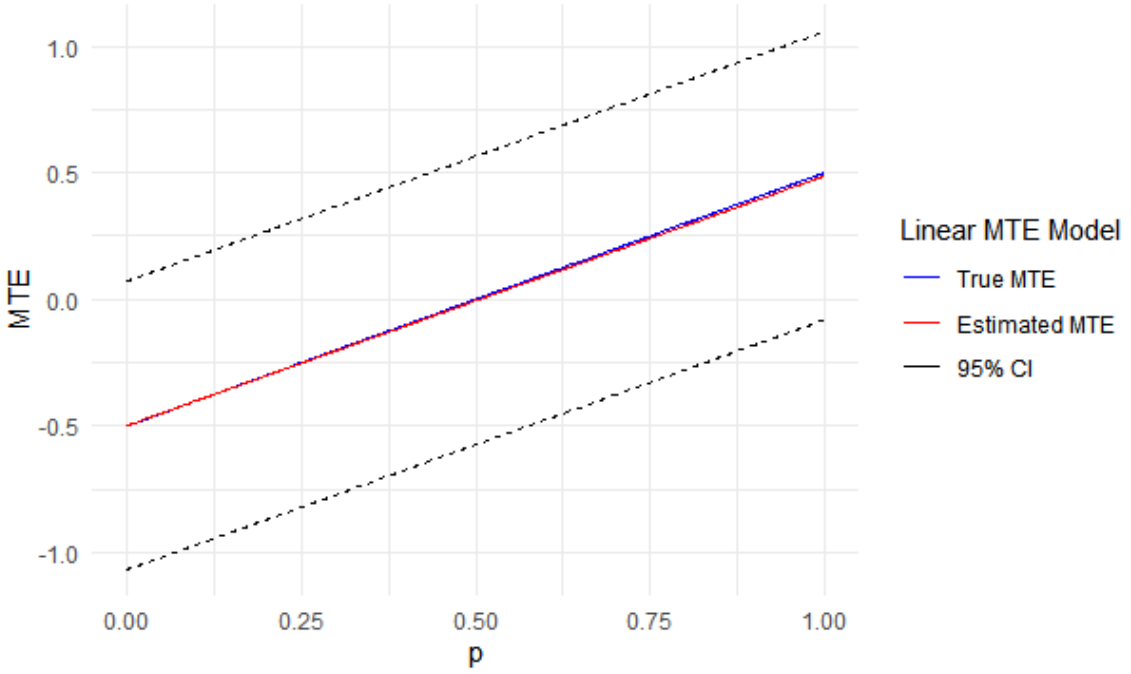
The estimator successfully recovers the MTE curve. Bias becomes greater in magnitude for larger values of p, which is expected as common support loses is strength for extreme values, but is never greater than the absolute value of 0.013, as shown in Figure 2.

In the LATE simulations, we focus on four empirically relevant situations. Table 1 displays the results for the simulations in the case where both the propensity score and the

[6]We define $U_1 \sim N(0,1)$ and $U_0 \sim N(0,1)$. The linear correaltion with U_D is defined in Section 3.2.1.

[7]This specification is consistent with the model presented in Olsen (1980), we use it for the sake of simplicity.

Figure 1: MTE Curve

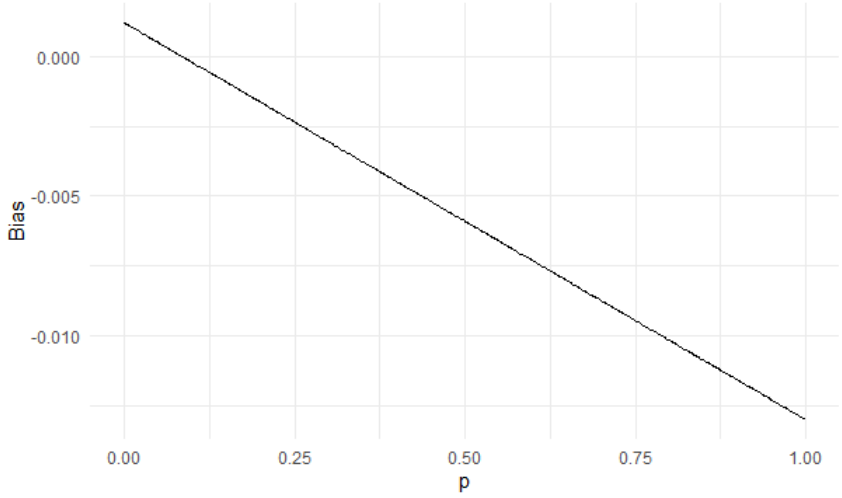


Note: This figure displays the MTE estimates based on assumptions 1-6.The MTE is evaluated at the mean value of the covariates. $P_D(G, T, X)$ is constructed through a linear probability model. The MTE estimates are based on a control function regression. The 95% precent confidence interval is computed from a bootstrap with 500 replications. The y-axis measures the value of the MTE from the DGP, whereas the x-axis measures the unobserved distaste for treatment.

control functions are correctly specified.

First, note that the TWFE fixed effect estimator is severely biased and its confidence interval has coverage probability close to zero. This result is expected, as it is well known that the TWFE regression model implicitly rules out covariate-specific trends and treatment effects (Sant'Anna and Zhao, 2020). Furthermore, the estimator is agnostic to imperfect compliance and does not account for the essential heterogeneity. The DR-DID estimator is also severely biased. It presents coverage rate of nearly 0.4, mostly due to the fact the bootstrapped confidence intervals have greater average length than the TWFE confidence interval. The DR-DID estimator accounts for covariate-specific trends but not for imperfect compliance, so the results are not unexpected. The WALD-DID is the first estimator we are considering that tries to tackle the "fuzzy" DiD design. To put it simply, it is a 2SLS strategy in which the instrument is given by the interaction of group and time variables. The WALD-DID is also severely biased in our setting. It recovers the LATE under the assumption of homogeneous treatment effects across groups. Since the assumption does not hold in our setting, the results are expected. Furthermore, the WALD-DID estimator presents the greatest average length for the confidence interval amongst the considered estimators, which

Figure 2: Bias of the CF-DID Estimator for the MTE



Note: This figure display the bias from the MTE estimates presented in Figure 1. The y-axis measures the value of the bias of the MTE estimates, whereas the x-axis measures the unobserved distaste for treatment.

is expected from the IV literature. The TC-WALD estimator presents a reasonably smaller bias, yet it still severely biased, as it relies on the assumption of homogeneous treatment effects within groups. Since our DGP allows for heterogeneous treatment effects withing groups as a function of the unobserved distaste for treatment, the bias is expected. Finally, the CF-DID and the RIPW estimators show little to no bias. In terms of efficiency, the performance of both estimators is also similar. Since we are dealing with parametric models and reasonably large simulated samples, we believe there is little difference in the performance of the estimators in terms of efficiency when both nuisance parameters are correct, which will not be necessarily true when working with non-parametric versions of the estimators.

Table 1: Control Function and Propensity Score are correct.

Estimator	Av. Bias	Med. Bias	RMSE	Cover	CIL
TWFE	-1.613	-1.614	1.615	0.000	0.208
DR-DID	1.611	1.612	1.615	0.327	2.632
WALD-DID	2.971	2.863	2.988	0.657	4.542
TC-WALD	-0.450	-0.439	0.449	0.768	1.822
CF-DID	-0.010	-0.011	0.209	0.953	1.088
RIPW	-0.012	-0.008	0.209	0.947	1.327

Note: Simulations based on 10,000 Monte Carlo experiments. TWFE is the two-way fixed effects estimator, DR-DID is the Doubly-Robust DiD estimator as proposed in Sant'anna and Zhao (2020), WALD-DID is the Wald-DiD estimator as used in Duflo (2001), TC-WALD is the Time-Corrected Wald Ratio proposed by De Chaisemartin and D'Haultefoeuille (2018), CF-DID and RIPW are our proposed estimator. "Av. Bias", "Med. Bias", "RMSE", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, 95% coverage probability, and 95% confidence interval length, respectively.

Table 2 shows the results for the simulations in which only the propensity score is correct.

We misspecify the control function by estimating a quadratic model for the MTE. In that case, the CF-DID estimator presents non-negligible bias, while the RIPW estimator still shows little to no bias.

Table 2: Only the Propensity Score is correctly specified.

Estimator	Av. Bias	Med. Bias	RMSE	Cover	CIL
TWFE	-1.613	-1.614	1.615	0.000	0.208
DR-DID	1.611	1.612	1.615	0.327	2.632
WALD-DID	2.971	2.863	2.988	0.657	4.542
TC-WALD	-0.450	-0.439	0.449	0.768	1.822
CF-DID	-0.141	-0.143	0.222	0.793	1.072
RIPW	-0.007	-0.001	0.209	0.949	1.329

Note: Simulations based on 10,000 Monte Carlo experiments. TWFE is the two-way fixed effects estimator, DR-DID is the Doubly-Robust DiD estimator as proposed in Sant'anna and Zhao (2020), WALD-DID is the Wald-DiD estimator as used in Duflo (2001), TC-WALD is the Time-Corrected Wald Ratio proposed by De Chaisemartin and D'Haultefoeuille (2018), CF-DID and RIPW are our proposed estimator. "Av. Bias", "Med. Bias", "RMSE", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, 95% coverage probability, and 95% confidence interval length, respectively.

In table 3 we present the results for the case in which only the control function is correctly specified. We estimate a linear model for the MTE, but using an incorrect propensity score to account for the unobserved heterogeneity. The CF-DID estimators exhibits non-negligible bias with magnitude similar to the one displayed in Table 2. The RIPW estimator remains unbiased.

Table 3: Only the Control Function is correctly specified.

Estimator	Av. Bias	Med. Bias	RMSE	Cover	CIL
TWFE	-1.613	-1.614	1.615	0.000	0.208
DR-DID	1.611	1.612	1.615	0.327	2.632
WALD-DID	2.971	2.863	2.988	0.657	4.542
TC-WALD	-0.450	-0.439	0.449	0.768	1.822
CF-DID	-0.160	-0.173	0.304	0.664	1.488
RIPW	-0.012	-0.013	0.205	0.941	1.617

Note: Simulations based on 10,000 Monte Carlo experiments. TWFE is the two-way fixed effects estimator, DR-DID is the Doubly-Robust DiD estimator as proposed in Sant'anna and Zhao (2020), WALD-DID is the Wald-DiD estimator as used in Duflo (2001), TC-WALD is the Time-Corrected Wald Ratio proposed by De Chaisemartin and D'Haultefoeuille (2018), CF-DID and RIPW are our proposed estimator. "Av. Bias", "Med. Bias", "RMSE", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, 95% coverage probability, and 95% confidence interval length, respectively.

When both nuisance parameters are incorrect, all estimators have non-negligible bias and all inference procedures are misleading. In this scenario, our estimators present the smaller biases when compared to the rest, and the CF-DID seems to perform the best in this case.

The simulations assert the desirable double-robustness property of the RIPW estimator. In terms of efficiency however, our estimators present a similar performance. Moreover, the

simulations show that we can consistently estimate the MTE curve when the nuisance parameters are correctly specified. There is no doubly-robust procedure for the estimation of the MTE, as it relies on the assumption that we can model the unobservable gains from treatment. Therefore, correctly specifying the control function is paramount for the identification of marginal treatment effects. In order to do so, researchers must rely on economic theory to model their control functions.

Table 4: Control Function and Propensity Score are incorrect.

Estimator	Av. Bias	Med. Bias	RMSE	Cover	CIL
TWFE	-1.613	-1.614	1.615	0.000	0.208
DR-DID	1.611	1.612	1.615	0.327	2.632
WALD-DID	2.971	2.863	2.988	0.657	4.542
TC-WALD	-0.450	-0.439	0.449	0.768	1.822
CF-DID	-0.137	-0.141	0.352	0.699	1.451
RIPW	0.174	0.266	0.334	0.799	1.652

Note: Simulations based on 10,000 Monte Carlo experiments. TWFE is the two-way fixed effects estimator, DR-DID is the Doubly-Robust DiD estimator as proposed in Sant'anna and Zhao (2020), WALD-DID is the Wald-DiD estimator as used in Duflo (2001), TC-WALD is the Time-Corrected Wald Ratio proposed by De Chaisemartin and D'Haultefoeuille (2018), CF-DID and RIPW are our proposed estimator. "Av. Bias", "Med. Bias", "RMSE", "Cover" and "CIL', stand for the average simulated bias, median simulated bias, simulated root mean-squared errors, 95% coverage probability, and 95% confidence interval length, respectively.

7 Empirical illustration: Returns to schooling in Indonesia

We illustrate the use of our estimators by revisiting Duflo (2001), which analyzes the returns to schooling in Indonesia by exploiting a major government school construction program as a natural experiment in a DiD design.

In 1973 the Indonesian government launched the INPRES program, a major primary school construction program. In the setting, year of birth plays the role of time. Men born between 1957 and 1962 are defined as cohort 0, as they should have finished primary school by the time the program was launched. Men born between 1968 and 1972 are defined as cohort 1 since they had the age to enroll in primary education after the program. Treatment and control groups are defined according to the number of primary schools per capita at each district. The author regresses the number of primary schools constructed on the number of school-age children in each district and define treatment districts as those with a positive residual in that regression.

The outcome of interest is the logarithm of wages and the treatment variable is the individual's years of schooling. Since our method is suited for binary treatments, we recategorize

the treatment variable in order to take value 1 for individuals that attended primary school and 0 for individuals that did not. Individuals with a greater educational attainment were excluded from the sample. We include a district's school enrollment rate in 1971 and the presence of a water and sanitation program as covariates in the regressions.

We estimate the LATE using the parametric control function and the RIPW, and compare the results obtained using the WALD-DID and the TC-WALD. We also use the estimates from the parametric control function to derive a linear MTE model.

7.1 LATE

Table 5 displays the value of the estimates obtained using the WALD-DID, the TC-WALD, the CF-DID and the RIPW. Standard errors for the first three estimators were obtained using 200 bootstrap replications of a bootstrap clustered at the district level, whereas the standard error for the RIPW was obtained trough the plug-in estimator from Section 5.

Table 5: Returns to primary school using the groups from Duflo (2001)

	WALD-DID	TC-WALD	CF-DID	RIPW
Returns to education	0.345	0.317	0.202	0.186
95% CI	[0.075, 1.432]	[0.062, 1.363]	[0.111, 0.367]	[0.116, 0 354]

Note: Sample size: 9113 observations. Standard errors account for clustering at the district level.

All point estimates are positive and statistically significant at conventional levels. The WALD-DID and the TC-WALD present larger point estimates than the ones obtained through the CF-DID and the RIPW, but not at a statistically significant level.

The RIPW estimator is the most flexible among the ones considered, as it relies on the least restrictive set of assumptions. Thus, we take it as the most credible estimate. Therefore, the results suggest that there is nearly a 19% increase in wages associated to primary school attendance.

7.2 Linear MTE model

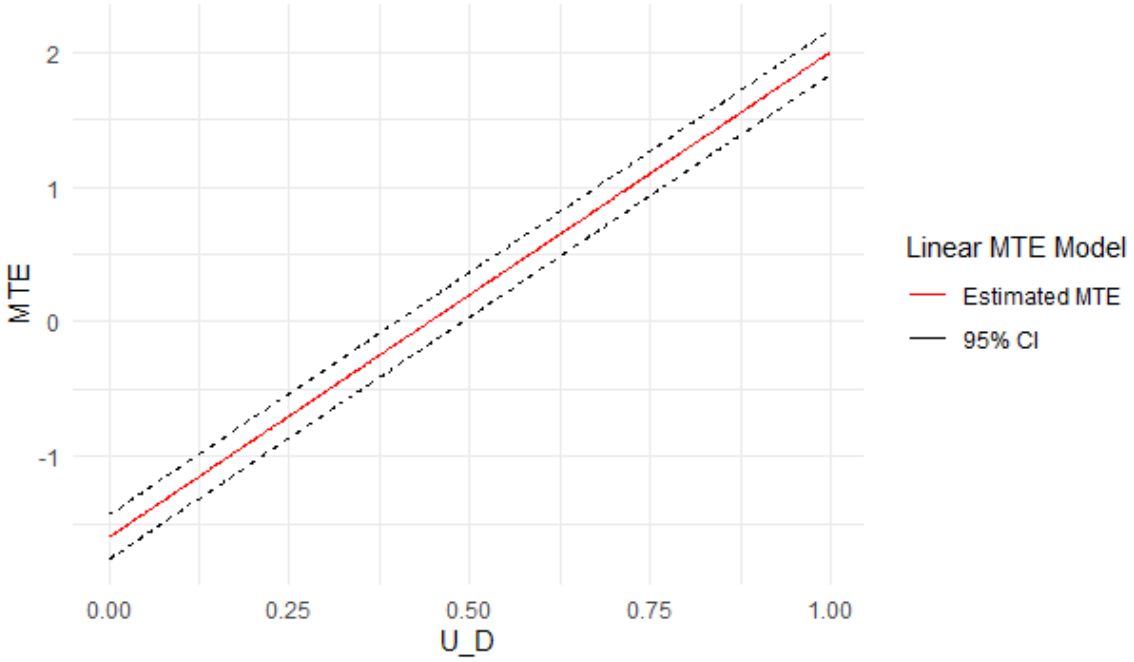
We define a linear MTE model following the specification outlined in Section 3.2.1. Figure 3 depicts the MTE curve evaluated at the mean values of the covariates in the sample. The figure reveals substantial heterogeneity in the effects of primary school attendance on future wages.

The MTE curve has an upward sloping shape, which indicates a pattern of adverse selection on gains. While individuals with low resistance to primary school attendance (i.e., low values of $U_D = p$) actually exhibit negative effects on future earnings, while there is

a substantive increase in wages for individuals with a high resistance to school attendance (high U_D). This pattern of reverse selection is also present in Cornelissen et al. (2018) which analyzes the returns of early child care attendance in Germany.

The results displayed in Table 5 exhibit the positive returns of primary school attendance for the treatment group switchers, but nevertheless mask substantial heterogeneity in the returns.

Figure 3: MTE curve for the returns to schooling



Note: Figure 3 displays the MTE curve of the effect of primary school attendance on future wages estimated by the CF-DID method using the data from Duflo (2001). The 95% confidence interval is based on bootstrapped standard errors.

The findings have important policy implications. First, they suggest that policies that successfully attract children with high resistance not currently enrolled in primary school may yield large returns. Thus, it follows that targeted programs can be more cost effective than universal primary school enrollment programs.

8 Conclusion

The Marginal Treatment Effect provides a choice-theoretic foundation that unifies the econometric literature on causal inference. Not only it has a clear economic interpretation, it also summarizes all other conventional treatment parameters.

In this paper we show how the difference-in-differences design can be used to identify

the MTE under a functional structure that allows for treatment heterogeneity based on the unobservable characteristics that drive selection into treatment using a control function estimation approach.

We propose two different estimators, that rely on different assumptions regarding the potential outcomes. First, we propose a semiparametric estimator that is valid if we assume that the potential outcomes are additively separable in a component that depends on covariates X and a component that is function of of the essential heterogeneity U_D. Second, we propose a parametric estimator that is valid if additive separability does not hold, but requires that covariates only affect potential outcomes through its unobservable component and that the unobserved heterogeneity has a known polynomial form. Thus, we provide for the applied researchers different estimators, from which he can choose the one that is most adequate for the particular empirical problem.

We derive the large sample properties of our parametric MTE estimator and illustrate the desirable finite sample properties of the estimator via a simulation exercise. Asymptotic properties for the semiparametric estimator are being derived still. For now we recommend to use boostrap for inference, which is standard for semiparametric MTE estimators in IV settings.

We provide simple tests to assess the validity of the identification assumptions in the setting, and to test the external validity of LATE estimates in a linear MTE model.

We also show that unit-specific reweighting of our parametric estimator's objective function improves the robustness of the resulting estimator for the LATE, providing thus a less demanding procedure for applied researchers to estimate the LATE in "fuzzy" DiD settings.

The Monte Carlo simulation studies assert the desirable finite-sample properties of the parametric control function DiD estimator and the RIPW estimator. Furthermore, results show that the conventional DiD established in the literature fail to recover consistent estimates for treatment effects in the presence of essential heterogeneity.

The empirical illustration illustrates the economic insights that can come from the estimation of the MTE curve. Applying the parametric control function DiD estimator in the setting of Duflo (2001) we find substantial heterogeneity in the effects of primary school attendance on future earnings, and a pattern of reverse selection on gains which could not be recovered by any other conventional policy evaluation parameter.

This work is only a first effort in the direction of building a complete theory for the MTE framework in DiD designs. So far we have only considered a parametric procedure for the 2×2 setting. A natural direction for the advance of this agenda is to propose identification results for the MTE in DiD settings with staggered adoption and dynamic treatment effects, following the recent advances in the literature (Sant'Anna and Callaway, 2021; Goodman-

Bacon, 2021; Borusyak et al., 2021).

Another necessary step in future research is to consider 2×2 settings and settings with multiple periods where the treatment variable is either multi-valued or continuous (Callaway et al., 2021).

References

Angrist, J. and Fernandez-Val, I. (2013). Extrapolate-ing: External validity and overidentification in the late framework. volume 3 of *Advances in Economics and Econometrics: Tenth World Congress*, pages 401–436. Cambridge University Press.

Arkhangelsky, D., Imbens, G. W., Lei, L., and Luo, X. (2021). Double-robust two-way-fixed-effects regression for panel data. *Working Paper*.

Björklund, A. and Moffitt, R. (1987). The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics*, 69(1):42–49.

Bonhomme, S. and Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economic and Statistics*, 93(2):479–494.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *Working Paper*.

Brinch, C., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4):985–1037.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2021). Difference-in-differences with a continuous treatment.

Carneiro, P., Heckman, J., and Vytlacil, E. (2011). Estimating Marginal Returns to Education. *American Economic Review*, 101(1):2754–2781.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, 41(1):47–60.

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409.

De Chaisemartin, C. and D'Haultefoeuille, X. (2018). Fuzzy Difference-in-Differences. *Review of Economic Studies*, 85(1):995–1028.

De Chaisemartin, C. and D'Haultefoeuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(1):795–811.

Field, E. (2007). Entitled to work: Urban property rights and labor supply in peru. *The Quarterly Journal of Economics*, 122(4):1561–1602.

Gentzkow, M., Shapiro, J. M., and Sinkinson, M. (2011). The effect of newspaper entry and exit on electoral politics. *American Economic Review*, 101(7):2980–3018.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Graham, B. S. and Pinto, C. C. d. X. (2021). Semiparametrically efficient estimation of the average linear regression function. *Journal of Econometrics*, 1(226):115–138.

Hansen, L. P. (1982). A more credible approach to parallel trends. *Econometrica*, 50(4):1029–1054.

Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 4(64):605–654.

Heckman, J. and Robb, R. (1985). Identification of Causal Effects Using Instrumental Variables. *Journal of Econometrics*, 30(1):239–267.

Heckman, J. and Vytlacil, E. (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4730–4734.

Heckman, J. and Vytlacil, E. (2001). Policy-Relevant Treatment Effects. *American Economic Review*, 91(2):107–111.

Heckman, J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738.

Heckman, J. and Vytlacil, E. (2007). Chapter 71 econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. volume 6 of *Handbook of Econometrics*, pages 4875–5143. Elsevier.

Heckman, J. J., Schmierer, D., and Urzua, S. (2010). Testing the correlated random coefficient model. *Journal of Econometrics*, 2(158):177–203.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier.

Olsen, R. J. (1980). A least square correction for selectivity bias. *Econometrica*, 48(7):1815–1820.

Rambachan, A. and Roth, J. (2022). A more credible approach to parallel trends. *Working Paper*.

Robins, J., Ronitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Roth, J. (2022). Pre-test with caution: Event-study estimates after testing for parallel trends. *American Economic Review (Forthcoming)*.

Rothe, C. and Firpo, S. (2018). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Journal of the American Statistical Association*, pages 1–40.

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

Sant'Anna, P. and Zhao, J. (2020). Doubly Robust Difference-in-Differences Estimators. *Journal of Econometrics*, 219(1):101–122.

Appendix: Main Proofs

Theorem 1

Proof. Standard errors of the two-stage estimator need to be adjusted for the fact that the we use the estimated propensity score from the first stage as a regressor in the second stage.

The asymptotic distribution of the second-stage estimates can be obtained by interpreting our two-stage procedur as a joint GMM estimator (Hansen, 1982).

Define W_{1i} as the vector of regressors from the selection into treatment equation and θ_1 as the vector of parameters to be estimated[8]. Furthermore, define W_{2i} as the vector of regressors used in the second stage and θ_2 as the vector of parameters associated to them.

The CF-DID estimator solves the population analogue of

[8]In the linear MTE model used in the simulations, $W_{1i} = (X_i, G_i, T_i, G_i \times T_i)$ and $\theta_1 = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$

$$E[f(\theta_1, \theta_2; W_{1i}, W_{2i})] = E\begin{bmatrix} W_{1i}(D_i - \theta_1^T W_{1i}) \\ W_{2i}(Y_i - \theta_2^T W_{2i}) \end{bmatrix} = 0$$

By Theorem 6.1 of Newey and McFadden (1994), and under standard regularity conditions,

$$\sqrt{n}(\widehat{\theta}_2 - \theta_2) \sim N(0, V_{CF})$$

where V_{CF} is the last element of

$$E\left[\frac{\partial f(\theta_1, \theta_2; W_{1i}, W_{2i})}{\partial(\theta_1, \theta_2)}\right]^{-1} E[f(\theta_1, \theta_2; W_{1i}, W_{2i}) f(\theta_1, \theta_2; W_{1i}, W_{2i})^T] E\left[\frac{\partial f(\theta_1, \theta_2; W_{1i}, W_{2i})}{\partial(\theta_1, \theta_2)}\right]^{-1T}$$

We construct our MTE estimates as a function of the parameters estimated in the second-stage of the procedure. For instance, in the case of our linear MTE model, the MTE estimates are given by $\Delta^{MTE}(\widehat{\theta}_2) = \widehat{\tau} + (\widehat{\pi}_1 - \widehat{\pi}_2)x(p - \frac{1}{2})$. Thus, it follows from the Delta Method that

$$\sqrt{n}(\Delta^{MTE}(\widehat{\theta}_2) - \Delta MTE(\theta_2)) \sim N(0, V_{MTE})$$

where $V_{MTE} = (\nabla_{\theta_2}\Delta^{MTE}(\theta_2))^T V_{CF}(\nabla_{\theta_2}\Delta^{MTE}(\theta_2))$, in which ∇_{θ_2} represents the gradient of the MTE function with respect to the parameters in θ_2. Consistency of the bootstrap follows directly from the consistency of the bootstrap for GMM estimators.

■

Theorem 2

We conduct the proof using the notation previously established in Section 5

Proof. Given an estimate $\widehat{m}_{ij}$ and an estimate $\widehat{P}_D(G, T, X)$ we consider the following RIPW estimator:

$$\begin{aligned} \widehat{\tau}(\pi) &= \underset{\mu,\tau}{argmin} \sum_{i=1}^{n} ((Y_i - \widehat{m}_{ij}) - \mu - \tau D_i)^2 \Theta_i \\ &= \underset{\mu,\tau}{argmin} \sum_{i=1}^{n} (\widetilde{Y}_i - \mu - \tau D_i)^2 \Theta_i \end{aligned}$$

The first order conditions with respect to μ and τ are respectively

$$\sum_{i=1}^{n} \Theta_i (\widetilde{Y}_i - \mu - \tau D_i) = 0$$

and

$$\sum_{i=1}^{n} \Theta_i D_i (\widetilde{Y}_i - \mu - \tau D_i) = 0$$

From the first order conditions we obtain the optimal $\widehat{\mu}$ as function of τ:

$$\widehat{\mu}(\tau) = \Gamma_{\Theta}^{-1}(\Gamma_Y - \tau \Gamma_D)$$

Substitute $\widehat{\mu}(\tau)$ into the first order condition with relation to τ and after some tedious algebraic manipulation we obtain

$$\widehat{\tau}(\pi) = \frac{\Gamma_{DY} - \Gamma_{\Theta}^{-1}(\Gamma_D \Gamma_Y)}{\Gamma_{DD} - \Gamma_{\Theta}^{-1}\Gamma_D^2}$$

■

Asymptotic Linear Expansion of the RIPW Estimator

We conduct the expansion using the notation introduced in Section 5. We begin the asymptotic expansion by noting that

$$B(\widehat{\tau} - \tau) = A - \tau B$$

By Lemma A.2 of Arkhangelsky et al. (2021), we have

$$|(\Gamma_{DY} - E[\Gamma_{DY}])(\Gamma_\Theta - E[\Gamma_\Theta])| + |(\Gamma_D) - E[\Gamma_D](\Gamma_Y - E[\Gamma_Y])| = O_p(n^{-q})$$

where $q \in (0, 1]$ denotes measures the strength of the correlation between unit's selection into treatment.

Let

$$\mathbf{V}_{i1} = \Theta_i \left\{ E[\Gamma_{DY}] - E[\Gamma_Y] D_i + E[\Gamma_\Theta] D_i \widetilde{Y}_i - E[\Gamma_D] \widetilde{Y}_i \right\}$$

Then we can write A as

$$A = E[\Gamma_{DY}]E[\Gamma_\Theta] - E[\Gamma_D]E[\Gamma_Y] + \frac{1}{n}\sum_{i=1}^{n} (\mathbf{V}_{i1} - E[\mathbf{V}_{i1}]) + O_p(n^{-q})$$

Similarly, we define

$$\mathbf{V}_{i2} = \Theta_i \left\{ E[\Gamma_{DD}] - E[\Gamma_D] D_i + E[\Gamma_\Theta] D_i^2 - E[\Gamma_\Theta] D_i \right\}$$

Since $\mathbf{V}_i = \mathbf{V}_{i1} - \tau \mathbf{V}_{i2}$, it follows that

$$B(\widehat{\tau}-\tau) = N_* + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{V}_i - E[\mathbf{V}_i]) + O_p(n^{-q})$$

Our plug-in estimator for the variance of the RIPW estimator and the inference procedure follow from Theorem A.5 of Arkhangelsky et al. (2021).

Theorem 3

Proof. Theorem A.2 of Arkhangelsky et al. (2021) show that $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{V}_i - E[\mathbf{V}_i]) = o_p(1)$. Thus, it follows that the asymptotic limit of $B(\widehat{\tau}-\tau)$ is equal to N_*. For consistency of the estimator , it remains to show that $N_* = o_p(1)$.

For the sake of simplicity in this exposition, we assume that $\tau = 0$, without loss of generality. Then

$$N_* = E[\Gamma_{DY}]E[\Gamma_\Theta] - E[\Gamma_D]E[\Gamma_Y]$$

We begin by noting that $\widetilde{Y}_i = \widetilde{Y}_i(D_i) + \tau D_i$. By Assumptions 1-6,

$$\begin{aligned} E[\Gamma_{DY}] &= \frac{1}{n}\sum_{i=1}^{n}[\Theta_i D_i \widetilde{Y}_i] = \frac{1}{n}\sum_{i=1}^{n}[\Theta_i D_i(\widetilde{Y}_i(D_i) + \tau D_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[\Theta_i D_i]E[\widetilde{Y}_i(D_i)] + \frac{1}{n}\sum_{i=1}^{n} E[\Theta_i D_i^2]\tau \end{aligned}$$

Analogously,

$$E[\Gamma_Y] = \frac{1}{n}\sum_{i=1}^{n} E[\Theta_i]E[\widetilde{Y}_i(D_i)] + \frac{1}{n}E[\Theta_i D_i]\tau$$

Thus, it follows that

$$\begin{aligned} N_* = &\frac{1}{n}\sum_{i=1}^{n}\left\{E[\Theta_i D_i]E[\Gamma_\Theta] - E[\Theta_i]E[\Gamma_D]\right\}E[\widetilde{Y}_i(D_i)] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left\{E[\Theta_i D_i^2]E[\Gamma_\Theta] - E[\Gamma_D]E[\Theta_i D_i]\right\}\tau \end{aligned}$$

1) Suppose that $\widehat{m}_{ij}(G_i, T_i, X_i, p_i) = m_{ij}(G, T, X, p)$ a.s.. Since we assume $\tau = 0$, it follows that

$$N_* = \frac{1}{n}\sum_{i=1}^{n}\left\{E[\Theta_i D_i]E[\Gamma_\Theta] - E[\Theta_i]E[\Gamma_D]\right\}E[\widetilde{Y}_i(D_i)]$$

By Assumptions 1-6, if $\widehat{m}_{ij}(G_i, T_i, X_i, p_i) = m_{ij}(G, T, X, p)$, then $E[\widetilde{Y}_i(D_i)] = 0$. Therefore, it is obvious from the expression above that $N_* = 0$.

2) Sppose now that $\widehat{P}_D(G_i, T_i, X_i) = P_D(G, T, X)$a.s.. Then for any function $f(.)$,

$$E[\Theta_i f(D_i)] = \sum_{D\in\{0,1\}} \frac{\pi(D;X)}{P_D(G,T,X)} P_D(G,T,X) f(D_i) = E_{D\sim\pi}[f(D)]$$

where $E_{D\sim\pi}[(.)]$ denotes to the expectation of a given variable reshaped by the density $\pi(D;X)$.

Therefore, we have that

$$E[\Theta_i D_i] = E_{D\sim\pi}[D] = E[\Gamma_D],\ E[\Theta_i] = E[\Gamma_\Theta] = 1$$

and

$$E[\Theta_i D_i^2] = E_{D\sim\pi}[D^2],\ E[\Theta_i D_i] = E_{D\sim\pi}[D]$$

Consequently, we obtain

$$E[\Theta_i D_i]E[\Theta_i] - E[\Theta_i]E[\Gamma_D] = E_{D\sim\pi}[D] - E_{D\sim\pi}[D] = 0$$

and

$$E[\Theta_i D_i^2]E[\Gamma_\Theta] - E[\Gamma_D]E[\Theta_i D_i] = E_{D\sim\pi}[(D - E_{D\sim\pi}[D])D]$$

As a result,

$$N_* = \frac{1}{n}\sum_{i=1}^{n} E_{D\sim\pi}[(D - E_{D\sim\pi}[D])D]\tau = 0$$

■