Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period[*]

Clément de Chaisemartin[†] Xavier D'Haultfœuille[‡] Félix Pasquier[§]
Gonzalo Vazquez-Bare[¶]

First version: January 18, 2022. This version: July 1, 2023

Abstract

We propose new difference-in-difference (DID) estimators for treatments continuously distributed at every time period, as is often the case of trade tariffs, or temperatures. We start by assuming that the data only has two time periods. We also assume that from period one to two, the treatment of some units, the switchers, changes, while the treatment of other units, the stayers, does not change. Then, our estimators compare the outcome evolution of switchers and stayers with the same value of the treatment at period one. Our estimators only rely on parallel trends assumptions, unlike commonly used two-way fixed effects regressions that also rely on homogeneous treatment effect assumptions. Comparing switchers and stayers with the same period-one treatment is important: unconditional comparisons of switchers and stayers implicitly assume constant treatment effects over time. With a continuous treatment, switchers cannot be matched to stayers with exactly the same period-one treatment, but comparisons of switchers and stayers with the same period-one treatment can still be achieved by non-parametric regression, or by propensity-score reweighting. We extend our results to applications with no stayers, more than two time periods, and where the treatment may have dynamic effects.

Keywords: differences-in-differences, continuous treatment, two-way fixed effects regressions, heterogeneous treatment effects, panel data, policy evaluation.

JEL Codes: C21, C23

[*] We are very grateful to Alberto Abadie, Federico Bugni, Ivan Canay, Matias Cattaneo, Daniel Herrera, Brian Jacob, Joel Horowitz, Thierry Mayer, Isabelle Méjean, Jean-Marc Robin, Jeffrey Wooldridge, and seminar participants at Bank of Portugal, IFAU, the International Panel Data Conference, Michigan, Northwestern, Reading, Sciences Po, the Stockholm School of Economics, Université Paris Dauphine, and York University for their helpful comments. Clément de Chaisemartin was funded by the European Union (ERC, REALLYCREDIBLE,GA N°101043899). Xavier D'Haultfœuille gratefully acknowledges financial support from the research grants Otelo (ANR-17-CE26-0015-041).

[†] Sciences Po Paris, clement.dechaisemartin@sciencespo.fr

[‡] CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr.

[§] CREST-ENSAE, felix.pasquier@ensae.fr.

[¶] University of California, Santa Barbara, gvazquez@econ.ucsb.edu.

1 Introduction

A popular method to estimate the effect of a treatment on an outcome is to compare over time units experiencing different evolutions of their exposure to treatment. In practice, this idea is implemented by estimating regressions that control for unit and time fixed effects. de Chaisemartin and D'Haultfœuille (2022) find that 26 of the 100 most cited papers published by the AER from 2015 to 2019 have used a two-way fixed effects (TWFE) regression to estimate the effect of a treatment on an outcome. de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), and Borusyak et al. (2021) have shown that TWFE regressions are not robust to heterogeneous effects: under a parallel trends assumption, those regressions may estimate a weighted sum of treatment effects across periods and units, with some negative weights. Owing to the negative weights, the treatment coefficient in TWFE regressions could be, say, negative, even if the treatment effect is positive for every unit × period. Importantly, the result in de Chaisemartin and D'Haultfœuille (2020) applies to binary, discrete, and continuous treatments.

Several alternative difference-in-difference (DID) estimators robust to heterogeneous effects have been recently proposed (see Table 2 of de Chaisemartin and d'Haultfoeuille, 2023, for a review of the estimators available to practitioners, depending on their research design). Some of them apply to designs with binary treatments (see Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2021). Others apply to designs with binary or discrete treatments (see de Chaisemartin and D'Haultfœuille, 2020, 2022). Finally, other estimators apply to designs where all units start with a treatment equal to 0, and then get treated with heterogeneous, potentially continuously distributed treatment intensities (see de Chaisemartin and D'Haultfœuille, 2022; Callaway et al., 2021). This last set of papers is most closely related to ours, but it does not consider the case where the treatment is continuously distributed at every period. The goal of this paper is to complement the literature, by proposing heterogeneity-robust DID estimators for such treatments. This extension is important: TWFE regressions have often been used to estimate the effect of treatments continuously distributed at every time period, such as trade tariffs (see Fajgelbaum et al., 2020) or precipitations (see Deschênes and Greenstone, 2007).

We assume that we have a panel data set, whose units could be geographical locations such as states or counties. We start by considering the case where the panel has two time periods. We assume that from period one to two, the treatment of some units, hereafter referred to as the switchers, changes. On the other hand, the treatment of other units, hereafter referred to as the stayers, does not change. Our first target parameter is the average, across all switchers, of the effect of moving their treatment from its period-one to its period-two value scaled by the difference between these two values. In other words, our target parameter is the average slope of switchers' potential outcome function, from their period-one to their period-two treatment, hereafter referred to as the Average Of Switchers' Slopes (AOSS). The AOSS can be interpreted as an average marginal effect of the treatment. We propose a regression-based estimator of the AOSS. First, one runs a (potentially non-parametric) regression of the outcome evolution on the period-one treatment, among stayers. Second, one uses that regression to predict switchers' counterfactual outcome evolution, had their treatment not changed. Third, one subtracts each switcher's counterfactual evolution to its actual evolution, and divides this DID by the switcher's treatment evolution, thus yielding an estimator of its period-two potential-outcome function's slope, between its period-one and its period-two treatment. Finally, one averages those estimators across all switchers.

When some switchers experience a small treatment change from period one to two, estimators of the AOSS may be noisy, owing to the small denominator of those switchers' slope estimator (see Graham and Powell, 2012). Accordingly, we also consider a second causal effect. This effect is a weighted average, across all switchers, of the slope of each switcher's potential outcome function from its period-one to its period-two treatment, where switchers receive a weight proportional to the absolute value of their treatment change from period one to two. Hereafter, we refer to this causal effect as the Weighted Average Of Switchers' Slopes (WAOSS). On top of avoiding the small-denominator problem, another advantage of this parameter is that it can be estimated by a regression-based estimator and by a propensity-score based estimator, thus implying that it can also be estimated by a doubly-robust estimator. Beyond those statistical advantages, we show that if treatments change because of the introduction of a policy, the WAOSS is the key quantity to compute in a cost-benefit analysis of the policy.

Overall, our key idea is to compare switchers and stayers with the same value of the treatment at period one. Comparing switchers and stayers with the same period-one treatment is important: unconditional comparisons of switchers and stayers implicitly assume constant treatment effects over time. With a continuous treatment, switchers cannot be matched to stayers with exactly the same period-one treatment, but comparisons of switchers and stayers with the same period-one treatment can still be achieved by a non-parametric regression, as implemented in our estimator of the AOSS and in our first estimator of the WAOSS, or by propensity score reweighting, as implemented in our second estimator of the WAOSS.

The estimators proposed so far require that there be some stayers, whose treatment does not change from period one to two. This assumption is likely to be met when the treatment is say, trade tariffs: tariffs' reforms rarely apply to all products, so it is likely that tariffs of at least some products stay constant over time. On the other hand, this assumption is unlikely to be met when the treatment is say, precipitations: for instance, US counties never experience the exact same amount of precipitations over two consecutive years. We show that our identification results can be extended to the case where there are no stayers, provided there are quasi-stayers, meaning units whose treatment barely changes from period one to two. This assumption is likely to hold when the treatment is, say, precipitations: for every pair of consecutive years, there are probably some US counties whose precipitations almost do not change.

Finally, we extend our results to applications with more than two time periods, and where the treatment may have dynamic effects. We also show how to test for pre-trends, and discuss how our estimators can be applied to discrete treatments taking a large number of values.

Some of our estimators can be computed by the `did_multiplegt` Stata and R packages, while other estimators can be computed by the `did_multiplegt_dyn` Stata package. Below, we give the syntax one should use to compute them.

Related Literature

As mentioned above, our paper is related to the recent literature on heterogeneity-robust DID estimators. The two most closely related papers are de Chaisemartin and D'Haultfœuille (2022) and Callaway et al. (2021), who also propose DID estimators of the effect of a continuous treatment. Here, we assume that the treatment is continuously distributed at every period, while they rule out that possibility. For instance, Callaway et al. (2021) assume that all units have a treatment equal to zero in period one. de Chaisemartin and D'Haultfœuille (2022) consider more general designs, but they still do not allow units' period-one's treatments to be continuously distributed. Accordingly, our paper does not overlap, with and complements those two papers.

Our paper builds upon several previous papers in the DID literature. First, it is closely related to the pioneering work of Graham and Powell (2012), who also propose DID estimators of the AOSS when the treatment is continuously distributed at every time period. They also allow for time-varying treatment effects, but rely on a linear treatment effect assumption (see their Equation (1)) and assume that units experience the same evolution of their treatment effect over time, a parallel-trends-on-treatment-effects assumption (see their Assumption 1.1(i) and (iii)). Our estimators, on the other hand, do not place any restriction on treatment effects.

The idea to compare switchers and stayers with the same baseline treatment also appears in de Chaisemartin and D'Haultfœuille (2018), de Chaisemartin and D'Haultfœuille (2020), and de Chaisemartin and D'Haultfœuille (2022), who had used that idea to form DID estimators of the effect of a binary or discrete treatment. With a non-continuous treatment, there will often be switchers and stayers with the exact same baseline treatment. With a continuous treatment, the sample will not contain switchers and stayers with the exact same baseline treatment, so this paper's contribution is to use non-parametric regression or propensity-score reweighting to compare switchers and stayers "with the exact same baseline treatment".

Finally, D'Haultfoeuille et al. (2021) also consider a DID-like estimator of the effect of a continuous treatment, but their estimator relies on a common change assumption akin to that in Athey and Imbens (2006) rather than on a parallel trends assumption. Our first identification result is also related to Hoderlein and White (2012), who show how to identify the average marginal effect of a continuous treatment with panel data. The main difference is that they rule out any systematic effect of time on the outcome.

2 Set-up and assumptions

A representative unit is drawn from an infinite super population, and observed at two time periods. This unit could be an individual or a firm, but it could also be a geographical unit, like a county or a region.[1] All expectations below are taken with respect to the distribution of variables in the super population. We are interested in the effect of a continuous and scalar treatment variable on that unit's outcome. Let D_1 (resp. D_2) denote the unit's treatment at period 1 (resp. 2), and let $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) be the set of values D_1 (resp. D_2) can take, i.e. its support. For any $d \in \mathcal{D}_1 \cup \mathcal{D}_2$, let $Y_1(d)$ and $Y_2(d)$ respectively denote the unit's potential outcomes at periods 1 and 2 with treatment d.[2] Finally, let Y_1 and Y_2 denote their observed outcomes at periods 1 and 2. Let $S = 1\{D_2 \neq D_1\}$ be an indicator equal to 1 if the unit's treatment changes from period one to two, i.e. if they are a switcher.

In what follows, all equalities and inequalities involving random variables are required to hold almost surely. For any random variables observed at the two time periods (X_1, X_2), let $\Delta X = X_2 - X_1$ denote the change of X from period 1 to 2.

We make the following assumptions.

Assumption 1 *(Parallel trends) For all $d \in \mathcal{D}_1$, $E(\Delta Y(d)|D_1 = d, D_2) = E(\Delta Y(d)|D_1 = d)$.*

Assumption 1 is a parallel trends assumption. It requires that $\Delta Y(d)$ be mean independent of D_2, conditonal on $D_1 = d$. It has one key implication we leverage for identification: the counterfactual outcome evolution switchers would have experienced if their treatment had not changed is identified by the outcome evolution of stayers with the same period-one treatment:

$$E(\Delta Y(d)|D_1 = d, S = 1) = E(\Delta Y(d)|D_1 = d, S = 0). \tag{1}$$

Accordingly, the DID estimators we propose below compare switchers and stayers with the same period-one treatment.

Instead, one could propose DID estimators comparing switchers and stayers, without conditioning on their period-one treatment. On top of Assumption 1, such estimators rest on two supplementary conditions:

[1] In that case, one may want to weight the estimation by counties' or regions' populations. Extending the estimators we propose to allow for such weighting would be straightforward.

[2] Throughout the paper, we implicitly assume that all potential outcomes have an expectation.

(i) $E(\Delta Y(d)|D_1 = d) = E(\Delta Y(d))$.

(ii) For all $(d, d') \in \mathcal{D}_1^2$, $E(\Delta Y(d)) = E(\Delta Y(d'))$.

(i) requires that all units experience the same evolution of their potential outcome with treatment d, while Assumption 1 only imposes that requirement for units with the same baseline treatment. Assumption 1 may be more plausible: units with the same period-one treatment may be more similar and more likely to be on parallel trends than units with different period-one treatments. (ii) requires that the trend affecting all potential outcomes be the same. Rearranging, (ii) is equivalent to assuming $E(Y_2(d) - Y_2(d')) = E(Y_1(d) - Y_1(d'))$: the treatment effect should be constant over time, a strong restriction on treatment effect heterogeneity. Assumption 1, on the other hand, does not impose any restriction on treatment effect heterogeneity, as it only restricts one potential outcome per unit. Overall, conditioning on D_1 has two advantages: i) it makes the parallel trends assumption more plausible, as in a DID-matching estimation strategy where parallel trends is often more plausible after conditioning on some covariates (see, e.g., Abadie, 2005); ii) it avoids assuming that treatment effects do not vary over time.

Finally, note that because it imposes parallel trends conditional on prior treatment, Assumption 1 is connected to the sequential ignorability assumption, another commonly-used identifying assumption in panel data models (see, e.g., Robins, 1986; Bojinov et al., 2021). Sequential ignorability requires that treatment be randomly assigned conditional on prior treatment and outcome, which implies parallel trends conditional on prior treatment and outcome.

Assumption 2 *(Bounded treatment, Lipschitz potential outcomes, bounded outcome trends)*

1. *$\mathcal{D}_1$ and $\mathcal{D}_2$ are bounded subsets of $\mathbb{R}$.*

2. *For all $t \in \{1, 2\}$ and for all $(d, d') \in \mathcal{D}_t^2$, there is a random variable $\overline{Y} \geq 0$ such that $|Y_t(d) - Y_t(d')| \leq \overline{Y}|d - d'|$, with $\sup_{(d_1,d_2)\in Supp(D_1,D_2)} E[\overline{Y}|D_1 = d_1, D_2 = d_2] < \infty$.*

Assumption 2 is a technical condition ensuring that all the expectations below are well defined. It requires that the set of values that the period-one and period-two treatments can take be bounded. It also requires that the potential outcome functions be Lipschitz (with an individual specific Lipschitz constant). This will automatically hold if $d \mapsto Y_2(d)$ is differentiable with respect to d and has a bounded derivative.

For estimation and inference, we assume we observe an iid sample with the same distribution as (Y_1, Y_2, D_1, D_2):

Assumption 3 *(iid sample) We observe $(Y_{i,1}, Y_{i,2}, D_{i,1}, D_{i,2})_{1\leq i\leq n}$, that are independent and identically distributed vectors with the same probability distribution as (Y_1, Y_2, D_1, D_2).*

Importantly, Assumption 3 allows for the possibility that Y_1 and Y_2 (resp. D_1 and D_2) are serially correlated, as is commonly assumed in DID studies (see Bertrand et al., 2004).

3 Estimating the average of switchers' potential-outcome's slopes

3.1 Target parameter

In this section, our target parameter is

$$\delta_1 := E\left(\frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}\middle| S = 1\right). \tag{2}$$

δ_1 is the average, across switchers, of the effect of moving their treatment from its period-one to its period-two value, scaled by the difference between these two values. In other words, δ_1 is the average of the slopes of switchers' potential outcome functions, between their period-one and their period-two treatments. Hereafter, δ_1 is referred to as the Average Of Switchers' Slopes (AOSS). Note that with a binary treatment such that all units are untreated at period 1 and some units get treated at period 2, the AOSS reduces to the standard average treatment effect on the treated. Thus, the AOSS generalizes that parameter to non-binary treatments and more complicated designs.

The AOSS averages effects of discrete rather than infinitesimal changes in the treatment as in Hoderlein and White (2012), for instance. But if one slightly reinforces Point 2 of Assumption 2 by supposing that $d \mapsto Y_2(d)$ is differentiable on $\mathcal{D}_1 \cup \mathcal{D}_2$, by the mean value theorem,

$$\frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1} = Y_2'\left(\tilde{D}\right)$$

for some $\tilde{D} \in (\min(D_1, D_2), \max(D_1, D_2))$. Then, the AOSS is an average marginal effect on switchers:

$$\delta_1 = E[Y_2'\left(\tilde{D}\right)|S = 1]. \tag{3}$$

The only difference with the usual average marginal effect on switchers $E[Y_2'(D_2)|S = 1]$ is that the derivative is evaluated at $\tilde{D}$ instead of D_2.

(3) implies that unlike TWFE regression coefficients, the AOSS satisfies the no-sign reversal property. If $Y_2'(d) \geq 0$ for all d, $\delta_1 \geq 0$: if increasing the treatment always increases the outcome of every switcher, the AOSS is positive.

However, the AOSS is a local effect. First, it only applies to switchers. Second, it measures the effect of changing their treatment from its period-one to its period-two value, not of other changes of their treatment. Still, the AOSS can be used to identify the effect of other treatment changes under shape restrictions on the potential outcome function. First, assume that the potential outcomes are linear: for $t \in \{1, 2\}$,

$$Y_t(d) = Y_t(0) + B_t d,$$

where B_t is a slope that may vary across units and may change over time. Then, $\delta_1 = E(B_2|S=1)$: the AOSS is equal to the average, across switchers, of the slopes of their potential outcome functions at period 2. Therefore, for all $d \neq d'$, $E(Y_2(d)-Y_2(d')|S=1) = (d-d')\delta_1$: under linearity, knowing the AOSS is sufficient to recover the ATE of any uniform treatment change among switchers. Accordingly, the AOSS can be used to evaluate other treatment changes than the one that was effectively implemented. Of course, this only holds under linearity, which may not be a plausible assumption. Assume instead that $d \mapsto Y_2(d)$ is convex. Then, for any $\epsilon > 0$,

$$E\left(Y_2(D_2+\epsilon) - Y_2(D_1)|S=1\right) \geq E\left(Y_2(D_2) - Y_2(D_1)|S=1\right) + \epsilon\delta_1.$$

$E\left(Y_2(D_2) - Y_2(D_1)|S=1\right)$ can be identified following the same steps as those we use to identify the AOSS below. Accordingly, under convexity one can use the AOSS to obtain a lower bound of the effect of changing the treatment from D_1 to a larger value than D_2. For instance, in Fajgelbaum et al. (2020), one can use this strategy to derive a lower bound of the effect of even larger tariffs' increases than those implemented by the Trump administration. Under convexity, one can also derive an upper bound of the effect of changing the treatment from D_1 to a lower value than D_2. And under concavity, one can derive an upper (resp. lower) bound of the effect of changing the treatment from D_1 to a larger (resp. lower) value than D_2.[3] Note that our results below concerning the identification and estimation of the AOSS hold even if the aforementioned linearity or convexity/concavity conditions fail. But those conditions are necessary to use the AOSS to identify or bound the effects of alternative policies.

3.2 Identification

To identify the AOSS, we use a DID estimand comparing switchers and stayers with the same period-one treatment. This requires that there be no value of the period-one treatment D_1 such that only switchers have that value, as stated formally below.

Assumption 4 *(Support condition for AOSS identification)* $P(S=1) > 0$ *and* $P(S=1|D_1) < 1$.

Note that Assumption 4 implies that $P(S=0) > 0$: while we assume that the treatments D_1 and D_2 are continuous, we also assume that the treatment is persistent, and thus ΔD has a mixed distribution with a mass point at zero.

To identify the AOSS, we also start by assuming that there are no quasi-stayers: the treatment of all switchers changes by at last c from period one to two, for some strictly positive c.

Assumption 5 *(No quasi-stayers)* $\exists c > 0$: $P(|\Delta D| > c|S=1) = 1$.

[3] See D'Haultfoeuille et al. (2021) for bounds of the same kind obtained under concavity or convexity.

We relax Assumption 5 just below.

Theorem 1 *If Assumptions 1-5 hold,*

$$\delta_1 = E\left(\frac{\Delta Y - E(\Delta Y|D_1, S=0)}{\Delta D}\middle| S=1\right).$$

Intuitively, the effect of changing switchers' treatment from its period-one to its period-two value is identified by a DID comparing their outcome evolution to that of stayers with the same period-one treatment. Then, this DID is normalized by ΔD, to recover the slope of switchers' potential outcome function, between their period-one and their period-two treatments.

If there are quasi-stayers, the AOSS is still identified. For any $\eta > 0$, let $S_\eta = 1\{|\Delta D| > \eta\}$ be an indicator equal to one for switchers whose treatment changes by at least η from period one to two.

Theorem 2 *If Assumptions 1-4 hold,*

$$\delta_1 = \lim_{\eta \downarrow 0} E\left(\frac{\Delta Y - E(\Delta Y|D_1, S=0)}{\Delta D}\middle| S_\eta=1\right).$$

If there are quasi-stayers whose treatment change is arbitrarily close to 0 (i.e. $f_{|\Delta D||S=1}(0) > 0$), the denominator of $(\Delta Y - E(\Delta Y|D_1, S=0))/\Delta D$ is very close to 0 for them. On the other hand,

$$\begin{aligned}&\Delta Y - E(\Delta Y|D_1, S=0)\\ =&Y_2(D_2) - Y_2(D_1) + \Delta Y(D_1) - E(\Delta Y(D_1)|D_1, S=0)\\ \approx&\Delta Y(D_1) - E(\Delta Y(D_1)|D_1, S=0),\end{aligned}$$

so the ratio's numerator may not be close to 0. Then, under weak conditions,

$$E\left(\left|\frac{\Delta Y - E(\Delta Y|D_1, S=0)}{\Delta D}\right| \middle| S=1\right) = +\infty.$$

Therefore, we need to trim quasi-stayers from the estimand in Theorem 1, and let the trimming go to 0 to still recover δ_1, as in Graham and Powell (2012) who consider a related estimand with some quasi-stayers. Accordingly, while the AOSS is still identified with quasi-stayers, it is irregularly identified by a limiting estimand.

3.3 Estimation and inference

With no quasi-stayers, $E\left((\Delta Y - E(\Delta Y|D_1, S=0))/\Delta D|S=1\right)$ can be estimated in three steps. First, one estimates $E(\Delta Y|D_1, S=0)$ using a non-parametric regression of ΔY_i on $D_{i,1}$ among stayers. Second, for each switcher, one computes $\widehat{E}(\Delta Y|D_1 = D_{i,1}, S = 0)$, its predicted outcome evolution given its baseline treatment, according to the non-parametric regression estimated among stayers. Third, one lets

$$\widehat{\delta}_1 := \frac{1}{n_s} \sum_{i:S_i=1} \frac{\Delta Y_i - \widehat{E}(\Delta Y|D_1 = D_{i,1}, S = 0)}{\Delta D_i},$$

where $n_s = \#\{i : S_i = 1\}$.

To estimate $E(\Delta Y|D_1, S = 0)$, we consider a series estimator based on polynomials in D_1, $(p_{k,K_n}(D_1))_{1\leq k\leq K_n}$. We make the following technical assumption.

Assumption 6 *(Conditions for asymptotic normality of AOSS estimator)*

1. *D_1 is continuously distributed on a compact interval I, with $\inf_{d\in I} f_{D_1}(d) > 0$.*
2. *$E[\Delta Y^2] < \infty$ and $d \mapsto E[\Delta Y^2|D_1 = d]$ is bounded on I.*
3. *$P(S = 1) > 0$ and $\sup_{d\in I} P(S = 1|D_1 = d) < 1$.*
4. *The functions $d \mapsto E[(1 - S)\Delta Y|D_1 = d]$, $d \mapsto E[S|D_1 = d]$ and $d \mapsto E[S/\Delta D|D_1 = d]$ are four times continuously differentiable.*
5. *The polynomials $d \mapsto p_{k,K_n}(d)$, $1 \leq k \leq K_n$, are orthonormal on I and $K_n^{12}/n \to +\infty$, $K_n^7/n \to 0$.*

Point 3 is a slight reinforcement of Assumption 4. In Point 5, $K_n^{12}/n \to \infty$ requires that K_n, the order of the polynomial in D_1 we use to approximate $E(\Delta Y|D_1, S = 0)$, goes to $+\infty$ when the sample size grows, thus ensuring that the bias of our series estimator of $E(\Delta Y|D_1, S = 0)$ tends to zero. $K_n^7/n \to 0$ ensures that K_n does not go to infinity too fast, thus preventing overfitting.

Theorem 3 *If Assumptions 1-3 and 5-6 hold,*

$$\sqrt{n}\left(\widehat{\delta}_1 - \delta_1\right) \xrightarrow{d} \mathcal{N}(0, V(\psi_1)),$$

where

$$\psi_1 := \frac{1}{E(S)} \left\{ \left(\frac{S}{\Delta D} - E\left(\frac{S}{\Delta D} \middle| D_1 \right) \frac{(1 - S)}{E[1 - S|D_1]} \right) [\Delta Y - E(\Delta Y|D_1, S = 0)] - \delta_1 S \right\}.$$

Theorem 3 shows that without quasi-stayers, the AOSS can be estimated at the $\sqrt{n}-$rate, and gives an expression of its estimator's asymptotic variance. With quasi-stayers, we conjecture that the AOSS cannot be estimated at the $\sqrt{n}-$rate. This conjecture is based on a result from

Graham and Powell (2012). Though their result applies to a broader class of estimands, it implies in particular that with quasi-stayers,

$$\lim_{\eta \downarrow 0} E\left(\frac{\Delta Y - E(\Delta Y|S=0)}{\Delta D}\middle| S_\eta = 1\right)$$

cannot be estimated at a faster rate than $n^{1/3}$. The estimand in the previous display is closely related to our estimand

$$\lim_{\eta \downarrow 0} E\left(\frac{\Delta Y - E(\Delta Y|D_1, S=0)}{\Delta D}\middle| S_\eta = 1\right)$$

in Theorem 2, and is equal to it if $E(\Delta Y|D_1, S = 0) = E(\Delta Y|S = 0)$. Then, even though the assumptions in Graham and Powell (2012) differ from ours, it seems reasonable to assume that their general conclusion still applies to our set-up: here as well, owing to δ_1's irregular identification, this parameter can probably not be estimated at the parametric $\sqrt{n}-$rate with quasi-stayers. This is one of the reasons that lead us to consider, in the next section, another target parameter that can be estimated at the parametric $\sqrt{n}-$rate with quasi-stayers.

Finally, on top of estimating the AOSS, a natural idea would be to also use the estimators $(\Delta Y_i - \widehat{E}(\Delta Y|D_1 = D_{i,1}, S = 0))/\Delta D_i$ to estimate the distribution of switchers' slopes, rather than just their average. Doing so is not straightforward, but may be achieved resorting to deconvolution techniques, under some assumptions (see, e.g., Arellano and Bonhomme, 2012). We do not pursue that route here.

4 Estimating a weighted average of switchers' potential-outcome's slopes

4.1 Target parameter

In this section, our target parameter is

$$\begin{aligned}\delta_2 :=& E\left(\frac{|D_2 - D_1|}{E(|D_2 - D_1||S=1)} \times \frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}\middle| S = 1\right)\\ =& \frac{E\left(\text{sgn}(D_2 - D_1)(Y_2(D_2) - Y_2(D_1))|S=1\right)}{E(|D_2 - D_1||S=1)}.\end{aligned}$$

δ_2 is a weighted average of the slopes of switchers' potential outcome functions from their period-one to their period-two treatments, where slopes receive a weight proportional to switchers' absolute treatment change from period one to two. δ_2 gives a weight larger than one to the slopes of switchers whose treatment increased more than the average among switchers ($|D_2 - D_1| > E(|D_2 - D_1||S = 1)$), and it gives a weight lower than one to the slopes of switchers whose treatment increased less than the average among switchers ($|D_2 - D_1| < E(|D_2 - D_1||S = 1)$). Accordingly, we refer to δ_2 as the Weighted Average Of Switchers' Slopes (WAOSS). All slopes are weighted positively, so the WAOSS satisfies the no-sign reversal property, like the AOSS. The WAOSS and AOSS may differ, if switchers' slopes are correlated with $|D_2 - D_1|$.

The AOSS and WAOSS serve different purposes. As discussed above, under shape restrictions on the potential outcome function, the AOSS can be used to identify or bound the effect of other treatment changes than the actual change switchers experienced from period one to two. The WAOSS cannot serve that purpose, but under some assumptions, it may be used to conduct a cost-benefit analysis of the treatment changes that took place from period one to two. To simplify the discussion, let us assume in the remainder of this paragraph that $D_2 \geq D_1$. Assume also that the outcome is a measure of output, such as agricultural yields or wages, expressed in monetary units. Finally, assume that the treatment is costly, with a cost linear in dose, uniform across units, and known to the analyst: the cost of giving d units of treatment to a unit at period t is $c_t \times d$ for some known $(c_t)_{t\in\{1,2\}}$. Then, D_2 is beneficial relative to D_1 if and only if

$$\begin{aligned} E(Y_2(D_2) - c_2 D_2) >& E(Y_2(D_1) - c_2 D_1) \\ \Leftrightarrow \delta_2 >& c_2, \end{aligned}$$

where the equivalence follows from the fact we momentarily assume $D_2 \geq D_1$. Then, comparing δ_2 to the per-unit treatment cost is sufficient to evaluate if changing the treatment from D_1 to D_2 was beneficial. For instance, when studying the effects of an increase in temperatures, due to climate change, on various economic outcomes (e.g. agricultural yields, labor productivity, etc.), one can compare the sum of the WAOSSes of all those outcomes, a measure of the average total economic impact of raising temperatures by one degree across switchers, to an estimate of the per-switcher cost of reducing emissions by a sufficient amount to reduce temperatures by a degree over the period under consideration.

4.2 Identification

We now show that δ_2 is identified by a regression-based and by a propensity-score based estimand. Though we do not formally show it here, δ_2 could also be identified by a doubly-robust estimand. This is a further advantage of considering the WAOSS rather than the AOSS: unlike the former, the latter parameter can only be identified by a regression-based estimand.

Let $S_+ = 1\{D_2 - D_1 > 0\}$, $S_- = 1\{D_2 - D_1 < 0\}$ and

$$\begin{aligned} \delta_{2+} :=& \frac{E\left(Y_2(D_2) - Y_2(D_1)|S_+ = 1\right)}{E(D_2 - D_1|S_+ = 1)}, \\ \delta_{2-} :=& \frac{E\left(Y_2(D_1) - Y_2(D_2)|S_- = 1\right)}{E(D_1 - D_2|S_- = 1)}. \end{aligned}$$

Hereafter, units with $S_+ = 1$ are referred to as “switchers up”, while units with $S_- = 1$ are referred to as “switchers down”. Thus, δ_{2+} is the WAOSS of switchers up, and δ_{2-} is the WAOSS of switchers down. One has

$$\begin{aligned}\delta_2 =& \frac{P(S_+ = 1|S = 1)E(D_2 - D_1|S_+ = 1)}{E(|D_2 - D_1||S = 1)}\delta_{2+} \\ &+ \frac{P(S_- = 1|S = 1)E(D_1 - D_2|S_- = 1)}{E(|D_2 - D_1||S = 1)}\delta_{2-}. \qquad (4)\end{aligned}$$

To identify δ_{2+} (resp. δ_{2-}) we use DID estimands comparing switchers up (resp. switchers down) to stayers with the same period-one treatment. This requires that there be no value of D_1 such that some switchers up (resp. switchers down) have that baseline treatment while there is no stayer with the same baseline treatment, as stated formally in Point 1 (resp. 2) of Assumption 7 below.

Assumption 7 *(Support conditions for WAOSS identification)*

1. $0 < P(S_+ = 1)$*, and* $0 < P(S_+ = 1|D_1)$ *implies that* $0 < P(S = 0|D_1)$*.*
2. $0 < P(S_- = 1)$*, and* $0 < P(S_- = 1|D_1)$ *implies that* $0 < P(S = 0|D_1)$*.*

Theorem 4 1. *If Assumptions 1-2 and Point 1 of Assumption 7 hold,*

$$\begin{aligned}\delta_{2+} &= \frac{E\left(\Delta Y - E(\Delta Y|D_1, S = 0)|S_+ = 1\right)}{E(\Delta D|S_+ = 1)} \qquad (5)\\ &= \frac{E\left(\Delta Y|S_+ = 1\right) - E\left(\Delta Y \frac{P(S_+=1|D_1)}{P(S=0|D_1)}\frac{P(S=0)}{P(S_+=1)}\Big|S = 0\right)}{E(\Delta D|S_+ = 1)}. \qquad (6)\end{aligned}$$

2. *If Assumptions 1-2 and Point 2 of Assumption 7 hold,*

$$\begin{aligned}\delta_{2-} &= \frac{E\left(\Delta Y - E(\Delta Y|D_1, S = 0)|S_- = 1\right)}{E(\Delta D|S_- = 1)} \qquad (7)\\ &= \frac{E\left(\Delta Y|S_- = 1\right) - E\left(\Delta Y \frac{P(S_-=1|D_1)}{P(S=0|D_1)}\frac{P(S=0)}{P(S_-=1)}\Big|S = 0\right)}{E(\Delta D|S_- = 1)}. \qquad (8)\end{aligned}$$

If Assumptions 1-2 and Assumption 7 hold,

$$\begin{aligned}\delta_2 =& \frac{E\left[sgn(\Delta D)\left(\Delta Y - E(\Delta Y|D_1, S = 0)\right)\right]}{E[|\Delta D|]} \qquad (9)\\ =& \frac{E\left[sgn(\Delta D)\Delta Y\right] - E\left[\Delta Y \frac{P(S_+=1|D_1)-P(S_-=1|D_1)}{P(S=0|D_1)}P(S = 0)\Big|S = 0\right]}{E[|\Delta D|]}. \qquad (10)\end{aligned}$$

Point 1 of Theorem 4 shows that δ_{2+}, the WAOSS of switchers-up, is identified by two estimands, a regression-based and a propensity-score-based estimand. Point 2 of Theorem 4 shows that δ_{2-}, the WAOSS of switchers down, is identified by two estimands similar to those identifying δ_{2+}, replacing S_+ by S_-. Finally, if the conditions in Point 1 and 2 of Theorem 4 jointly hold, it directly follows from (4) that δ_2, the WAOSS of all switchers, is identified by a weighted average of the estimands in Equations (5) and (7), and by a weighted average of the estimands in Equations (6) and (8). Those weighted averages simplify into the expressions given in Point 3 of Theorem 4.

4.3 Estimation and inference

The regression-based estimands identifying δ_{2+} and δ_{2-} can be estimated following almost the same steps as in Section 3.3. Specifically, let

$$\widehat{\delta}^r_{2+} := \frac{\frac{1}{n_+}\sum_{i:S_{i+}=1}\left(\Delta Y_i - \widehat{E}(\Delta Y|D_1 = D_{i,1}, S = 0)\right)}{\frac{1}{n_+}\sum_{i:S_{i+}=1}\Delta D_i}$$
$$\widehat{\delta}^r_{2-} := \frac{\frac{1}{n_-}\sum_{i:S_{i-}=1}\left(\Delta Y_i - \widehat{E}(\Delta Y|D_1 = D_{i,1}, S = 0)\right)}{\frac{1}{n_-}\sum_{i:S_{i-}=1}\Delta D_i},$$

where $n_+ = \#\{i : S_{i+} = 1\}$ and $n_- = \#\{i : S_{i-} = 1\}$, and where $\widehat{E}(\Delta Y|D_1, S = 0)$ is the series estimator of $E(\Delta Y|D_1, S = 0)$ defined in Section 3.3 of the paper. Then, let

$$\widehat{w}_+ = \frac{\frac{n_+}{n} \times \frac{1}{n_+}\sum_{i:S_{i+}=1}\Delta D_i}{\frac{n_+}{n} \times \frac{1}{n_+}\sum_{i:S_{i+}=1}\Delta D_i - \frac{n_-}{n} \times \frac{1}{n_-}\sum_{i:S_{i-}=1}\Delta D_i},$$

and let

$$\widehat{\delta}^r_2 = \widehat{w}_+\widehat{\delta}^r_{2+} + (1 - \widehat{w}_+)\widehat{\delta}^r_{2-}$$

be the corresponding estimator of δ_2.

We now propose estimators of the propensity-score-based estimands identifying δ_{2+} and δ_{2-} in Equations (6) and (8). Let $\widehat{P}(S_+ = 1) = n_+/n$ (resp. $\widehat{P}(S_- = 1) = n_-/n$, $\widehat{P}(S = 0) = (n - n_s)/n$) be an estimator of $P(S_+ = 1)$ (resp. $P(S_- = 1)$, $P(S = 0)$). Let $\widehat{P}(S_+ = 1|D_1)$ (resp. $\widehat{P}(S_- = 1|D_1)$, $\widehat{P}(S = 0|D_1)$) be a non-parametric estimator of $P(S_+ = 1|D_1)$ (resp. $P(S_- = 1|D_1)$, $P(S = 0|D_1)$) using a series logistic regression of S_{i+} (resp. S_{i-}, $1 - S_i$) on polynomials in D_1 $(p_{k,K_n}(D_1))_{1\leq k\leq K_n}$. We make the following technical assumption.

Assumption 8 *(Technical conditions for asymptotic normality of propensity-score WAOSS estimator)*

1. *D_1 is continuously distributed on a compact interval I, with* $\inf_{d\in I} f_{D_1}(d) > 0$.
2. *$E[\Delta Y^2] < \infty$ and $d \mapsto E[\Delta Y^2|D_1 = d]$ is bounded on I*

3. $0 < E[S_+] < 1$, $0 < E[S_-] < 1$, $E[S] > 0$ *and* $\sup_{d\in I} E[S|D_1 = d] < 1$.

4. *The functions* $d \mapsto E[\Delta Y(1-S)|D_1 = d]$, $d \mapsto E[S|D_1 = d]$, $d \mapsto E[S_+|D_1 = d]$ *and* $d \mapsto E[S_-|D_1 = d]$ *are four times continuously differentiable.*

5. *The polynomials* $d \mapsto p_{k,K_n}(d)$, $k \leq 1 \leq K_n$ *are orthonormal on* I *and* $K_n = Cn^\nu$ *where* $1/10 < \nu < 1/6$.

Let

$$\widehat{\delta}_{2+}^{ps} := \frac{\frac{1}{n_+}\sum_{i:S_{i+}=1}\Delta Y_i - \frac{1}{n-n_s}\sum_{i:S_i=0}\Delta Y_i \frac{\widehat{P}(S_+=1|D_1=D_{i1})}{\widehat{P}(S=0|D_1=D_{i1})}\frac{\widehat{P}(S=0)}{\widehat{P}(S_+=1)}}{\frac{1}{n_+}\sum_{i:S_{i+}=1}\Delta D_i}$$
$$\widehat{\delta}_{2-}^{ps} := \frac{\frac{1}{n_-}\sum_{i:S_{i-}=1}\Delta Y_i - \frac{1}{n-n_s}\sum_{i:S_i=0}\Delta Y_i \frac{\widehat{P}(S_-=1|D_1=D_{i1})}{\widehat{P}(S=0|D_1=D_{i1})}\frac{\widehat{P}(S=0)}{\widehat{P}(S_-=1)}}{\frac{1}{n_-}\sum_{i:S_{i-}=1}\Delta D_i},$$

and let

$$\widehat{\delta}_2^{ps} = \widehat{w}_+\widehat{\delta}_{2+}^{ps} + (1-\widehat{w}_+)\widehat{\delta}_{2-}^{ps}$$

be the corresponding estimator of δ_2. Let

$$\psi_{2+} := \frac{1}{E(\Delta D S_+)}\left\{\left(S_+ - E(S_+|D_1)\frac{(1-S)}{E(1-S|D_1)}\right)(\Delta Y - E(\Delta Y|D_1, S=0)) - \delta_{2+}\Delta D S_+\right\}$$
$$\psi_{2-} := \frac{1}{E(\Delta D S_-)}\left\{\left(S_- - E(S_-|D_1)\frac{(1-S)}{E(1-S|D_1)}\right)(\Delta Y - E(\Delta Y|D_1, S=0)) - \delta_{2-}\Delta D S_-\right\}$$
$$\psi_2 := \frac{1}{E(|\Delta D|)}\left\{\left(S_+ - S_- - E(S_+ - S_-|D_1)\frac{(1-S)}{E(1-S|D_1)}\right)\right.$$
$$\left.\times(\Delta Y - E(\Delta Y|D_1, S=0)) - \delta_2|\Delta D|(S_+ - S_-)\right\}.$$

Theorem 5 1. *If Assumptions 1-3 and 6 hold,*

$$\sqrt{n}\left((\widehat{\delta}_{2+}^r, \widehat{\delta}_{2-}^r)' - (\delta_{2+}, \delta_{2-})'\right) \xrightarrow{d} \mathcal{N}(0, V((\psi_{2+}, \psi_{2-})')).$$

and

$$\sqrt{n}\left(\widehat{\delta}_2^r - \delta_2\right) \xrightarrow{d} \mathcal{N}(0, V(\psi_2).$$

2. *If Assumptions 1-3 and 8 hold,*

$$\sqrt{n}\left((\widehat{\delta}_{2+}^{ps}, \widehat{\delta}_{2-}^{ps})' - (\delta_{2+}, \delta_{2-})'\right) \xrightarrow{d} \mathcal{N}(0, V((\psi_{2+}, \psi_{2-})')).$$

and

$$\sqrt{n}\left(\widehat{\delta}_2^{ps} - \delta_2\right) \xrightarrow{d} \mathcal{N}(0, V(\psi_2).$$

Point 1 (resp. 2) of Theorem 5 shows that $\widehat{\delta}^r_{2+}$, $\widehat{\delta}^r_{2-}$ and $\widehat{\delta}^r_2$ (resp. $\widehat{\delta}^{ps}_{2+}$, $\widehat{\delta}^{ps}_{2-}$ and $\widehat{\delta}^{ps}_2$) are $\sqrt{n}-$consistent and jointly asymptotically normal.

$\widehat{\delta}^r_2$ can be computed by the `did_multiplegt` Stata command. To do so, the syntax is:
`did_multiplegt Y G T D, controls(`$(1\{t=2\}p_{k,K_n}(D_1))_{1\leq k\leq K_n}$`).`
Essentially, one just needs to control for the interaction of a period-two indicator and the polynomial in D_1 one uses to estimate switchers' counterfactual trend.

5 Extensions

5.1 No stayers

So far we have assumed that $P(S=0)>0$, meaning that there are units whose treatment does not change over time. We now show that when this is not the case, the WAOSS can still be estimated, provided there are quasi-stayers, meaning units whose treatment "barely" changes between periods 1 and 2, a requirement we formalize below. The AOSS can also still be estimated without stayers and with quasi-stayers, but we conjecture that the resulting estimator will converge at an even slower rate than the estimator of the WAOSS we propose below, so we do not consider it here.

To identify the WAOSS without stayers, we use a DID estimand comparing movers and quasi-stayers with the same period-one treatment. This requires that there be no value of the period-one treatment D_1 such that some movers but no quasi-stayers have that value, as stated formally below.

Assumption 9 *(Support condition without stayers) $P(S=1)=1$ and for all $\eta>0$, $P(S_\eta=0|D_1)>0$.*

Assumption 9 implies that for all $\eta>0$, $P(S_\eta=0)>0$. This formalizes our requirement that there be quasi-stayers: for any η, including very small ones, there must be a strictly positive proportion of units whose treatment changes by less than η. Because $S_\eta\leq S$ for all $\eta>0$, Assumption 9 is weaker than Assumption 4. In particular, it does not require that $P(S=0)>0$.

Theorem 6 *If Assumptions 1, 2, and 9 hold,*

$$\delta_2=\frac{E\left[sgn(\Delta D)\left(\Delta Y-\lim_{\eta\downarrow 0}E(\Delta Y|D_1,S_\eta=0)\right)\right]}{E[|\Delta D|]}.$$

Theorem 6 shows that without stayers, δ_2 is identified by a limiting regression-based estimand similar to that in (9), except that movers are compared to quasi-stayers with the same baseline treatment, letting quasi-stayers' ΔD go to zero, to ensure that in the limit, their outcome evolution only reflects the effect of time and not the effect of their treatment change. Without stayers, δ_2 could also be identified by a limiting propensity-score-based estimand. However, while we can rely on an existing statistical result to study the asymptotic distribution of an estimator of the limiting regression-based estimand in Theorem 6, we are not aware of a similar result for a limiting propensity-score-based estimand. Accordingly, we only consider the limiting regression-based estimand.

To estimate δ_2, remark that δ_2 is a simple function of $E(\Delta Y)$, $E(\Delta D)$ and the more complicated term

$$\theta_0 := E\left[\text{sgn}(\Delta D)\lim_{\eta\downarrow 0} E(\Delta Y|D_1, S_\eta = 0)\right].$$

Let $g(u, d_1) := E[Y_2 - Y_1|D_2 - D_1 = u, D_1 = d_1]$. Under regularity conditions, we have $\theta_0 = E\left[g(0, D_1)\right]$. Thus, θ_0 corresponds to the marginal integration of the nonparametric regression function g with respect to the distribution of D_1. We then follow Linton and Nielsen (1995) and consider a plug-in estimator where g is estimated by a local linear estimator. Specifically, $\widehat{g}(u, d_1) := \widehat{\beta}_0(u, d_1)$ with

$$(\widehat{\beta}_0(u, d_1), \widehat{\beta}_1(u, d_1), \widehat{\beta}_2(u, d_1)) := \arg\min_{(\beta_0,\beta_1,\beta_2)\in\mathbb{R}^3} \sum_{i=1}^{n} k_{b_1}(\Delta D_i - u)k_{b_2}(D_{1i} - d_1) \times \left(\Delta Y_i - \beta_0 - \beta_1\Delta D_i - \beta_2 D_{1i}\right)^2,$$

where $k_b(t) := k(t/b)/b$, k is a kernel function and b_1, b_2 are bandwidths. Then, we let $\widehat{\theta} := \frac{1}{n}\sum_{i=1}^{n} \text{sgn}(\Delta D_i)\widehat{g}(0, D_{1i})$ and the estimator of δ_2 is finally

$$\widehat{\delta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} \Delta Y_i - \widehat{\theta}}{\frac{1}{n}\sum_{i=1}^{n} \Delta D_i}.$$

We conjecture, following Linton and Nielsen (1995) or Corollary 2 in Kong et al. (2010), that $\widehat{\delta}_2$ is asymptotically normal but with a nonparametric rate of convergence corresponding to a univariate nonparametric regression.

5.2 Discrete treatments

While in this paper we focus on continuous treatments, our results can also be applied to discrete treatments. In Section 4 of their Web Appendix, de Chaisemartin and D'Haultfœuille (2020) already propose a DID estimator of the effect of a discrete treatment. The plug-in estimator of δ_2 one can form following Theorem 4 and using simples averages to estimate the non-parametric regressions or the propensity scores is numerically equivalent to the estimator therein. This paper still makes two contributions relative to de Chaisemartin and D'Haultfœuille (2020) when the treatment is discrete. First, the estimator based on Theorem 1 was not proposed therein. Second, with a discrete treatment taking a large number of values, the estimator in de Chaisemartin and

D'Haultfœuille (2020) may not be applicable as it requires finding switchers and stayers with the exact same period-one treatment, which may not always be feasible. Instead, one can follow Theorem 4, using a parametric model to estimate the regressions or the propensity scores entering the estimands in that theorem.

5.3 More than two time periods

In this section, we assume the representative unit is observed at $T > 2$ time periods. Let $(D_1, ..., D_T)$ denote the unit's treatments and $\mathcal{D}_t = \text{Supp}(D_t)$ for all $t \in \{1, ..., T\}$. For any $t \in \{1, ..., T\}$, and for any $d \in \mathcal{D}_t$ let $Y_t(d)$ denote the unit's potential outcome at period t with treatment d. Finally, let Y_t denote their observed outcome at t. For any $t \in \{2, ..., T\}$, let $S_t = 1\{D_t \neq D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment switches from period $t-1$ to t. Let also $S_{+,t} = 1\{D_t > D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment increases from period $t-1$ to t, and let $S_{-,t} = 1\{D_t < D_{t-1}\}$ be an indicator equal to 1 if the unit's treatment decreases.

In this section, we will assume that the assumptions made in the paper, rather than just holding for $t = 1$ and $t = 2$, actually hold for all pairs of consecutive time periods $(t-1, t)$. For instance, we replace Assumption 1 by the following condition.

Assumption 10 *(Parallel trends) For all $t \geq 2$, for all $d \in \mathcal{D}_{t-1}$, $E(\Delta Y_t(d)|D_{t-1} = d, D_t) = E(\Delta Y_t(d)|D_{t-1} = d)$.*

To preserve space, we do not restate our other assumptions with more than two periods.

Let

$$\delta_{1t} = E\left(\frac{Y_t(D_t) - Y_t(D_{t-1})}{D_t - D_{t-1}}\middle|S_t = 1\right),$$

$$\delta_{2+t} = \frac{E\left(Y_t(D_t) - Y_t(D_{t-1})|S_{+,t} = 1\right)}{E(D_t - D_{t-1}|S_{+,t} = 1)},$$

$$\delta_{2-t} = \frac{E\left(Y_t(D_{t-1}) - Y_t(D_t)|S_{-,t} = 1\right)}{E(D_{t-1} - D_t|S_{-,t} = 1)}.$$

Theorem 7 *If Assumption 10 and generalizations of Assumptions 2-5 to more than two periods hold,*

$$\sum_{t=2}^{T} \frac{P(S_t = 1)}{\sum_{k=2}^{T} P(S_k = 1)} \delta_{1t} = \sum_{t=2}^{T} \frac{P(S_t = 1)}{\sum_{k=2}^{T} P(S_k = 1)} E\left(\frac{\Delta Y_t - E(\Delta Y_t|D_{t-1}, S_t = 0)}{\Delta D_t}\middle|S_t = 1\right).$$

Theorem 8 *1. If Assumption 10 and generalizations of Assumption 2 and Point 1 of Assumption 7 to more than two periods hold,*

$$\sum_{t=2}^{T} \frac{P(S_{+,t}=1)E(\Delta D_t|S_{+,t}=1)}{\sum_{k=2}^{T} P(S_{+,k}=1)E(\Delta D_k|S_{+,k}=1)}\delta_{2+t}$$
$$= \sum_{t=2}^{T} \frac{P(S_{+,t}=1)E(\Delta D_t|S_{+,t}=1)}{\sum_{k=2}^{T} P(S_{+,k}=1)E(\Delta D_k|S_{+,k}=1)} \frac{E\left(\Delta Y_t - E(\Delta Y_t|D_{t-1}, S_t=0)|S_{+,t}=1\right)}{E(\Delta D_t|S_{+,t}=1)}$$
$$= \sum_{t=2}^{T} \frac{P(S_{+,t}=1)E(\Delta D_t|S_{+,t}=1)}{\sum_{k=2}^{T} P(S_{+,k}=1)E(\Delta D_k|S_{+,k}=1)} \frac{E\left(\Delta Y_t|S_{+,t}=1\right) - E\left(\Delta Y_t \frac{P(S_{+,t}=1|D_{t-1})}{P(S_t=0|D_{t-1})} \frac{P(S_t=0)}{P(S_{+,t}=1)}\Big|S_t=0\right)}{E(\Delta D|S_{+,t}=1)}.$$

2. If Assumption 10 and generalizations of Assumption 2 and Point 2 of Assumption 7 to more than two periods hold,

$$\sum_{t=2}^{T} \frac{P(S_{-,t}=1)E(-\Delta D_t|S_{-,t}=1)}{\sum_{k=2}^{T} P(S_{-,k}=1)E(-\Delta D_k|S_{-,k}=1)}\delta_{2-t}$$
$$= \sum_{t=2}^{T} \frac{P(S_{-,t}=1)E(-\Delta D_t|S_{-,t}=1)}{\sum_{k=2}^{T} P(S_{-,k}=1)E(-\Delta D_k|S_{-,k}=1)} \frac{E\left(\Delta Y_t - E(\Delta Y_t|D_{t-1}, S_t=0)|S_{-,t}=1\right)}{E(\Delta D_t|S_{-,t}=1)}$$
$$= \sum_{t=2}^{T} \frac{P(S_{-,t}=1)E(-\Delta D_t|S_{-,t}=1)}{\sum_{k=2}^{T} P(S_{-,k}=1)E(-\Delta D_k|S_{-,k}=1)} \frac{E\left(\Delta Y_t|S_{-,t}=1\right) - E\left(\Delta Y_t \frac{P(S_{-,t}=1|D_{t-1})}{P(S_t=0|D_{t-1})} \frac{P(S_t=0)}{P(S_{-,t}=1)}\Big|S_t=0\right)}{E(\Delta D|S_{-,t}=1)}.$$

Theorems 7 and 8 are straightforward generalizations of Theorems 1 and 4 to settings with more than two time periods. Note that we propose different weights to aggregate the AOSS and WAOSS across time periods. For the AOSS, the weights are just proportional to the proportion of switchers between $t-1$ and t. For the WAOSS, the weights are proportional to the proportion of switchers times their average treatment switch.

A regression-based estimator following the first point of Theorem 8 can be computed as follows. First, one restricts the sample to periods 1 and 2, and to stayers and switchers in whose treatment increased. Then, one computes our regression-based estimator in that subsample, using the same syntax as that given in the last paragraph of Section 4.3 in the paper. Then, one repeats the same procedure, restricting the sample to periods 2 and 3, to periods 3 and 4, ..., to periods $T-1$ and T, and replacing D_1 by D_{t-1}. Finally, one aggregates the estimators computed for every pair of consecutive time periods, using the sample equivalents of the weights in the first point of Theorem 8. A regression-based estimator following the second point of Theorem 8 can be computed similarly, restricting the sample to stayers and switchers out for each pair of consecutive time periods.

5.4 Testing for pre-trends

With several time periods, one can test the following condition, which is closely related to Assumption 10:

Assumption 11 *(Testable parallel trends) For all $t \geq 2, t \leq T-1$, for all $d \in \mathcal{D}_{t-1}$, $E(\Delta Y_t(d)|D_{t-1} = d, D_t, D_{t+1}) = E(\Delta Y_t(d)|D_{t-1} = d)$.*

To test that condition, one can compute a placebo version of the estimators described in the previous subsection, replacing ΔY_t by ΔY_{t-1}, and restricting the sample, for each pair of consecutive time periods $(t-1, t)$, to units whose treatment did not change between $t-2$ and $t-1$. Thus, the placebo compares the average ΔY_{t-1} of the $t-1$-to-t switchers and stayers, restricting attention to $t-2$-to-t stayers. The placebo we propose generalizes that in de Chaisemartin and D'Haultfœuille (2020) to applications with a continuous treatment. Finally, note that in applications with no stayers, it is less straightforward to propose placebo estimators of our parallel trends assumption. The actual estimator already does not converge at the parametric rate. A placebo would compare the average ΔY_{t-1} of $t-1$-to-t switchers and quasi-stayers, restricting attention to $t-2$-to-t quasi-stayers. This placebo may converge at an even slower rate than the actual estimator, unless a strictly positive proportion of units are $t-2$-to-$t-1$ stayers or quasi-stayers.

5.5 Allowing for dynamic effects

In this subsection, we allow for dynamic effects. The results below generalize those in de Chaisemartin and D'Haultfœuille (2022), who allow for dynamic effects, but require that groups' period-one treatment take a finite number of values. Here, we assume that groups' period-one treatment is continuously distributed. Though this is not of essence, to ease exposition we require that the representative unit's treatment can never get lower than their period-one treatment:

Assumption 12 *(Lowest treatment at period one) For all t, $D_t \geq D_1$.*

It may be the case that Assumption 12 fails, for instance because for some units, the treatment is at its lowest at period one, but for other units the treatment is at its highest at period one. In that case, one can split the sample in two, and compute the estimators based on the results below in the first subsample, and the negative of those estimators in the second subsample. There may also be some units that have a treatment higher than their period-one treatment at some time periods, but a lower treatment at other time periods. One may have to discard such units, as the dynamic treatment effects of those units may conflate together effects of increases and decreases of the treatment, and may not be interpretable, see the discussion of Assumption 5 in de Chaisemartin and D'Haultfœuille (2022) for further details.

For all $\boldsymbol{d} \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, let $Y_t(\boldsymbol{d})$ denote the potential outcome of the representative unit at period t, if their treatments from period one to T are equal to $\boldsymbol{d}$. This dynamic potential outcome framework is similar to that in Robins (1986). It allows for the possibility that the outcome at time t be affected by past and future treatments. Let $\boldsymbol{D} = (D_1, ..., D_T)$ be a $1 \times T$ vector stacking the representative unit's treatments from period one to T.

Assumption 13 *(No Anticipation) For all $\boldsymbol{d} \in \mathcal{D}_1 \times ... \times \mathcal{D}_T$, $Y_t(\boldsymbol{d}) = Y_t(d_1, ..., d_t)$.*

Assumption 13 requires that the current outcome do not depend on future treatments, the so-called no-anticipation hypothesis. Abbring and Van den Berg (2003) have discussed that assumption in the context of duration models, and Malani and Reif (2015), Botosaru and Gutierrez (2018), and Sun and Abraham (2021) have discussed it in the context of DID models.

Let $F = \min\{t : D_t \neq D_{t-1}\}$ denote the first date at which the representative unit's treatment changes, with the convention that $F = T+1$ if their treatment never changes. For all $d \in \mathcal{D}_1$, we let $\boldsymbol{d} = (d, ..., d)$ denote a $1 \times T$ vector with coordinates equal to d. We also let $\boldsymbol{D}_1 = (D_1, ..., D_1)$ denote a $1 \times T$ vector with coordinates equal to D_1, the unit's period-one treatment. We make the following assumptions, which generalize Assumptions 1 and 4 to settings with more than two time periods and dynamic effects.

Assumption 14 *(Parallel trends, allowing for dynamic effects) For all $t \geq 2$, for all $d \in \mathcal{D}_1$, $E(Y_t(\boldsymbol{D}_1) - Y_{t-1}(\boldsymbol{D}_1)|\boldsymbol{D}) = E(Y_t(\boldsymbol{D}_1) - Y_{t-1}(\boldsymbol{D}_1)|D_1)$.*

Assumption 15 *(Support condition, allowing for dynamic effects) For all $t \geq 2$, $0 < P(F = t)$, and for all $t' > t$, the support of $D_1|F = t$ is included in that of $D_1|F > t - 1 + \ell'$: $0 < P(F = t|D_1) \Rightarrow 0 < P(F > t - 1 + \ell'|D_1)$.*

Assumption 15 requires that the probability that the representative unit never changes its treatment be strictly positive. When that condition fails, results below still hold, till the last period where the probability of having never changed treatment is still strictly positive.

Theorem 9 below generalizes Theorems 7 and 8 to allow for dynamic effects. Let

$$\delta_\ell = E\left(Y_{F-1+\ell}(\boldsymbol{D}) - Y_{F-1+\ell}(\boldsymbol{D}_1)|F \leq T - \ell + 1\right).$$

The effects δ_ℓ identified in Theorem 9 correspond to the non-normalized event-study effects δ_ℓ studied in de Chaisemartin and D'Haultfœuille (2022). We refer the reader to that paper for further details on those effects and their interpretation.

Theorem 9 *If Assumptions 13, 14 and 15 hold, for any $\ell \in \{1, ..., T-1\}$,*

$$\begin{aligned} &\delta_\ell \\ &= \sum_{t=2}^{T-\ell+1} \frac{P(F = t)}{\sum_{k=2}^{T-\ell+1} P(F = k)} E\left(Y_{t-1+\ell} - Y_{t-1} - E(Y_{t-1+\ell} - Y_{t-1}|D_1, F > t - 1 + \ell)|F = t\right) \quad (11) \\ &= \sum_{t=2}^{T-\ell+1} \frac{P(F = t)}{\sum_{k=2}^{T-\ell+1} P(F = k)} \left(E\left(Y_{t-1+\ell} - Y_{t-1}|F = t\right)\right. \\ &- E\left(\left.(Y_{t-1+\ell} - Y_{t-1}) \frac{P(F = t|D_1)}{P(F > t - 1 + \ell|D_1)} \frac{P(F > t - 1 + \ell)}{P(F = t)}\right| F > t - 1 + \ell\right)\Bigg). \quad (12) \end{aligned}$$

(11) shows that with a continuous period-one treatment, δ_ℓ is identified by a regression-based estimand, while (12) shows that it is also identified by a propensity-score-reweighting-based estimand. Those estimands generalize the estimands proposed in de Chaisemartin and D'Haultfœuille (2022) when the period-one treatment is discrete.

The non-normalized event-study effects compare, in period $t-1+\ell$, the actual outcome of units whose treatment changed for the first time ℓ periods ago to the counterfactual outcome they would have obtained if they had instead kept their period-one treatment from period one to t. Assumption 12 ensures that those effects can be interpreted as effects of increasing the treatment, but they may aggregate together the effects of many different treatment trajectories. de Chaisemartin and D'Haultfœuille (2022) also propose normalized event-study estimands, that equal weighted averages of the slopes of the potential outcome function with respect to the current treatment and its lags. Those normalized event-study effects are also identified by regression-based and propensity-score-based estimands similar to those in Theorem 9.

A regression-based estimator following Theorem 9 can be computed by the `did_multiplegt_dyn` Stata command. To do so, the syntax is:
`did_multiplegt_dyn Y G T D, dynamic(`ℓ`) controls(`$(\sum_{t'=2}^{T}1\{t=t'\}p_{k,K_n}(D_1))_{1\leq k\leq K_n}$`),`
where ℓ is the number of event-study effects one wishes to estimate. Essentially, one just needs to control for the interaction of time fixed effects and the polynomial in D_1 one wishes to use to estimate switchers' counterfactual trend. To estimate normalized event-study effects, one just needs to add `normalized` to the previous command.

6 Future work

In future work, we will use our results to revisit Fajgelbaum et al. (2020) and Deschênes and Greenstone (2007).

7 Proofs

Hereafter, Supp(X) denotes the support of X. Note that under Assumption 2, one can show that for all $(t,t')\in\{0,1\}^2$, $E(Y_t(D_{t'}))$ exists.

7.1 Theorem 1

The result is just a special case of Theorem 2, under Assumption 5 □

7.2 Theorem 2

First, observe that the sets $\{S_\eta = 1\}$ are decreasing for the inclusion and $\{S = 1\} = \cup_{\eta>0}\{S_\eta = 1\}$. Then, by continuity of probability measures,

$$\lim_{\eta\downarrow 0} P(S_\eta = 1) = P(S = 1) > 0, \tag{13}$$

where the inequality follows by Assumption 4. Thus, there exists $\underline{\eta} > 0$ such that for all $\eta \in (0, \underline{\eta})$, $P(S_\eta = 1) > 0$. Hereafter, we assume that $\eta \in (0, \underline{\eta})$.

We have $\text{Supp}(D_1|S_\eta = 1) \subseteq \text{Supp}(D_1|S = 1)$ and by Assumption 4, $\text{Supp}(D_1|S = 1) \subseteq \text{Supp}(D_1|S = 0)$. Thus, for all $(d_1, d_2) \in \text{Supp}(D_1, D_2|S_\eta = 1)$, $d_1 \in \text{Supp}(D_1|S = 0)$, so $E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, S = 0) = E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_1)$ is well-defined. Moreover, for almost all such (d_1, d_2),

$$\begin{aligned}E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_2) =& E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_1)\\ =& E(\Delta Y|D_1 = d_1, S = 0),\end{aligned} \tag{14}$$

where the first equality follows from Assumption 1. Now, by Point 2 of Assumption 2, $[Y_2(D_2) - Y_2(D_1)]/\Delta D$ admits an expectation. Moreover,

$$\begin{aligned}& E\left(\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle| S_\eta = 1\right)\\ =& E\left(\frac{E(Y_2(D_2) - Y_1(D_1)|D_1, D_2) - E(Y_2(D_1) - Y_1(D_1)|D_1, D_2)}{\Delta D}\middle| S_\eta = 1\right)\\ =& E\left(\frac{E(\Delta Y|D_1, D_2) - E(\Delta Y|D_1, S = 0)}{\Delta D}\middle| S_\eta = 1\right)\\ =& E\left(\frac{\Delta Y - E(\Delta Y|D_1, S = 0)}{\Delta D}\middle| S_\eta = 1\right),\end{aligned} \tag{15}$$

where the first equality follows from the law of iterated expectations, the second follows from (14), and the third again by the law of iterated expectations. Next,

$$\delta_1 = \Pr(S_\eta = 1|S = 1) E\left[\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle| S_\eta = 1\right] + E\left[(1 - S_\eta)\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle| S = 1\right].$$

Moreover,

$$\begin{aligned}\left|E\left[(1 - S_\eta)\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle| S = 1\right]\right| \leq& E\left[(1 - S_\eta)\left|\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\right|\middle| S = 1\right]\\ \leq& E\left[(1 - S_\eta)\overline{Y}|S = 1\right],\end{aligned}$$

where the second inequality follows by Assumption 2. Now, by (13) again, $\lim_{\eta\downarrow 0}(1 - S_\eta)\overline{Y} = 0$ a.s. Moreover, $(1 - S_\eta)\overline{Y} \leq \overline{Y}$ with $E[\overline{Y}|S = 1] < \infty$. Then, by the dominated convergence theorem,

$$\lim_{\eta\downarrow 0} E\left[(1 - S_\eta)\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle|S = 1\right] = 0.$$

We finally obtain

$$\delta_1 = \lim_{\eta\downarrow 0} E\left[\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle|S_\eta = 1\right]. \tag{16}$$

The result follows by combining (15) and (16) □

7.3 Theorem 3

Let $\Delta Y = Y_2 - Y_1$, $\Delta D = D_2 - D_1$, $\mu_1(D_1) = E[(1 - S)Y|D_1]$, $\mu_2(D_1) = E[1 - S|D_1]$. In what follows we let $\mu(D_1) = (\mu_1(D_1), \mu_2(D_1))'$. From Theorem 1, the parameter δ_1 is characterized by the condition:

$$0 = E\left[\frac{S}{\Delta D}\left(\Delta Y - \delta_1\Delta D - \frac{\mu_1(D_1)}{\mu_2(D_1)}\right)\right]$$

Define:

$$g(Z, \delta, \mu) = \frac{S}{\Delta D}\left(\Delta Y - \frac{\mu_1(D_1)}{\mu_2(D_2)}\right) - S\delta_1$$

where $Z = (Y_1, Y_2, D_1, D_2)$. Also define:

$$\mathcal{L}(Z, \mu, \delta_1, \tilde{\mu}) = -\frac{S}{\Delta D}\cdot\frac{1}{\tilde{\mu}_2(D_1)}\left(\mu_1(D_1) - \frac{\tilde{\mu}_1(D_1)}{\tilde{\mu}_2(D_1)}\mu_2(D_1)\right)$$

We verify conditions 6.1 to 6.3, 5.1(i) and 6.4(ii) to 6.6 in Newey (1994). Following his notation, we let $\mu_0 = (\mu_{10}, \mu_{20})'$ and δ_{10} represent the true parameters, and $g(Z, \mu) = g(Z, \delta_{10}, \mu)$.

Step 1. We verify condition 6.1. First, since S is binary $E[(S - E[S|D_1])^2|D_1] = V[S|D_1] \leq 1/4$. On the other hand, $E[((1 - S)\Delta Y - E[(1 - S)\Delta Y|D_1])^2|D_1] \leq E[\Delta Y^2|D_1] < \infty$ by part 2 of Assumption 6. Thus, condition 6.1 holds.

Step 2. We verify condition 6.2. Since $p^K(d_1)$ is a power series, the support of D_1 is compact and the density of D_1 is uniformly bounded below, by Lemma A.15 in Newey (1995) for each K there exists a constant nonsingular matrix A_K such that for $P^K(d_1) = A_K p^K(d_1)$, the smallest eigenvalue of $E[P^K(D_1)P^K(D_1)']$ is bounded away from zero uniformly over K, and $P^K(D_1)$ is a subvector of $P^{K+1}(D_1)$. Since the series-based propensity scores estimators are invariant to nonsingular linear transformations, we do not need to distinguish between $P^K(d_1)$ and $p^K(d_1)$ and thus conditions 6.2(i) and 6.2(ii) are satisfied. Finally, because $p_{1K}(d_1) \equiv 1$ for all K, for a vector $\tilde{\gamma} = (1, 0, 0, \ldots, 0)$ we have that $\tilde{\gamma}' p^k(d_1) = \tilde{\gamma}_1 \neq 0$ for all d_1. Since A_K is nonsingular, letting $\gamma = A_K^{-1\prime}\tilde{\gamma}$, $\gamma' P^k(d_1) = \tilde{\gamma}' A_K^{-1} P^K(d_1)$ is a non-zero constant for all d_1 and thus condition 6.2(iii) holds.

Step 3. We verify condition 6.3 for $d = 0$. Since $p^K(d_1)$ is a power series, the support of D_1 is compact and the functions to be estimated have 4 continuous derivatives, by Lemma A.12 in Newey (1995) there is a constant $C > 0$ such that there is π with $\left\|\mu - (p^K)'\pi\right\| \leq CK^{-\alpha}$, where in our case $\alpha = s/r = 4$ since the dimension of the covariates is 1 and the unknown functions are 4 times continuously differentiable. Thus, condition 6.3 holds.

Step 4. We verify condition 5.1(i). By part 3 of Assumption 6, $\mu_{20}(D_1) = E[1 - S|D_1] = 1 - E[S|D_1] \geq 1 - c_M$ for some constant c_M>0. Let $C = 1 - c_M$. For μ such that $\|\mu - \mu_0\|_\infty < C/2$,

$$
\begin{aligned}
&|g(Z,\mu) - g(Z,\mu_0) - \mathcal{L}(Z, \mu - \mu_0, \delta_{10}, \mu_0)| \\
&= \left|\frac{S}{\Delta D}\right| \left|\frac{\mu_1(D_1)}{\mu_2(D_1)} - \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)} - \frac{1}{\mu_{20}(D_1)}\left(\mu_1(D_1) - \mu_{10}(D_1) - \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)}(\mu_2(D_1) - \mu_{20}(D_1))\right)\right| \\
&\leq \frac{1}{c}\left|\frac{\mu_1(D_1)}{\mu_2(D_1)} - \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)} - \frac{1}{\mu_{20}(D_1)}\left(\mu_1(D_1) - \mu_{10}(D_1) - \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)}(\mu_2(D_1) - \mu_{20}(D_1))\right)\right| \\
&\leq \frac{1}{c} \cdot \frac{2\left(1 + |\mu_{10}(D_1)| \,/\, |\mu_{20}(D_1)|\right)}{C^2} \max\left\{|\mu_1(D_1) - \mu_{10}(D_1)|\,, |\mu_2(D_1) - \mu_{20}(D_1)|\right\}^2 \\
&\leq \frac{1}{c} \cdot \frac{2\left(1 + |\mu_{10}(D_1)| \,/\, |\mu_{20}(D_1)|\right)}{C^2} \|\mu - \mu_0\|_\infty^2
\end{aligned}
$$

where the first inequality follows from Assumption 5 and the second inequality follows from Lemma S3 in the Web Appendix of de Chaisemartin and D'Haultfœuille (2018). Thus, condition 5.1(i) holds.

Step 5. We verify condition 6.4(ii). First, $E[(1 + |\mu_{10}(D_1)| \,/\, |\mu_{20}(D_1)|)^2] < \infty$. For power series, by Lemma A.15 in Newey (1995), $\zeta_d(K) = \sup_{|\lambda|=d, x\in I}\left\|\partial^\lambda p^K(x)\right\| \leq CK^{1+2d}$ so setting $d = 0$,

$$
\zeta_0(K)\left((K/n)^{1/2} + K^{-\alpha}\right) \leq CK\left((K/n)^{1/2} + K^{-\alpha}\right) = C\left(\sqrt{\frac{K^3}{n}} + K^{1-\alpha}\right) \to 0
$$

since $\alpha = 4 > 1/2$, $K^7/n \to 0$ and $K \to \infty$. Finally,

$$
\sqrt{n}\zeta_0(K)^2\left(\frac{K}{n} + K^{-2\alpha}\right) \leq C^2\sqrt{n}K^2\left(\frac{K}{n} + K^{-2\alpha}\right) = C\left(\sqrt{\frac{K^6}{n}} + \sqrt{\frac{n}{K^{4\alpha-4}}}\right) \to 0
$$

since $K^7/n \to 0$ and for $\alpha = 4$, $K^{4\alpha-4}/n = K^{12}/n \to \infty$. Hence condition 6.4(ii) holds.

Step 6. We verify condition 6.5 for $d = 1$ and where $|\mu|_d = \sup_{|\lambda|\leq d, x\in I} \left\|\partial^\lambda \mu(x)\right\|$. Since $E[(1 + |\mu_{10}(D_1)| \,/\, |\mu_{20}(D_1)|)^2] < \infty$,

$$\begin{aligned}|\mathcal{L}(Z, \mu, \delta_{10}, \mu_0)| &= \left| \frac{S}{\Delta D} \cdot \frac{1}{\mu_{20}(D_1)} \left(\mu_1(D_1) - \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)} \mu_2(D_1) \right) \right| \\ &\leq \frac{1}{c(1 - c_M)} \left(1 + \left| \frac{\mu_{10}(D_1)}{\mu_{20}(D_1)} \right| \right) |\mu|_1 .\end{aligned}$$

Next, the same linear transformation of p^K as in Step 2, namely P^K is, by Lemma A.15 in Newey (1995), such that $\left|P_k^K\right|_d \leq CK^{1/2+2d}$. As a result, $\left(\sum_k \left|P_k^K\right|_1^2\right)^{1/2} \leq CK^{1+2d}$. Then, for $d = 1$,

$$\left(\sum_k \left|P_k^K\right|_1^2\right)^{1/2} \left(\sqrt{\frac{K}{n}} + K^{-\alpha}\right) \leq CK^3 \left(\sqrt{\frac{K}{n}} + K^{-\alpha}\right) = C\left(\sqrt{\frac{K^7}{n}} + K^{3-\alpha}\right) \to 0$$

since $K^7/n \to 0$ and $K^{3-\alpha} = K^{-1} \to 0$ for $\alpha = 4$. Thus, condition 6.5 holds.

Step 7. We verify condition 6.6. Condition 6.6(i) holds for

$$\delta(D_1) = [-E[S/\Delta D|D_1]/\mu_{20}(D_1)](1, -\mu_{10}(D_1)/\mu_{20}(D_1)).$$

Because the involved functions are continuously differentiable, by Lemma A.12 from Newey (1995) there exist π_K and ξ_K such that:

$$E\left[\left\|\delta(D_1) - \xi_K p^K(D_1)\right\|^2\right] \leq \left\|\delta - \xi_K p^K\right\|_\infty^2 \leq CK^{-2\alpha}$$

and

$$E\left[\left\|\mu_0(D_1) - \pi_K p^K(D_1)\right\|^2\right] \leq \left\|\mu_0 - \pi_K p^K\right\|_\infty^2 \leq CK^{-2\alpha}$$

were we recall that $\alpha = 4$. Thus, the first part of condition 6.6(ii) follows from

$$nE\left[\left\|\delta(D_1) - \xi_K p^K(D_1)\right\|^2\right] E\left[\left\|\mu_0(D_1) - \pi_K p^K(D_1)\right\|^2\right] \leq CnK^{-16} \to 0.$$

Next,

$$\zeta_0(K)^4 \frac{K}{n} \leq C\frac{K^5}{n} \to 0$$

and finally

$$\zeta_0(K)^2 E\left[\left\|\mu_0(D_1) - \pi_K p^K(D_1)\right\|^2\right] \leq CK^{2-2\alpha} \to 0$$

and

$$E\left[\left\|\delta(D_1) - \xi_K p^K(D_1)\right\|^2\right] \leq CK^{-2\alpha} \to 0.$$

Thus, condition 6.6 holds.

By inspection of the proof of Theorem 6.1 in Newey (1994), condition 6.4(ii) implies 5.1(ii) therein, conditions 6.5 and 6.2 imply 5.2 therein, and condition 6.6 implies 5.3 therein. Then, conditions 5.1-5.3 inNewey (1994) hold, and thus by his Lemma 5.1,

$$\frac{1}{\sqrt{n}}\sum_i g(Z_i, \delta_{10}, \hat{\mu}) = \frac{1}{\sqrt{n}}\sum_i [g(Z_i, \mu_0) + \alpha(Z_i)] + o_P(1) \to_d \mathcal{N}(0, V)$$

where

$$\alpha(Z) = \delta(D_1)\begin{bmatrix}\Delta Y(1-S) - \mu_{10}(D_1)\\ (1-S) - \mu_{20}(D_1)\end{bmatrix} = -\frac{E\left(\frac{S}{\Delta D}\middle| D_1\right)}{E[1-S|D_1]}(1-S)(\Delta Y - \mu_0(D_1))$$

and $V = E\left[(g(Z_i, \mu_0) + \alpha(Z_i))\,(g(Z_i, \mu_0) + \alpha(Z_i))'\right]$. Finally note that:

$$\sqrt{n}(\hat{\delta}_1 - \delta_{10}) = \frac{n}{\sum_i S_i}\cdot\frac{1}{\sqrt{n}}\sum_i g(Z_i, \delta_{10}, \hat{\mu}) = \frac{1}{E[S]}\cdot\frac{1}{\sqrt{n}}\sum_i [g(Z_i, \mu_0) + \alpha(Z_i)] + o_P(1)$$

and the result follows defining $\psi_1 = [g(Z_i, \mu_0) + \alpha(Z_i)]/E[S]$. □

7.4 Theorem 4

We only prove the first point, as the proof of the second point is similar and (9)-(10) follow by combining these two points. Moreoer, the proof of (5) is similar to the proof of Theorem 1 so it is omitted. We thus focus on (6) hereafter.

For all $d_1 \in \text{Supp}(D_1|S_+ = 1)$, by Point 1 of Assumption 7, $d_1 \in \text{Supp}(D_1|S = 0)$. Thus, $E(\Delta Y|D_1 = d_1, S = 0)$ is well-defined. Then, using the same reasoning as that used to show (14) above, we obtain

$$E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, S_+ = 1) = E(\Delta Y|D_1 = d_1, S = 0).$$

Now, let $\text{Supp}(D_1|S_+ = 1)^c$ be the complement of $\text{Supp}(D_1|S_+ = 1)$. For all $d_1 \in \text{Supp}(D_1|S = 0) \cap \text{Supp}(D_1|S_+ = 1)^c$, $P(S_+ = 1|D_1 = d_1) = 0$. Then, with the convention that $E(\Delta Y|D_1 = d_1, S_+ = 1)P(S_+ = 1|D_1 = d_1) = 0$,

$$\begin{aligned}&E(\Delta Y|D_1 = d_1, S = 0)P(S_+ = 1|D_1 = d_1)\\ =&E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, S_+ = 1)P(S_+ = 1|D_1 = d_1).\end{aligned}$$

Combining the two preceding displays implies that for all $d_1 \in \text{Supp}(D_1|S = 0)$,

$$\begin{aligned}&E(\Delta Y|D_1 = d_1, S = 0)P(S_+ = 1|D_1 = d_1)\\ =&E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, S_+ = 1)P(S_+ = 1|D_1 = d_1).\end{aligned}$$

Hence, by repeated use of the law of iterated expectation,

$$\begin{aligned}
&E\left(\Delta Y \frac{P(S_+=1|D_1)}{P(S=0|D_1)}\frac{P(S=0)}{P(S_+=1)}\middle| S=0\right)\\
=&E\left(E[Y_2(D_1)-Y_1(D_1)|D_1,S_+=1]\frac{P(S_+=1|D_1)}{P(S=0|D_1)}\frac{P(S=0)}{P(S_+=1)}\middle| S=0\right)\\
=&E\left(E[Y_2(D_1)-Y_1(D_1)|D_1,S_+=1]\frac{P(S_+=1|D_1)}{P(S=0|D_1)}\frac{1-S}{P(S_+=1)}\right)\\
=&E\left(E[Y_2(D_1)-Y_1(D_1)|D_1,S_+=1]\frac{P(S_+=1|D_1)}{P(S_+=1)}\right)\\
=&E\left(E[Y_2(D_1)-Y_1(D_1)|D_1,S_+=1]\frac{S_+}{P(S_+=1)}\right)\\
=&E\left(Y_2(D_1)-Y_1(D_1)|S_+=1\right).
\end{aligned}$$

The result follows after some algebra. □

7.5 Theorem 5

We prove the result for the propensity-score-based estimator and drop the "ps" subscript to reduce notation. Let $\mu_1(d)=E[S_+|D_1=d]$, $\mu_2(d)=E[1-S|D_1=d]$, $\mu_3(d)=E[S_-|D_1=d]$ and $\mu_Y(D_1)=E[\Delta Y(1-S)|D_1]$. The logit series estimators of the unknown functions $\mu_j(d)$ are given by $\hat{\mu}_j(d)=\Lambda(P^K(d)'\hat{\pi}_j)$ where $\Lambda(z)=1/(1+e^{-z})$ is the logit function and

$$0=\sum_i(S_{ji}-\Lambda(P^K(D_{1i})'\hat{\pi}_j))P^K(D_{1i})$$

for S_{ji} equal to $1-S_i$, S_{i+} or S_{i-}. Under Assumption 8, there exists a constant $\pi_{j,K}$ that satisfies:

$$\left\|\log\left(\frac{\mu_j}{1-\mu_j}\right)-(P^K)'\pi_{j,K}\right\|_\infty=O(K^{-\alpha})$$

and we let $\mu_{ji,K}=\Lambda(P^K(D_{1i})'\pi_{j,K})$. We suppress the n subscript on K to reduce notation and let $\mu_{ji}:=\mu_j(D_{1i})$ and $\hat{\mu}_{ji}:=\hat{\mu}_j(D_{1i})$. Under Assumption 8 part 1, Lemma A.15 in Newey (1995) ensures that the smallest eigenvalue of $E[P^K(D_1)P^K(D_1)']$, is bounded away from zero uniformly over K. In addition, Cattaneo (2010) shows that under Assumption 8, the multinomial logit series estimator satisfies:

$$\|\mu_{j,K}-\mu_j\|_\infty=O(K^{-\alpha}),\quad \|\hat{\pi}_j-\pi_{j,K}\|=O_P\left(\sqrt{\frac{K}{n}}+K^{-\alpha+1/2}\right)$$

and

$$\|\hat{\mu}_j-\mu_j\|_\infty=O_P\left(\zeta(K)\left(\sqrt{\frac{K}{n}}+K^{-\alpha+1/2}\right)\right)$$

where $\zeta(K) = \sup_{d \in I} \left\| P^K(d) \right\|$. Newey (1994) also shows that for orthonormal polynomials, $\zeta(K)$ is bounded above by CK for some constant C, which implies in our case that $\|\hat{\mu}_j - \mu_j\|_\infty = O_P\left(K\left(\sqrt{\frac{K}{n}} + K^{-\alpha+1/2}\right)\right)$. Throughout the proof, we also use the fact that by a second-order mean value expansion, there exists a $\tilde{\pi}_j$ such that:

$$\begin{aligned} \hat{\mu}_{ji} - \mu_{ji,K} &= \Lambda(P^K(D_{1i})'\hat{\pi}_j) - \Lambda(P^K(D_{1i})'\pi_{j,K}) \\ &= \dot{\Lambda}(P^K(D_{1i})'\pi_{j,K})P^K(D_{1i})'(\hat{\pi}_j - \pi_{j,K}) + \ddot{\Lambda}(P^K(D_{1i})'\tilde{\pi}_j)(P^K(D_{1i})'(\hat{\pi}_j - \pi_{j,K}))^2 \end{aligned}$$

where both $\dot{\Lambda}(z)$ and $\ddot{\Lambda}(z)$ are bounded.

We start by considering the δ_{2+} parameter and omit the "ps" superscript to reduce notation. Recall that

$$\hat{\delta}_{2+} = \frac{1}{\sum_i \Delta D_i S_{i+}} \sum_i \left\{ \Delta Y_i S_{i+} - \Delta Y_i (1 - S_i) \frac{\hat{\mu}_{1i}}{\hat{\mu}_{2i}} \right\}.$$

Thus,

$$\sqrt{n}(\hat{\delta}_{2+} - \delta_{2+}) = \frac{1}{E[\Delta D S_+]} \cdot \frac{1}{\sqrt{n}} \sum_i \left\{ \Delta Y_i S_{i+} - \Delta Y_i (1 - S_i) \frac{\hat{\mu}_{1i}}{\hat{\mu}_{2i}} - \delta_{2+} E[\Delta D S_+] \right\} + o_P(1).$$

Define:

$$V_i = \Delta Y_i S_{i+} - \Delta Y_i (1 - S_i) \frac{\hat{\mu}_{1i}}{\hat{\mu}_{2i}} - \delta_{2+} E[\Delta D S_+].$$

Let $\psi_{2+,i}$ be the influence function defined in the statement of the theorem. Using the identity:

$$\frac{1}{\hat{b}} - \frac{1}{b} = -\frac{1}{b^2}(\hat{b} - b) + \frac{1}{b^2 \hat{b}}(\hat{b} - b)^2$$

we have, after some rearranging,

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_i V_i &= E[\Delta D S_+] \cdot \frac{1}{\sqrt{n}} \sum_i \psi_{2+,i} \\ &\quad - \frac{1}{\sqrt{n}} \sum_i \left(\frac{\Delta Y_i (1 - S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}} \right) (\hat{\mu}_{1i} - \mu_{1i}) \\ &\quad + \frac{1}{\sqrt{n}} \sum_i (\Delta Y_i (1 - S_i) - \mu_{Yi}) \frac{\mu_{1i}}{\mu_{2i}^2} (\hat{\mu}_{2i} - \mu_{2i}) \\ &\quad - \frac{1}{\sqrt{n}} \sum_i \Delta Y_i (1 - S_i) \frac{\mu_{1i}}{\mu_{2i}^2 \hat{\mu}_{2i}} (\hat{\mu}_{2i} - \mu_{2i})^2 \\ &\quad + \frac{1}{\sqrt{n}} \sum_i \frac{\Delta Y_i (1 - S_i)}{\mu_{2i}^2} (\hat{\mu}_{1i} - \mu_{1i})(\hat{\mu}_{2i} - \mu_{2i}) \\ &\quad - \frac{1}{\sqrt{n}} \sum_i \frac{\Delta Y_i (1 - S_i)}{\mu_{2i}^2 \hat{\mu}_{2i}} (\hat{\mu}_{1i} - \mu_{1i})(\hat{\mu}_{2i} - \mu_{2i})^2 \\ &\quad + \frac{1}{\sqrt{n}} \sum_i \frac{\mu_{Yi}}{\mu_{2i}} (S_{i+} - \hat{\mu}_{1i}) \\ &\quad - \frac{1}{\sqrt{n}} \sum_i \frac{\mu_{Y_i} \mu_{1i}}{\mu_{2i}^2} (1 - S_i - \hat{\mu}_{2i}). \end{aligned}$$

which we rewrite as:

$$\frac{1}{\sqrt{n}}\sum_i V_i = E[\Delta D S_+]\cdot\frac{1}{\sqrt{n}}\sum_i \psi_{2+,i} + \sum_{j=1}^{7} A_{j,n}$$

where each $A_{j,n}$ represents one term on the above display. We now bound each one of these terms.

Term 1. For the first term, we have that:

$$\begin{aligned}
-A_{1,n} &= \frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)(\hat{\mu}_{1i}-\mu_{1i})\\
&= \frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)(\hat{\mu}_{1i}-\mu_{1i,K})\\
&+ \frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)(\mu_{1i,K}-\mu_{1i})\\
&= A_{11,n} + A_{12,n}.
\end{aligned}$$

Now, by a second-order mean value expansion,

$$\begin{aligned}
A_{11,n} &= \frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)\dot{\Lambda}(P^K(D_{1i})'\pi_{j,K})P^K(D_{1i})'(\hat{\pi}_K-\pi_K)\\
&+ \frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)\ddot{\Lambda}(P^K(D_{1i})'\tilde{\pi})(P^K(D_{1i})'(\hat{\pi}_K-\pi_K))^2\\
&= A_{111,n} + A_{112,n}.
\end{aligned}$$

Next note that

$$|A_{111,n}| \le \|\hat{\pi}_K-\pi_K\|\left\|\frac{1}{\sqrt{n}}\sum_i\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)\dot{\Lambda}(P^K(D_{1i})'\pi_{j,K})P^K(D_{1i})'\right\|.$$

Now, $\|\hat{\pi}_K-\pi_K\| = O_P\left(\left(\sqrt{K/n}+K^{-\alpha+1/2}\right)\right)$. Let

$$U_i = (U_i^1,...U_i^K)' := \left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)\dot{\Lambda}(P^K(D_{1i})'\pi_{j,K})P^K(D_{1i})'.$$

We have $E[U_i]=E[E[U_i|D_{1i}]]=0$ and

$$\begin{aligned}
E\left[\|U_i\|^2\right] \le & E\left[\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)^2\left\|P^K(D_{1i})\right\|^2\right]\\
\le & CE\left[\left\|P^K(D_{1i})\right\|^2\right]\\
= & CE\left[\text{trace}\{P^K(D_{1i})'P^K(D_{1i})\}\right]\\
= & C\times\text{trace}\left(E\left[P^K(D_{1i})P^K(D_{1i})'\right]\right)\\
= & CK, \qquad (17)
\end{aligned}$$

since the polynomials can be chosen such that $E\left[P^K(D_{1i})P^K(D_{1i})'\right] = I_K$, see Newey (1997), page 161. Hence,

$$\begin{aligned}
E\left[\left\|\frac{1}{\sqrt{n}}\sum_i U_i\right\|^2\right] =& E\left[\sum_{j=1}^{K}\left(\frac{1}{\sqrt{n}}\sum_i U_i^j\right)^2\right] \\
=& \sum_{j=1}^{K}\frac{1}{n}\sum_{i,i'} E\left[U_i^j U_{i'}^j\right] \\
=& \sum_{j=1}^{K}\frac{1}{n}\sum_{i=1}^{n} E\left[U_i^{j2}\right] \\
=& E\left[\|U_1\|^2\right].
\end{aligned}$$

Therefore, by Markov's inequality,

$$A_{111,n} = O_P\left(K^{1/2}\left(\sqrt{\frac{K}{n}} + K^{-\alpha+1/2}\right)\right).$$

Next,

$$\begin{aligned}
|A_{112,n}| &\leq C\sqrt{n}\,\|\hat{\pi}_K - \pi_K\|^2 \frac{1}{n}\sum_i \left|\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right| \left\|P^K(D_{1i})\right\|^2 \\
&= O_P\left[\sqrt{n}\left(\frac{K}{n} + K^{-2\alpha+1}\right) E\left(\left|\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right| \left\|P^K(D_{1i})\right\|^2\right)\right] \\
&= O_P\left(\sqrt{n}K\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right),
\end{aligned}$$

where the first inequality follows by Cauchy-Schwarz inequality, the second by Markov's inequality and the third by the same reasoning as to obtain (17). Hence,

$$A_{11,n} = O_P\left(K^{1/2}\left(\sqrt{\frac{K}{n}} + K^{-\alpha+1/2}\right)\right) + O_P\left(\sqrt{n}K\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right).$$

Finally, for $A_{12,n}$ we have that

$$E\left[\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)(\mu_{1i,K} - \mu_{1i})\right| D_1\left.\vphantom{\frac{1}{1}}\right] = 0$$

and

$$E\left[\left\|\left(\frac{\Delta Y_i(1-S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}}\right)(\mu_{1i,K} - \mu_{1i})\right\|^2\right] \leq C\,\|\mu_{1,K} - \mu_1\|_\infty^2 = O(K^{-2\alpha})$$

and therefore

$$A_{1,n} = O_P\left(K^{1/2}\left(\sqrt{\frac{K}{n}} + K^{-\alpha+1/2}\right)\right) + O_P\left(\sqrt{n}K\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right) + O_P(K^{-\alpha}).$$

Term 2. This follows by the same argument as that of Term 1 and we obtain:

$$A_{2,n} = O_P\left(K^{1/2}\left(\sqrt{\frac{K}{n}} + K^{-\alpha+1/2}\right)\right) + O_P\left(\sqrt{n}K\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right) + O_P(K^{-\alpha}).$$

Term 3. For the third term, since μ_{2i} is uniformly bounded and $\hat{\mu}_2$ converges uniformly to μ_2, for n large enough

$$|A_{3,n}| \leq \sqrt{n}\,\|\hat{\mu}_2 - \mu_2\|_\infty^2 \frac{1}{C}\frac{1}{n}\sum_i |\Delta Y_i(1 - S_i)| = O_P\left(\sqrt{n}K^2\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right).$$

Term 4. For the fourth term,

$$|A_{4,n}| \leq \sqrt{n}\,\|\hat{\mu}_1 - \mu_1\|_\infty \|\hat{\mu}_2 - \mu_2\|_\infty \frac{1}{C}\frac{1}{n}\sum_i |\Delta Y_i(1 - S_i)| = O_P\left(\sqrt{n}K^2\left(\frac{K}{n} + K^{-2\alpha+1}\right)\right)$$

Term 5. For the fifth term, since μ_{2i} is uniformly bounded and $\hat{\mu}_2$ converges uniformly to μ_2, for n large enough

$$|A_{5,n}| \leq \sqrt{n}\,\|\hat{\mu}_1 - \mu_1\|_\infty \|\hat{\mu}_2 - \mu_2\|_\infty^2 \frac{1}{C}\frac{1}{n}\sum_i |\Delta Y_i(1 - S_i)| = O_P\left(\sqrt{n}K^3\left(\left(\frac{K}{n}\right)^{3/2} + K^{-3\alpha+3/2}\right)\right).$$

Term 6. For the sixth term, let $\gamma_{6,K}$ be the population coefficient from a (linear) series approximation to the function $\mu_Y(D_1)/\mu_2(D_1)$. Then we have that

$$\begin{aligned} A_{6,n} &= \frac{1}{\sqrt{n}}\sum_i \left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(S_{i+} - \hat{\mu}_{1i}) + \frac{1}{\sqrt{n}}\sum_i P^K(D_{1i})'\gamma_{6,K}(S_{i+} - \hat{\mu}_{1i}) \\ &= \frac{1}{\sqrt{n}}\sum_i \left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(S_{i+} - \hat{\mu}_{1i}) \end{aligned}$$

because the last term in the second line equals zero by the first-order conditions of the logit series estimator. Next, we have that

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_i \left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(S_{i+} - \hat{\mu}_{1i}) &= \frac{1}{\sqrt{n}}\sum_i \left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(S_{i+} - \mu_{1i}) \\ &\quad - \frac{1}{\sqrt{n}}\sum_i \left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(\hat{\mu}_{1i} - \mu_{1i}) \\ &= A_{61,n} + A_{62,n}. \end{aligned}$$

Now, for $A_{61,n}$, we have that

$$E\left[\left(\frac{\mu_{Yi}}{\mu_{2i}} - P^K(D_{1i})'\gamma_{6,K}\right)(S_{i+} - \mu_{1i})\middle| D_1\right] = 0$$

and

$$E\left[(S_{i+}-\mu_{1i})^2\left\|\left(\frac{\mu_{Yi}}{\mu_{2i}}-P^K(D_{1i})'\gamma_{6,K}\right)\right\|^2\right]\leq O(K^{-2\alpha})$$

so that

$$A_{61,n}=O_P(K^{-\alpha}).$$

On the other hand, for $A_{62,n}$, we have that

$$|A_{62,n}|\leq\sqrt{n}\left\|\frac{\mu_Y}{\mu_2}-(P^K)'\gamma_{6,K}\right\|_\infty\|\hat{\mu}_1-\mu_1\|_\infty=O_P\left(\sqrt{n}K^{1-\alpha}\left(\sqrt{\frac{K}{n}}+K^{-\alpha+1/2}\right)\right)$$

from which

$$A_{6,n}=O_P\left(\sqrt{n}K^{1-\alpha}\left(\sqrt{\frac{K}{n}}+K^{-\alpha+1/2}\right)+K^{-\alpha}\right).$$

Term 7. This follows by the same argument as that of Term 6 and we obtain

$$A_{7,n}=O_P\left(\sqrt{n}K^{1-\alpha}\left(\sqrt{\frac{K}{n}}+K^{-\alpha+1/2}\right)+K^{-\alpha}\right).$$

Collecting all the terms, if follows that under the conditions

$$\frac{K^6}{n}\to 0,\quad \frac{K^{4\alpha-6}}{n}\to\infty,\quad \alpha>3$$

we obtain

$$\sqrt{n}(\hat{\delta}_{2+}-\delta_{2+})=\frac{1}{\sqrt{n}}\sum_i\psi_{2+,i}+o_P(1).$$

Setting $\alpha=4$, this implies

$$\frac{K^6}{n}\to 0,\quad \frac{K^{10}}{n}\to\infty.$$

These conditions are satisfied when $K=n^\nu$ for $1/(4\alpha-6)<\nu<1/6$ or in this case $1/10<\nu<1/6$.

By an analogous argument, we can show that under the same conditions

$$\sqrt{n}(\hat{\delta}_{2-}-\delta_{2-})=\frac{1}{\sqrt{n}}\sum_i\psi_{2-,i}+o_P(1)$$

and the result follows by a multivariate CLT. Finally, notice that letting $\mu_{1-}(d)=E[S_-|D_1=d]$ and $\hat{\mu}_{ji-}=\hat{\mu}_{1-}(D_{1i})$, and using that $\text{sgn}(\Delta D_i)=S_{i+}-S_{i-}$, after some simple manipulations:

$$\hat{\delta}_2=\frac{1}{\sum_i|\Delta D_i|}\sum_i\left\{\Delta Y_i(S_{i+}-S_{i-})-\Delta Y_i(1-S_i)\left(\frac{\hat{\mu}_{1i}-\hat{\mu}_{1i-}}{\hat{\mu}_{2i}}\right)\right\}$$

which is analogous to $\hat{\delta}_{2+}$ replacing S_{i+} by $(S_{i+}-S_{i-})$ and the denominator by $\sum_i|\Delta D_i|$. Thus, under the same conditions

$$\sqrt{n}(\hat{\delta}_2-\delta_2)=\frac{1}{\sqrt{n}}\sum_i\psi_{2,i}+o_P(1)$$

where $\psi_{2,i}$ is defined in the statement of the theorem. □

7.6 Theorem 6

First, remark that

$$\delta_2 = \frac{E[\text{sgn}(\Delta D)(\Delta Y - (Y_2(D_1) - Y_1(D_1)))]}{E[|\Delta D|]}.$$

Thus, it suffices to show that a.s.,

$$\lim_{\eta \downarrow 0} E\left(\Delta Y | D_1, S_\eta = 0\right) = E\left(Y_2(D_1) - Y_1(D_1) | D_1, D_2\right). \tag{18}$$

Fix $\eta > 0$. By Assumption 9, $P(S_\eta = 0|D_1) > 0$. Thus, $E\left(\Delta Y | D_1, S_\eta = 0\right)$ is well-defined. Moreover,

$$\begin{aligned} E\left(\Delta Y | D_1, S_\eta = 0\right) =& E\left(Y_2(D_2) - Y_2(D_1) | D_1, S_\eta = 0\right) \\ &+ E\left(Y_2(D_1) - Y_1(D_1) | D_1, S_\eta = 0\right). \end{aligned} \tag{19}$$

Now, by Jensen's inequality and Point 2 of Assumption 2,

$$\begin{aligned} \left|E\left(Y_2(D_2) - Y_2(D_1) | D_1, S_\eta = 0\right)\right| \leq & E\left(|Y_2(D_2) - Y_2(D_1)| \; | D_1, S_\eta = 0\right) \\ \leq & E\left(\overline{Y} |D_2 - D_1| \; | D_1, S_\eta = 0\right) \\ \leq & \eta E\left[\sup_{(d_1, d_2) \in \text{Supp}(D_1, D_2)} E\left(\overline{Y} | D_1 = d_1, D_2 = d_2\right) | D_1, S_\eta = 0\right] \\ \leq & \overline{K} \eta \end{aligned} \tag{20}$$

for some $\overline{K} < \infty$. Next, by Assumption 1,

$$\begin{aligned} E\left(Y_2(D_1) - Y_1(D_1) | D_1, S_\eta = 0\right) &= E\left(Y_2(D_1) - Y_1(D_1) | D_1\right) \\ &= E\left(Y_2(D_1) - Y_1(D_1) | D_1, D_2\right). \end{aligned}$$

Combined with (19)-(20), this yields (18) □

7.7 Theorem 7

Using the same steps as in the proof of Theorem 1, one can show that for all $t \geq 2$,

$$\delta_{1t} = E\left(\frac{Y_t - Y_{t-1} - E(Y_t - Y_{t-1} | D_{t-1}, S_t = 0)}{D_t - D_{t-1}} \middle| S_t = 1\right).$$

This proves the result □

7.8 Theorem 8

Using the same steps as in the proof of Theorem 1, one can show that for all $t \geq 2$,

$$\delta_{2+t} = \frac{E\left(Y_t - Y_{t-1}|S_{+,t}=1\right) - E\left((Y_t - Y_{t-1})\frac{P(S_{+,t}=1|D_{t-1})}{P(S_t=0|D_{t-1})}\frac{P(S_t=0)}{P(S_{+,t}=1)}\Big|S_t=0\right)}{E(D_t - D_{t-1}|S_{+,t}=1)},$$

$$\delta_{2-t} = \frac{E\left(Y_t - Y_{t-1}|S_{-,t}=1\right) - E\left((Y_t - Y_{t-1})\frac{P(S_{-,t}=1|D_{t-1})}{P(S_t=0|D_{t-1})}\frac{P(S_t=0)}{P(S_{-,t}=1)}\Big|S_t=0\right)}{E(D_{t-1} - D_t|S_{-,t}=1)}.$$

This proves the result □

7.9 Theorem 9

We start by proving (11). For all $t \leq T - \ell + 1$, for every d_1 in the support of $D_1|F=t$,

$$\begin{aligned} E(Y_{t-1+\ell} - Y_{t-1}|D_1 = d_1, F > t-1+\ell) =& E(Y_{t-1+\ell}(\boldsymbol{d}_1) - Y_{t-1}(\boldsymbol{d}_1)|D_1 = d_1) \\ =& E(Y_{t-1+\ell}(\boldsymbol{d}_1) - Y_{t-1}(\boldsymbol{d}_1)|D_1 = d_1, F = t). \qquad (21) \end{aligned}$$

It follows from Assumption 15 that d_1 belongs to the support of $D_1|F=t$, so $E(Y_{t-1+\ell} - Y_{t-1}|D_1 = d_1, F > t-1+\ell)$ is well defined. The first and second equalities follow from Assumptions 13 and 14 and the law of iterated expectations.

Then,

$$\begin{aligned} & E\left(Y_{t-1+\ell} - Y_{t-1} - E(Y_{t-1+\ell} - Y_{t-1}|D_1, F > t-1+\ell)|F=t\right) \\ =& E\left(Y_{t-1+\ell} - Y_{t-1} - E(Y_{t-1+\ell}(\boldsymbol{D}_1) - Y_{t-1}(\boldsymbol{D}_1)|D_1, F=t)|F=t\right) \\ =& E\left(Y_{t-1+\ell}(\boldsymbol{D}) - Y_{t-1}(\boldsymbol{D}_1) - E(Y_{t-1+\ell}(\boldsymbol{D}_1) - Y_{t-1}(\boldsymbol{D}_1)|D_1, F=t)|F=t\right) \\ =& E\left(E(Y_{t-1+\ell}(\boldsymbol{D}) - Y_{t-1}(\boldsymbol{D}_1) - Y_{t-1+\ell}(\boldsymbol{D}_1) + Y_{t-1}(\boldsymbol{D}_1)|D_1, F=t)|F=t\right) \\ =& E\left(Y_{t-1+\ell}(\boldsymbol{D}) - Y_{t-1+\ell}(\boldsymbol{D}_1)|F=t\right). \end{aligned}$$

The first equality follows from (21). The second equality follows from the definition of F and Assumption 13. The third and fourth equalities follow from the law of iterated expectations. This proves the result □

We now prove (12). It follows from the definition of F and Assumption 13 that

$$E\left(Y_{t-1+\ell} - Y_{t-1}|F=t\right) = E\left(Y_{t-1+\ell}(\boldsymbol{D}) - Y_{t-1}(\boldsymbol{D}_1)|F=t\right).$$

Then,

$$\begin{aligned} & E\left((Y_{t-1+\ell} - Y_{t-1})\frac{P(F=t|D_1)}{P(F>t-1+\ell|D_1)}\frac{P(F>t-1+\ell)}{P(F=t)}\Bigg|F>t-1+\ell\right) \\ =& E\left(E(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|D_1, F=t)\frac{P(F=t|D_1)}{P(F>t-1+\ell|D_1)}\frac{P(F>t-1+\ell)}{P(F=t)}\Bigg|F>t-1+\ell\right) \\ =& E\left(Y_t(\boldsymbol{D}_1) - Y_{t-\ell-1}(\boldsymbol{D}_1)|F=t\right). \end{aligned}$$

The first equality follows from the law of iterated expectations and (21). The second equality follows from the same steps as in the proof of Theorem 4. Combining the two previous displays proves the result □

References

Abadie, A. (2005, 01). Semiparametric difference-in-differences estimators. *Review of Economic Studies 72*(1), 1–19.

Abbring, J. H. and G. J. Van den Berg (2003). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–1517.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020.

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

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bojinov, I., A. Rambachan, and N. Shephard (2021). Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics 12*(4), 1171–1196.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. arXiv preprint arXiv:2108.12419.

Botosaru, I. and F. H. Gutierrez (2018). Difference-in-differences when the treatment status is observed in only one period. *Journal of Applied Econometrics 33*(1), 73–90.

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2021). Difference-in-differences with a continuous treatment. arXiv preprint arXiv:2107.02637.

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

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

de Chaisemartin, C. and X. D'Haultfœuille (2018). Fuzzy differences-in-differences. *The Review of Economic Studies 85*(2), 999–1028.

de Chaisemartin, C. and X. D'Haultfœuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

de Chaisemartin, C. and X. D'Haultfœuille (2022). Difference-in-differences estimators of intertemporal treatment effects. NBER Working paper 29873.

de Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal Forthcoming.*

Deschênes, O. and M. Greenstone (2007). The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *American economic review 97*(1), 354–385.

D'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2021). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. arXiv preprint arXiv:2104.14458.

Fajgelbaum, P. D., P. K. Goldberg, P. J. Kennedy, and A. K. Khandelwal (2020). The return to protectionism. *The Quarterly Journal of Economics 135*(1), 1–55.

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

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

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

Linton, O. and J. P. Nielsen (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika 82*(1), 93–100.

Malani, A. and J. Reif (2015). Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics 124*, 1–17.

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

Newey, W. K. (1995). Convergence rates for series estimators. In G. Maddala, P. Phillips, and T. Srinivasan (Eds.), *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*. Basil Blackwell.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics 79*(1), 147–168.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical modelling 7*(9-12), 1393–1512.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*, 175–199.