Difference-in-Differences for Continuous Treatments and Instruments with Stayers

Clément de Chaisemartin, Xavier d'Haultfoeuille, Félix Pasquier, Doulo Sow, Gonzalo Vazquez-Bare

Difference-in-Differences for Continuous Treatments and Instruments with Stayers*

Clément de Chaisemartin† Xavier D'Haultfœuille‡ Félix Pasquier§
Doulo Sow¶ Gonzalo Vazquez-Bare‖

First version: January 18, 2022. This version: January 10, 2025

Abstract

We propose difference-in-differences estimators in designs where the treatment is continuously distributed at every period, as is often the case when one studies the effects of taxes, tariffs, or prices. We assume that between consecutive periods, the treatment of some units, the switchers, changes, while the treatment of other units remains constant. We show that under a placebo-testable parallel-trends assumption, averages of the slopes of switchers' potential outcomes can be nonparametrically estimated. We generalize our estimators to the instrumental-variable case. We use our estimators to estimate the price-elasticity of gasoline consumption.

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

JEL Codes: C21, C23

*We are very grateful to Matias Cattaneo, Joachim Freyberger, Basile Grassi, Daniel Herrera, Thomas Le Barbanchon, Thierry Mayer, Isabelle Méjean, Andres Santos and seminar participants at AgroParisTech, Bank of Portugal, Bocconi University, Bonn University, CERDI, GAEL, IFAU, the International Panel Data Conference, the IZA-CREST Conference on Labor Market Policy Evaluation, Michigan University, Northwestern University, Reading University, Sciences Po, the 2023 North American Summer Meeting of the Econometric Society, 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). Views and opinions expressed are those of the authors and do not reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

†Sciences Po Paris, clement.dechaisemartin@sciencespo.fr
‡CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr.
§CREST-ENSAE, felix.pasquier@ensae.fr.
¶Sciences Po, doulo.sow@sciencespo.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 estimate a two-way fixed effects (TWFE) regression that controls for unit and time fixed effects:

$$Y_{i,t} = \alpha_i + \gamma_t + \beta_{TWFE} D_{i,t} + u_{i,t},$$

where $D_{i,t}$ is the treatment of unit i at time t. de Chaisemartin and D'Haultfœuille (2023a) find that 26 of the 100 most cited papers published by the American Economic Review from 2015 to 2019 have estimated at least one TWFE regression. de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), and Borusyak et al. (2024) have shown that under a parallel trends assumption, TWFE regressions are not robust to heterogeneous effects: they may estimate a weighted sum of treatment effects across periods and units, with some negative weights. Owing to the negative weights, β_{TWFE} could be, say, negative, even if the treatment effect is positive for every unit $\times$ period. Importantly, the result in de Chaisemartin and D'Haultfœuille (2020) applies to binary, discrete, and continuous treatments.

Several alternative heterogeneity-robust difference-in-difference (DID) estimators have been proposed (see Table 2 of de Chaisemartin and D'Haultfœuille, 2023b). Some apply to binary and staggered treatments (see Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024). Some 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, 2023a; Callaway et al., 2021; de Chaisemartin and D'Haultfœuille, 2024). However, treatments continuously distributed at every period, including the first one in the data, are ubiquitous in applied work. For instance, taxes (see Li et al., 2014) or tariffs (see Fajgelbaum et al., 2020) are often continuously distributed at all periods. No difference-in-difference estimator robust to heterogeneous effects is available for such designs. Proposing such estimators is the purpose of this paper.

We assume that we have a panel data set, whose units could be geographical locations such as counties. We start by considering the case where the panel has two time periods. 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. We consider two target parameters. The first one is the average slope of switchers' period-two potential outcome function, from their period-one to their period-two treatment, hereafter referred to as the Average of Slopes (AS). Our second target is a weighted average of switchers' slopes, where switchers receive a weight proportional to the absolute value of their treatment change, hereafter referred to as the Weighted Average of Slopes (WAS). We propose a novel parallel trends assumption on the outcome evolution of switchers and stayers with the same period-one treatment, in the counterfactual where switchers' treatment would not have

changed. Because it conditions on units' period-one treatment, this parallel-trends assumption does not impose any restriction on effect heterogeneity. This parallel trends assumption is also placebo testable, by comparing switchers' and stayers' outcome evolutions before switchers' treatment changes. We view the possibility of placebo-testing it as an important advantage of our assumption, as placebo tests are an essential step in establishing the credibility of an identifying assumption in observational studies (Imbens et al., 2001; Imbens and Xu, 2024). We show that under our placebo-testable parallel-trends assumption, the AS and the WAS are identified. This contrasts with other target parameters, like the dose-response function or the average marginal effect, which can only be identified under non-placebo-testable assumptions.

Economically, the AS and the WAS can serve different purposes, so neither parameter dominates the other. Under shape restrictions on the potential outcome function, the AS can be used to infer the effect of other treatment changes than those that took place from period one to two. Instead, the WAS can be used to conduct a cost-benefit analysis of the treatment changes that effectively took place. On the other hand, when it comes to estimation, the WAS unambiguously dominates the AS. First, we show that it can be estimated at the standard parametric rate even if switchers can experience an arbitrarily small change of their treatment between consecutive periods. Second, we show that under some conditions, the asymptotic variance of the WAS estimator is strictly lower than that of the AS estimator. Third, unlike the AS, the WAS is amenable to doubly-robust estimation, which comes with a number of advantages.

Then, we consider the instrumental-variable (IV) case. For instance, one may be interested in estimating the price-elasticity of a good's consumption. If prices respond to demand shocks, the counterfactual consumption trends of units experiencing and not experiencing a price change may not be parallel. On the other hand, the counterfactual consumption trends of units experiencing and not experiencing a tax change may be parallel. Then, taxes may be used as an instrument for prices. In such cases, we show that the reduced-form WAS effect of the instrument on the outcome divided by the first-stage WAS effect is equal to a weighted average of switchers' outcome-slope with respect to the treatment, where switchers with a larger first-stage effect receive more weight. Hereafter, we refer to this effect as the IV-WAS effect. The ratio of the reduced-form and first-stage AS effects is also equal to a weighted average of slopes, with arguably less natural weights, so in the IV case the WAS seems both economically and statistically preferable to the AS. Importantly, we show that the reduced-form parallel-trends assumption implicitly restricts treatment-effect heterogeneity. Such restrictions can be alleviated by controlling for groups baseline treatment in the IV specification, but the resulting estimator still restricts effects' heterogeneity across units.

We consider other extensions. First, we extend our results to applications with more than two time periods. Importantly, with several time periods our estimators rely on a parallel-trends assumption over consecutive periods, rather than over the entire duration of the panel. Second,

we propose a placebo estimator of the parallel-trends assumption underlying our estimators.

Finally, we use the yearly, 1966 to 2008 US state-level panel dataset of Li et al. (2014) to estimate the effect of gasoline taxes on gasoline consumption and prices. Using the WAS estimators, we find a significantly negative effect of taxes on gasoline consumption, and a significantly positive effect on prices. The AS estimators are close to, and not significantly different from, the WAS estimators, but they are also markedly less precise: their standard errors are almost three times larger than that of the WAS estimators. This shows that our theoretical result on the precision ranking of the WAS and AS estimators, which we derive under strong assumptions, still holds in a real-life application where those assumptions probably do not hold. The precision losses attached to using the AS have consequences. The AS estimator of the effect of taxes on prices is not statistically significant, so with that estimator one cannot use taxes as an instrument to estimate the price-elasticity of consumption, because the instrument does not have a first-stage. This contrasts with the WAS, whose first-stage t-stat is around 7. We compute an IV-WAS estimator of the price elasticity of gasoline consumption, and find a fairly small elasticity of -0.67, in line with previous literature (for instance, Hausman and Newey, 1995, find a long-run elasticity of -0.81). Our placebo estimators are small, insignificant, and fairly precisely estimated, thus suggesting that our parallel trends assumption is plausible.

Our estimators are computed by the `did_multiplegt_stat` Stata package, available from the SSC repository. Our package allows estimators with control variables and weights, see the help file and the package's companion paper for further details.

Related Literature.

Our paper builds upon several previous papers in the panel data literature. Chamberlain (1982) seems to be the first paper to have proposed an estimator of the AS parameter. Under the assumption of no counterfactual time trend, the estimator therein is a before-after estimator. Then, our paper is closely related to the work of Graham and Powell (2012), who also propose DID estimators of the AS (see their Equation (21)) when the treatment is continuously distributed at every time period. Their estimators rely on a linear effect assumption and assume that units experience the same evolution of their treatment effect over time, a parallel-trends-on-treatment-effects assumption. By contrast, our estimator of the AS does not place any restriction on treatment effects. But our main contribution to this literature is to introduce the WAS, and to contrast the pros and cons of the AS and WAS estimators. Our results are also related to Hoderlein and White (2012), who consider the average marginal effect of a continuous treatment with panel data. However, their target parameters and identifying assumptions are different. For instance, they rule out systematic changes of the outcome over time.

With respect to the aforementioned heterogeneity-robust DID literature, we make two contribu-

tions. First, in the non-IV case we propose estimators that can be used even if units' treatment varies at baseline. Thus we usefully complement previous literature, that has mostly focused on the case where all units have a baseline treatment equal to zero (see Sun and Abraham, 2021; Callaway and Sant'Anna, 2021; Borusyak et al., 2024; Callaway et al., 2021; de Chaisemartin and D'Haultfœuille, 2024). One exception predating this paper is de Chaisemartin and D'Haultfœuille (2020), who allow for a non-binary discrete treatment at baseline in their Web Appendix, and propose estimators comparing switchers and stayers with the same baseline treatment. However, that paper does not allow for continuously distributed treatments, and comparing switchers and stayers with the same baseline treatment is no longer feasible with a continuously distributed treatment.[1] Second, in the IV case, previous IV-DID literature has only considered classical designs with two periods and binary instrument and treatment (de Chaisemartin, 2010; Hudson et al., 2017), as well as fuzzy DID designs, a special case of IV-DIDs (de Chaisemartin and D'Haultfœuille, 2018). Instead, this paper proposes broadly applicable IV-DID estimators, and also highlights that IV-DID estimators impose restrictions on treatment effect-heterogeneity, which can be mitigated by controlling for the baseline treatment.

Importantly, our estimators require that there be some stayers, whose treatment does not change between consecutive periods. This assumption is unlikely to be met when the treatment is say, precipitations: for instance, US counties never experience the exact same precipitations over two consecutive years. In de Chaisemartin et al. (2023), we discuss the (non-trivial) extension of the results in this paper to applications without stayers.

Organization of the paper. In Section 2, we present the set-up, introduce notation and discuss our main assumptions. In Section 3, we introduce the AS and discuss its identification and estimation. Section 4 then turns to the WAS. Section 5 extends our previous results to an instrumental variable set-up. We consider other extensions in Section 6. Finally, our application is developed in Section 7. The proofs are collected in the appendix.

2 Set-up, assumptions, and building-block identification result

2.1 Set-up

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,

[1] Another exception, posterior to this paper, is de Chaisemartin and D'Haultfœuille (2023a), who extend the estimators in this paper to models with dynamic effects in Section 1.10 of their Web Appendix.

like a county or a region.[2] 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. 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.

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, and let Y_1 and Y_2 denote their observed outcomes at periods 1 and 2. Our potential outcome notation assumes that Y_1 does not depend on units' period-two treatment, thus ruling out anticipation effects, a commonly-made assumption in the DID literature. Our notation also assumes that Y_2 does not depend on units' period-one treatment, thus ruling out dynamic effects. When the treatment is continuously distributed at period one, allowing for dynamic effects opens up the so-called initial-conditions problem. As units receive heterogeneous doses at period one, they may have experienced treatment changes before period one. With dynamic effects such changes may still affect their outcome over the study period, but they cannot be accounted for because they are not observed. Ruling out dynamic effects allows us to abstract from this thorny issue, but could yield misleading results if dynamic effects are present. To alleviate this concern, in Section 6.3 we propose a modified version of our estimators, robust to dynamic effects up to a pre-specified treatment lag.

In what follows, all equalities and inequalities involving random variables are required to hold almost surely. Finally, for any random variable 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.

2.2 Assumptions

We make the following assumptions.[3]

Assumption 1 *(Parallel trends) For all $d_1 \in \mathcal{D}_1$, $E(\Delta Y(d_1)|D_1 = d_1, D_2) = E(\Delta Y(d_1)|D_1 = d_1)$.*

Assumption 1 is a parallel trends assumption, requiring that $\Delta Y(d_1)$ be mean independent of D_2, conditional on $D_1 = d_1$.

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

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

[2] 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 is a mechanical extension.

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

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 ensures 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 a unit-specific Lipschitz constant). This will automatically hold if $d \mapsto Y_2(d)$ is differentiable with respect to d and has a bounded derivative.

Finally, 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).

2.3 Building-block identification result

Assumption 1 implies the following lemma, our building-block identification result.

Lemma 1 *If Assumption 1 holds, then for all $(d_1, d_2) \in \mathcal{D}_1 \times \mathcal{D}_2$ such that $d_1 \neq d_2$ and $P(S|D_1 = d_1) < 1$,*

$$\begin{aligned} TE(d_1, d_2|d_1, d_2) :=& E\left(\frac{Y_2(d_2) - Y_2(d_1)}{d_2 - d_1}\middle| D_1 = d_1, D_2 = d_2\right) \\ =& E\left(\frac{\Delta Y - E(\Delta Y|D_1 = d_1, S = 0)}{d_2 - d_1}\middle| D_1 = d_1, D_2 = d_2\right). \end{aligned}$$

Proof:

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

where the second equality follows from Assumption 1. This proves the result □

Intuitively, under Assumption 1 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. If a unit's treatment changes from two to five, we can recover its counterfactual outcome evolution if its treatment had not changed, by using the average

outcome evolution of all stayers with a baseline treatment of two. Then, a DID estimand comparing switchers' and stayers' outcome evolutions identifies $E\left(Y_2(d_2) - Y_2(d_1)|D_1 = d_1, D_2 = d_2\right)$, and we can scale that effect by $d_2 - d_1$ to identify a slope rather than an unnormalized effect.

Note that in a canonical DID design where $\mathcal{D}_1 = 0$ and $\mathcal{D}_2 \in \{0, 1\}$, Lemma 1 only applies to $(d_1, d_2) = (0, 1)$, $\text{TE}(0, 1|0, 1)$ reduces to the ATT, and the estimand reduces to the canonical DID estimand comparing the outcome evolutions of treated and untreated units. Note also that in a design where all units are untreated at period one, $d_1 = 0$, and

$$\text{TE}(0, d_2|0, d_2) = E\left(\frac{Y_2(d_2) - Y_2(0)}{d_2}\middle| D_2 = d_2\right),$$

an effect closely related to the $ATT(d|d)$ effect in Callaway et al. (2021). Thus, the effects we consider are extensions of those effects to applications with a treatment continuous at all periods.

Lemma 1 shows that under Assumption 1,

$$(d_1, d_2) \mapsto \text{TE}(d_1, d_2|d_1, d_2)$$

is identified. Of course, one may be interested in other parameters, like

$$(d, d') \mapsto \text{TE}(d, d') := E\left(\frac{Y_2(d) - Y_2(d')}{d - d'}\right),$$

a function which, unlike $(d_1, d_2) \mapsto \text{TE}(d_1, d_2|d_1, d_2)$, applies to the entire population rather than to specific subpopulation that depend on (d_1, d_2). Alternatively, one could also be interested in the average marginal effect

$$E\left(Y_2'(D_2)\right).$$

What is the appeal of $\text{TE}(d_1, d_2|d_1, d_2)$ with respect to those other parameters? Conditional on $D_1 = d_1, D_2 = d_2$, $Y_2(d_2)$ is observed, so estimating $\text{TE}(d_1, d_2|d_1, d_2)$ only requires estimating $Y_2(d_1)$, switchers' counterfactual outcomes if their treatment had not changed. By definition, $Y_1(d_1)$ is observed. If the data contains a third period 0 and the treatment of some units does not change from period 0 to 1, then $Y_0(d_1)$ is also observed for some switchers and stayers. Then, as explained in further details in Section 6.2, one can placebo-test Assumption 1, by comparing the outcome evolutions of switchers and stayers from period 0 to 1. This shows that $\text{TE}(d_1, d_2|d_1, d_2)$ is identified under a placebo-testable parallel-trends assumption. On the other hand, estimating $\text{TE}(d, d')$ requires estimating, for most units, *two* unobserved counterfactual outcomes. This cannot be achieved under a placebo-testable assumption as we only observe *one* potential outcome at each date. When $D_1 = 0$, Callaway et al. (2021) propose a "strong parallel-trends" assumption under which the dose-response function, a parameter closely related to $\text{TE}(0, d')$ are identified, but their "strong parallel-trends" assumption is not placebo testable. Similarly, estimating $E\left(\frac{Y_2(D_2) - Y_2(d')}{D_2 - d'}\right)$ requires estimating $Y_2(d')$, which cannot be achieved under a placebo-testable

assumption because $Y_1(d')$ is not observed for all units. As $Y_2'(D_2) = \lim_{d' \to D_2} \frac{Y_2(D_2)-Y_2(d')}{D_2-d'}$, the same issue applies to $E\left(Y_2'(D_2)\right)$.

Variability in $\text{TE}(d_1, d_2|d_1, d_2)$ across values of (d_1, d_2) conflates a dose-response relationship that may be of economic interest, and a selection bias due to the fact that units with different period one and two treatments may have heterogeneous treatment effects (Callaway et al., 2021). Moreover, Lemma 1 shows that estimating $\text{TE}(d_1, d_2|d_1, d_2)$ requires estimating the values of two conditional expectations with respect to continuous variables, at points $D_1 = d_1, D_2 = d_2$ and $D_1 = d_1$. Unless one is willing to make parametric functional-form assumptions, the resulting estimator will converge at a slower rate than the standard $\sqrt{n}-$ parametric rate. For these two reasons, in this paper we focus on averages of the slopes $\text{TE}(d_1, d_2|d_1, d_2)$, that can be estimated non-parametrically at the standard $\sqrt{n}-$ parametric rate, and we do not focus on the function $(d_1, d_2) \mapsto \text{TE}(d_1, d_2|d_1, d_2)$.

Finally, our DID estimands compare switchers and stayers with the same period-one treatment. Instead, one could propose estimands comparing switchers and stayers, without conditioning on their period-one treatment. To recover the counterfactual outcome trend of a switcher going from two to five units of treatment, one could use a stayer with treatment equal to three at both dates. On top of Assumption 1, such estimands rest on two supplementary conditions:

(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: to rationalize a DID estimand comparing a switcher going from two to five units of treatment to a stayer with treatment equal to three, $E(\Delta Y(2))$ and $E(\Delta Y(3))$ should be equal. Rearranging, (ii) is equivalent to

$$E(Y_2(d) - Y_2(d')) = E(Y_1(d) - Y_1(d')) : \tag{1}$$

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.

3 Estimating the average of switchers' slopes

3.1 Target parameter

In this section, our target parameter is

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

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 Slopes (AS).

The AS is a local effect: it only applies to switchers, and 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 AS can be used to point or partially identify the effect of other treatment changes under shape restrictions. 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, $\delta_1 = E\left(B_2|S = 1\right)$: the AS 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 AS is sufficient to recover the ATE of any uniform treatment change among switchers. 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 $\epsilon > 0$,

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

Accordingly, under convexity one can use the AS to obtain lower bounds of the effect of changing the treatment from D_2 to larger values than D_2. For instance, in Fajgelbaum et al. (2020), one can use this strategy to derive a lower bound of the effect of increasing tariffs' to even higher levels than those decided by the Trump administration. Under convexity, one can also use the AS to derive an upper bound of the effect of changing the treatment from D_1 to a lower value than D_1. And under concavity, one can derive an upper (resp. lower) bound of the effect of changing the treatment from D_2 (resp. D_1) to a larger (resp. lower) value.[4] Importantly, the AS is identified even if those linearity or convexity/concavity conditions fail. But those non-placebo testable conditions are necessary to use the AS to identify or bound the effects of alternative policies.

[4]See D'Haultfœuille et al. (2023) for bounds of the same kind obtained under concavity or convexity.

3.2 Identification

To identify the AS, 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 AS identification)* $P(S=1)>0$, $P(S=1|D_1)<1$.

Assumption 4 implies that $P(S=0)>0$, meaning that there are stayers whose treatment does not change. While we assume that 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 AS, 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)* $\exists c>0$*:* $P(|\Delta D|>c|S=1)=1$.

We relax Assumption 5 just below.

Theorem 1 *If Assumptions 1-5 hold,*

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

If there are quasi-stayers, the AS is still identified. For any $\eta>0$, let $S_\eta = 1\{|\Delta D|>\eta\}$ be an indicator 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 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, as in Graham and Powell (2012) who consider a related estimand with some quasi-stayers. Accordingly, with quasi-stayers the AS 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 AS 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 AS 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 AS 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.

4 Estimating a weighted average of switchers' slopes

4.1 Target parameter

In this section, our target parameter is

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

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. Accordingly, we refer to δ_2 as the Weighted Average of Slopes (WAS). $\delta_2 = \delta_1$ if and only if

$$\text{cov}\left(\frac{Y_2(D_2) - Y_2(D_1)}{D_2 - D_1}, |D_2 - D_1|\middle| S=1\right) = 0: \tag{3}$$

the WAS and AS are equal if and only if switchers' slopes are uncorrelated with $|D_2 - D_1|$.

Economically, the AS and WAS serve different purposes. As discussed above, under shape restrictions on the potential outcome function, the AS 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 WAS 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 $E(Y_2(D_2) - c_2 D_2) > E(Y_2(D_1) - c_2 D_1)$ or, equivalently,

$$\delta_2 > c_2 :$$

comparing δ_2 to c_2 is sufficient to evaluate if changing the treatment from D_1 to D_2 was beneficial.

4.2 Identification

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

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

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 WAS of switchers up, and δ_{2-} is the WAS of switchers down. One has

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

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 in Point 1 (resp. 2) of Assumption 7 below.

Assumption 7 *(Support conditions for WAS 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,*

$$\delta_{2+} = \frac{E\left(\Delta Y - E(\Delta Y|D_1, S = 0)|S_+ = 1\right)}{E(\Delta D|S_+ = 1)} \tag{5}$$

$$= \frac{E\left(\Delta Y|S_+ = 1\right) - E\left(\Delta Y \frac{P(S_+=1|D_1)}{P(S=0|D_1)} \frac{P(S=0)}{P(S_+=1)} \Big| S = 0\right)}{E(\Delta D|S_+ = 1)}. \tag{6}$$

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

$$\delta_{2-} = \frac{E\left(\Delta Y - E(\Delta Y|D_1, S = 0)|S_- = 1\right)}{E(\Delta D|S_- = 1)} \tag{7}$$

$$= \frac{E\left(\Delta Y|S_- = 1\right) - E\left(\Delta Y \frac{P(S_-=1|D_1)}{P(S=0|D_1)} \frac{P(S=0)}{P(S_-=1)} \Big| S = 0\right)}{E(\Delta D|S_- = 1)}. \tag{8}$$

3. If Assumptions 1-2 and Assumption 7 hold,

$$\delta_2 = \frac{E\left[sgn(\Delta D)\left(\Delta Y - E(\Delta Y|D_1, S = 0)\right)\right]}{E[|\Delta D|]} \tag{9}$$

$$= \frac{E\left[sgn(\Delta D)\Delta Y\right] - E\left[\Delta Y \frac{P(S_+=1|D_1) - P(S_-=1|D_1)}{P(S=0|D_1)} P(S = 0) \Big| S = 0\right]}{E[|\Delta D|]}. \tag{10}$$

Point 1 of Theorem 4 shows that δ_{2+}, the WAS 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 WAS 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 WAS 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. Point 3 of Theorem 4 also implies that δ_2 is identified by the following doubly-robust estimand:

$$\frac{E\left[\left(S_+ - S_- - \frac{P(S_+=1|D_1) - P(S_-=1|D_1)}{P(S=0|D_1)}(1 - S)\right)\left(\Delta Y - E(\Delta Y|D_1, S = 0)\right)\right]}{E[|\Delta D|]}. \tag{11}$$

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

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

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 *(Conditions for asymptotic normality of propensity-score WAS 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

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

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

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

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

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

and

$$\sqrt{n} \left(\widehat{\delta}^r_2 - \delta_2 \right) \xrightarrow{d} \mathcal{N}(0, V(\psi_2)).$$

2. If Assumptions 1-3 and 8 hold,

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

and

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

Based on (11), we can also estimate δ_2 using the following doubly-robust estimator:

$$\hat{\delta}^{dr}_2 = \frac{\sum_i \left(S_{i+} - S_{i-} - \frac{\hat{P}(S_+=1|D_1=D_{1i}) - P(S_{i-}=1|D_1=D_{1i})}{P(S_i=0|D_1=D_{1i})} (1-S_i) \right) (\Delta Y_i - \hat{E}(\Delta Y_i|D_1 = D_{1i}, S_i = 0))}{\sum_i |\Delta D_i|}.$$

This estimator has an important advantage. While our regression-based (resp. propensity-score-based) estimator is nominally non-parametric, in practice it requires choosing a polynomial order to estimate $E(\Delta Y|D_1, S=0)$ (resp. $P(S_+ = 1|D_1)$ and $P(S_- = 1|D_1)$), and the rate conditions on K_n in Assumptions 6 (resp. 8) do not give specific guidance on the choice of this tuning parameter. With the doubly-robust estimator above, one can choose this tuning parameter in a data-driven manner, using cross-validation (CV). Results in Section 4 of Andrews (1991) imply that a series estimator of a nonparametric regression model with a polynomial order chosen by CV is optimal: the ratio of its mean-squared error and that of an oracle estimator using the best polynomial order given the sample size converges to one. Then, as D_1 is a scalar variable, series estimators of $E(\Delta Y|D_1, S=0)$, $P(S_+ = 1|D_1)$, and $P(S_- = 1|D_1)$ with CV-chosen polynomial orders converge at a rate strictly faster than $n^{1/4}$, as long as one assumes that those nuisance

functions are twice differentiable. Then, we conjecture that one can show, following arguments similar to those in Farrell (2015), that $\hat{\delta}_2^{dr}$ with CV-chosen polynomial orders in the estimation of the nuisance functions is $\sqrt{n}-$consistent, with asymptotic variance $V(\psi_2)$.

Finally, we now show that under some assumptions, the asymptotic variance of the WAS estimator is lower than that of the AS estimator.

Proposition 1 *If Assumption 1 holds,* $(Y_2(D_2)-Y_2(D_1))/(D_2-D_1)=\delta$ *for some real number* δ, $V(\Delta Y(D_1)|D_1,D_2)=\sigma^2$ *for some real number* $\sigma^2>0$, $D_2\geq D_1$, *and* $\Delta D \perp\!\!\!\perp D_1$,

$$\begin{aligned} V(\psi_1) =& \sigma^2\left[\frac{E(1/(\Delta D)^2|S=1)}{P(S=1)}+\frac{\left(E(1/\Delta D|S=1)\right)^2}{P(S=0)}\right] \\ \geq& \sigma^2\frac{1}{\left(E(\Delta D|S=1)\right)^2}\left[\frac{1}{P(S=1)}+\frac{1}{P(S=0)}\right]=V(\psi_2), \end{aligned}$$

with equality if and only if $V(\Delta D|S=1)=0$.

Of course, the constant treatment effect and the homoscedasticity assumptions underlying Proposition 1 are strong, but one often has to make strong assumptions to be able to rank estimators' variances. The question then is whether this ranking still holds in real-life applications, where those assumptions are unlikely to hold. Put differently, all models are wrong but some are useful, and the question is whether Proposition 1 is useful. In our empirical application, we find that the variance of $\hat{\delta}_1$ is indeed much larger than that of $\hat{\delta}_2^{dr}$, as predicted by Proposition 1.

5 Instrumental-variable estimation

There are instances where the parallel-trends condition in Assumption 1 is implausible, but one has at hand an instrument satisfying a similar parallel-trends condition. For instance, one may be interested in estimating the price-elasticity of a good's consumption, but prices respond to supply and demand shocks, and therefore do not satisfy Assumption 1. On the other hand, taxes may not respond to supply and demand shocks and may satisfy a parallel-trends assumption.

5.1 Notation and assumptions

Let (Z_1,Z_2) denote the instrument's values at period one and two and $\mathcal{Z}_t$ be the support of Z_t. For any $z\in\mathcal{Z}_1\cup\mathcal{Z}_2$, let $D_1(z)$ and $D_2(z)$ respectively denote the unit's potential treatments at periods 1 and 2 with instrument z. Let $SC=1\{D_2(Z_2)\neq D_2(Z_1), Z_2\neq Z_1\}$ be an indicator equal to 1 for switchers-compliers, namely units whose instrument changes from period one to two and whose treatment is affected by that change in the instrument.

We replace Assumption 1 by the following assumption.[5]

Assumption 9 *(Reduced-form and first-stage parallel trends) For all* $z \in \mathcal{Z}_1$*,*

1. $E(Y_2(D_2(z)) - Y_1(D_1(z))|Z_1 = z, Z_2, D_1) = E(Y_2(D_2(z)) - Y_1(D_1(z))|Z_1 = z, D_1)$.
2. $E(D_2(z) - D_1(z)|Z_1 = z, Z_2, D_1) = E(D_2(z) - D_1(z)|Z_1 = z, D_1)$.

Point 1 of Assumption 9 requires that $Y_2(D_2(z)) - Y_1(D_1(z))$, units' outcome evolutions in the counterfactual where their instrument does not change from period one to two, be mean independent of Z_2, conditional on Z_1 and D_1. Unlike Assumption 1, this condition imposes some restrictions on treatment effect heterogeneity, and the goal of conditioning on D_1 is to minimize the stringency of those restrictions. To see this, note that the two following conditions are sufficient for Point 1 of Assumption 9 to hold:

$$E(Y_2(D_1(z)) - Y_1(D_1(z))|Z_1 = z, Z_2, D_1) = E(Y_2(D_1(z)) - Y_1(D_1(z))|Z_1 = z, D_1) \quad (12)$$
$$E(Y_2(D_2(z)) - Y_2(D_1(z))|Z_1 = z, Z_2, D_1) = E(Y_2(D_2(z)) - Y_2(D_1(z))|Z_1 = z, D_1). \quad (13)$$

(12) requires that $Y_2(D_1(z)) - Y_1(D_1(z))$, units' outcome evolutions in the counterfactual where their instrument and their treatment does not change from period one to two, be mean independent of Z_2, conditional on Z_1 and D_1. Thanks to the conditioning on D_1, (12) is a standard parallel trends assumption that does not impose any restriction on treatment effect heterogeneity, like Assumption 1. If D_1 was not conditioned upon, (12) would require parallel trends among units with different baseline treatments, which implicitly assumes homogeneous treatment effects over time, as discussed in Section 2. (13), on the other hand, is a restriction on treatment effect heterogeneity across units. Essentially, it requires that switching the treatment from $D_1(Z_1)$ to $D_2(Z_1)$, the natural treatment change happening over time even without any change in the instrument, has an effect on the outcome that is mean independent of Z_2 conditional on Z_1 and D_1. Importantly, note that Point 1 of Assumption 9 is placebo testable, by comparing the outcome evolutions of instrument-switchers and instrument-stayers before instrument-switchers experience a change of their instrument. Finally, Point 2 of Assumption 9 requires that units' treatment evolutions under Z_1 be mean independent of Z_2, conditional on Z_1 and D_1. Because D_1 is conditioned upon, this parallel trends condition is equivalent to a sequential exogeneity assumption (see Robins, 1986; Bojinov et al., 2021).

Point 1 of Assumption 9 is related to identifying assumptions previously proposed in the literature. de Chaisemartin (2010) and Hudson et al. (2017) also consider IV-DID estimands, in classical designs with two periods and a binary instrument that turns on for some units at period two. Both papers introduce a "reduced-form" parallel trends assumption similar to Point 1 of

[5]Note that with our notation where potential outcomes do not depend on z, we also implicitly impose the usual exclusion restriction.

Assumption 9, but without noting that it imposes restrictions on effects' heterogeneity, even in the simple designs considered by those papers.

We also make the following assumptions.

Assumption 10 *(Monotonicity and strictly positive first-stage) i) For all* $(z, z') \in \mathcal{Z}_2^2$, $z \geq z' \Rightarrow D_2(z) \geq D_2(z')$, *and ii)* $E(|D_2(Z_2) - D_2(Z_1)|) > 0$.

i) is a monotonicity assumption similar to that in Imbens and Angrist (1994). It requires that increasing the period-two instrument weakly increases the period-two treatment. This condition is plausible when the instrument is taxes and the treatment is prices, as is the case in our application. ii) requires that the instrument has a strictly positive first stage.

Assumption 11 *(Bounded instrument, Lipschitz and bounded reduced-form potential outcomes and potential treatments)*

1. $\mathcal{Z}_1$ *and* $\mathcal{Z}_2$ *are bounded subsets of* $\mathbb{R}$.
2. *For all* $t \in \{1, 2\}$ *and for all* $(z, z') \in \mathcal{Z}_t^2$, *there is a random variable* $\overline{Y} \geq 0$ *such that* $|Y_t(D_t(z)) - Y_t(D_t(z'))| \leq \overline{Y}|z - z'|$, *with* $\sup_{(z_1,z_2)\in Supp(Z_1,Z_2)} E[\overline{Y}|Z_1 = z_1, Z_2 = z_2] < \infty$.
3. *For all* $t \in \{1, 2\}$ *and for all* $(z, z') \in \mathcal{Z}_t^2$, *there is a random variable* $\overline{D} \geq 0$ *such that* $|D_t(z) - D_t(z')| \leq \overline{D}|z - z'|$, *with* $\sup_{(z_1,z_2)\in Supp(Z_1,Z_2)} E[\overline{D}|Z_1 = z_1, Z_2 = z_2] < \infty$.

Assumption 11 is an adaptation of Assumption 2 to the IV setting we consider in this section.

Assumption 12 *(iid sample) We observe* $(Y_{i,1}, Y_{i,2}, D_{i,1}, D_{i,2}, Z_{i,1}, Z_{i,2})_{1\leq i\leq n}$, *that are independent and identically distributed with the same probability distribution as* $(Y_1, Y_2, D_1, D_2, Z_1, Z_2)$.

5.2 Target parameter

In this section, our target parameter is

$$\delta_{IV} := E\left(\frac{|D_2(Z_2) - D_2(Z_1)|}{E(|D_2(Z_2) - D_2(Z_1)||SC = 1)} \times \frac{Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))}{D_2(Z_2) - D_2(Z_1)}\middle| SC = 1\right).$$

δ_{IV} is a weighted average of the slopes of compliers-switchers' period-two potential outcome functions, from their period-two treatment under their period-one instrument, to their period-two treatment under their period-two instrument. Slopes receive a weight proportional to the absolute value of compliers-switchers' treatment response to the instrument change. δ_{IV} is just equal to the reduced-form WAS effect of the instrument on the outcome, divided by the first-stage WAS effect of the instrument on the treatment. With a binary instrument, such that $Z_1 = 0$ and $Z_2 \in \{0, 1\}$, our IV-WAS effect coincides with that identified in Corollary 2 of Angrist et al. (2000), in a cross-sectional IV model.

We could also consider a reduced-form AS divided by a first-stage AS. The resulting target is a weighted average of the slopes $\frac{Y_2(D_2(Z_2))-Y_2(D_2(Z_1))}{D_2(Z_2)-D_2(Z_1)}$, with weights proportional to $\frac{D_2(Z_2)-D_2(Z_1)}{Z_2-Z_1}$. It seems more natural to us to weight compliers-switchers' slopes by the absolute value of their first-stage than by the slope of their first-stage.[6]

5.3 Identification

Let $S^I = 1\{Z_2 - Z_1 \neq 0\}$, $S^I_+ = 1\{Z_2 - Z_1 > 0\}$, and $S^I_- = 1\{Z_2 - Z_1 < 0\}$.

Assumption 13 *(Support conditions for IV-WAS identification)*

1. $0 < P(S^I_+ = 1)$, *and* $0 < P(S^I_+ = 1|Z_1, D_1)$ *implies that* $0 < P(S^I = 0|Z_1, D_1)$.
2. $0 < P(S^I_- = 1)$, *and* $0 < P(S^I_- = 1|Z_1, D_1)$ *implies that* $0 < P(S^I = 0|Z_1, D_1)$.

Theorem 6 *If Assumptions 9-11 and 13 hold,*

$$\delta_{IV} = \frac{E\left[sgn(\Delta Z)\left(\Delta Y - E(\Delta Y|Z_1, D_1, S^I = 0)\right)\right]}{E\left[sgn(\Delta Z)\left(\Delta D - E(\Delta D|Z_1, D_1, S^I = 0)\right)\right]} \tag{14}$$

$$= \frac{E\left[sgn(\Delta Z)\Delta Y\right] - E\left[\Delta Y \frac{P(S^I_+=1|Z_1,D_1)-P(S^I_-=1|Z_1,D_1)}{P(S^I=0|Z_1,D_1)} P(S^I = 0)\middle| S^I = 0\right]}{E\left[sgn(\Delta Z)\Delta D\right] - E\left[\Delta D \frac{P(S^I_+=1|Z_1,D_1)-P(S^I_-=1|Z_1,D_1)}{P(S^I=0|Z_1,D_1)} P(S^I = 0)\middle| S^I = 0\right]}. \tag{15}$$

The regression-based (resp. propensity-score-based) estimand identifying δ_{IV} is just equal to the regression-based (resp. propensity-score-based) estimand identifying the reduced-form WAS effect of the instrument on the outcome controlling for D_1, divided by the regression-based (resp. propensity-score-based) estimand identifying the first-stage WAS effect controlling for D_1.

5.4 Estimation and inference

Let

$$\widehat{\delta}^r_{IV} = \frac{\frac{1}{n}\sum_{i=1}^n \text{sgn}(\Delta Z_i)\left(\Delta Y_i - \widehat{E}(\Delta Y|Z_1 = Z_{i,1}, D_1 = D_{i,1}, S^I = 0)\right)}{\frac{1}{n}\sum_{i=1}^n \text{sgn}(\Delta Z_i)\left(\Delta D_i - \widehat{E}(\Delta D|Z_1 = Z_{i,1}, D_1 = D_{i,1}, S^I = 0)\right)}, \tag{16}$$

where $\widehat{E}(\Delta Y|Z_1, D_1, S^I = 0)$ and $\widehat{E}(\Delta D|Z_1, D_1, S^I = 0)$ are series estimators of $E(\Delta Y|Z_1, D_1, S^I = 0)$ and $E(\Delta D|Z_1, D_1, S^I = 0)$ defined analogously to the series estimator in Section 3.3.

[6] If the first-stage effect is homogenous and linear, the weights in the IV-AS effect reduce to one, and one recovers a standard AS effect. However, linearity and homogeneity of the first-stage effect are strong assumptions.

Let us define

$$\widehat{\delta}_{IV}^{ps} = \frac{\frac{1}{n}\sum_{i=1}^{n} \text{sgn}(\Delta Z_i)\Delta Y_i - \frac{1}{n}\sum_{i:S_i^I=0} \Delta Y_i \frac{\widehat{P}(S_+^I=1|Z_1=Z_{i1},D_1=D_{i1}) - \widehat{P}(S_-^I=1|Z_1=Z_{i1},D_1=D_{i1})}{\widehat{P}(S^I=0|Z_1=Z_{i1},D_1=D_{i1})}}{\frac{1}{n}\sum_{i=1}^{n} \text{sgn}(\Delta Z_i)\Delta D_i - \frac{1}{n}\sum_{i:S_i^I=0} \Delta D_i \frac{\widehat{P}(S_+^I=1|Z_1=Z_{i1},D_1=D_{i1}) - \widehat{P}(S_-^I=1|Z_1=Z_{i1},D_1=D_{i1})}{\widehat{P}(S^I=0|Z_1=Z_{i1},D_1=D_{i1})}}, \quad (17)$$

where $\widehat{P}(S_+^I = 1|Z_1, D_1)$ (resp. $\widehat{P}(S_-^I = 1|Z_1, D_1)$, $\widehat{P}(S^I = 0|Z_1, D_1)$) is a series logistic regression estimator of $P(S_+^I = 1|Z_1, D_1)$ (resp. $P(S_-^I = 1|Z_1, D_1)$, $P(S^I = 0|Z_1, D_1)$) defined analogously to the series logistic regression estimators in Section 4.3.

For any variable X, let

$$\begin{aligned}
\delta_X &= E\left[\text{sgn}(\Delta Z)\left(\Delta X - E(\Delta X|Z_1, D_1, S^I = 0)\right)\right] \\
\psi_X &= \frac{1}{E(|\Delta Z|)}\left\{\left(S_+^I - S_-^I - E(S_+^I - S_-^I|Z_1, D_1)\frac{(1-S^I)}{E(1-S^I|Z_1, D_1)}\right)\right. \\
&\qquad \left.\times (\Delta X - E(\Delta X|Z_1, D_1, S^I = 0)) - \delta_X |\Delta Z|\right\}.
\end{aligned}$$

Then, let

$$\psi_{IV} = \frac{\psi_Y - \delta_{IV}\psi_D}{\delta_D}.$$

Under technical conditions similar to those in Assumptions 6 and 8, one can show that

$$\begin{aligned}
\sqrt{n}\left(\widehat{\delta}_{IV}^{r} - \delta_{IV}\right) &\xrightarrow{d} \mathcal{N}(0, V(\psi_{IV})), \\
\sqrt{n}\left(\widehat{\delta}_{IV}^{ps} - \delta_{IV}\right) &\xrightarrow{d} \mathcal{N}(0, V(\psi_{IV})).
\end{aligned}$$

6 Extensions

In this section, we return to the case where the treatment, rather than an instrument, satisfies a parallel-trends condition. Combining the extensions below with the IV case is possible.

6.1 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 $S_{+,t} = 1\{D_t > D_{t-1}\}$ and $S_{-,t} = 1\{D_t < D_{t-1}\}$. We 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:

Assumption 14 *(Parallel trends) For all $t \geq 2$, for all $d \in \mathcal{D}_{t-1}$, $E(\Delta Y_t(d)|D_{t-1} = d, D_t) = E(\Delta Y_t(d)|D_{t-1} = d)$.*

Assumption 14 requires that $E(\Delta Y_t(d)|D_{t-1} = d, D_t = d')$ be constant across d': groups of units with the same period-$t-1$ treatment but different period-t treatments all have the same expected outcome evolution in the counterfactual where their period-$t-1$ treatment would not have changed. Importantly, note that because Assumption 14 is conditional on D_{t-1}, it cannot be "chained" across pairs of time periods: it requires parallel trends over pairs of consecutive time periods, not over the entire duration of the panel. To preserve space, we do not restate our other assumptions with more than two periods.

Let

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

Let

$$\delta_1^{T\geq 3} = \sum_{t=2}^{T} \frac{P(S_t = 1)}{\sum_{k=2}^{T} P(S_k = 1)} \delta_{1,t},$$
$$\delta_2^{T\geq 3} = \sum_{t=2}^{T} \frac{E(|\Delta D_t|)}{\sum_{k=2}^{T} E(|\Delta D_k|)} \delta_{2,t}$$

be generalizations of the AS and WAS effects to applications with more than two periods. Note that in line with the spirit of the two effects, we propose different weights to aggregate the AS and WAS across time periods. For the AS, the weights are just proportional to the proportion of switchers between $t-1$ and t. For the WAS, the weights are proportional to the average absolute value of the treatment switch from $t-1$ to t.

Theorem 7 *If Assumption 14 and generalizations of Assumptions 2-5 to more than two periods hold,*

$$\delta_1^{T\geq 3} = \sum_{t=2}^{T} \frac{P(S_t = 1)}{\sum_{k=2}^{T} P(S_k = 1)} E\left(\frac{\Delta Y_t - E(\Delta Y_t|D_{t-1}, S_t = 0)}{\Delta D_t}\middle| S_t = 1\right).$$

Theorem 8 *If Assumption 14 and generalizations of Assumptions 2 and 7 to more than two periods hold,*

$$\begin{aligned}\delta_2^{T\geq 3} &= \sum_{t=2}^{T} \frac{E(|\Delta D_t|)}{\sum_{k=2}^{T} E(|\Delta D_k|)} \frac{E\left(sgn(\Delta D_t)\left(\Delta Y_t - E(\Delta Y_t|D_{t-1}, S_t = 0)\right)\right)}{E(|\Delta D_t|)} \\ &= \sum_{t=2}^{T} \frac{E(|\Delta D_t|)}{\sum_{k=2}^{T} E(|\Delta D_k|)} \frac{E\left[sgn(\Delta D_t)\Delta Y_t\right] - E\left[\Delta Y_t \frac{P(S_{+,t}=1|D_{t-1}) - P(S_{-,t}=1|D_{t-1})}{P(S_t=0|D_{t-1})} P(S_t = 0)\middle| S_t = 0\right]}{E(|\Delta D_t|)}.\end{aligned}$$

Theorems 7 and 8 are straightforward generalizations of Theorems 1 and 4 to settings with more than two time periods.

Let

$$\psi_{1,t} = \frac{1}{E(S_t)} \left\{ \left(\frac{S_t}{\Delta D_t} - E\left(\frac{S_t}{\Delta D_t} \middle| D_{t-1} \right) \frac{(1-S_t)}{E[1-S_t|D_{t-1}]} \right) [\Delta Y_t - E(\Delta Y_t | D_{t-1}, S_t = 0)] - \delta_{1,t} S_t \right\},$$

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

After some algebra, one can show that the influence function of the AS estimator with several periods is

$$\psi_1^{T \geq 3} := \frac{\sum_{t=2}^{T} (P(S_t = 1)\psi_{1,t} + (\delta_{1,t} - \delta_1^{T \geq 3})(S_t - P(S_t = 1)))}{\sum_{t=2}^{T} P(S_t = 1)}, \tag{18}$$

while the influence function of the WAS estimators with several periods is

$$\psi_2^{T \geq 3} := \frac{\sum_{t=2}^{T} E(|\Delta D_t|)\psi_{2,t} + (\delta_{2,t} - \delta_2^{T \geq 3})(|\Delta D_t| - E(|\Delta D_t|))}{\sum_{t=2}^{T} E(|\Delta D_t|)}. \tag{19}$$

Importantly, those influence functions allow the unit's treatments and outcomes to be arbitrarily serially correlated.

6.2 Placebo tests

With several time periods, one can test the following condition, which is closely related to Assumption 14:

Assumption 15 *(Testable parallel trends) For all $t \geq 3, t \leq T$, for all $d \in \mathcal{D}_{t-1}$, $E(\Delta Y_{t-1}(d)|D_{t-2} = D_{t-1} = d, D_t) = E(\Delta Y_{t-1}(d)|D_{t-2} = D_{t-1} = d)$.*

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-$t-1$ stayers. If one finds that from $t-2$-to-$t-1$, $t-1$-to-t switchers and stayers are on parallel trends, this lends credibility to Assumption 14.

Assumption 14 can only be placebo tested among $t-2$-to-$t-1$ stayers. Then, as a robustness check one may restrict the estimation of δ_1 and δ_2 to $t-2$-to-$t-1$ stayers, to ensure that effects are only estimated in a subsample for which the identifying assumption can be placebo tested. The resulting estimator relies on the following identifying assumption:

$$\forall t \geq 3, t \leq T, d \in \mathcal{D}_{t-1}: \; E(\Delta Y_t(d)|D_{t-2} = D_{t-1} = d, D_t) = E(\Delta Y_t(d)|D_{t-2} = D_{t-1} = d),$$

the exact analogue of Assumption 15 but one period ahead.

6.3 Estimators robust to dynamic effects up to a pre-specified treatment lag.

Importantly, the robustness check in the previous section also yields an estimator robust to dynamic effects up to one treatment lag. If units' current and first treatment lag affect their current outcome, our $t-1$-to-t estimators in the subsample of $t-2$-to-$t-1$ stayers are unbiased for effects of the current treatment on the outcome under the following assumption:

$$\forall t \geq 3, t \leq T, d \in \mathcal{D}_{t-1}: \; E(Y_t(d,d) - Y_{t-1}(d,d) | D_{t-2} = D_{t-1} = d, D_t) = E(Y_t(d,d) - Y_{t-1}(d,d) | D_{t-2} = D_{t-1} = d).$$

Similarly, if one wants to allow for effects of the first and second treatment lags on the outcome, one just needs to restrict the estimation sample to $t-3$-to-$t-1$ stayers. However, the more robustness to dynamic effects one would like to have, the smaller the estimation sample becomes.

7 Application

Data and research questions. We use the yearly 1966-to-2008 panel dataset of Li et al. (2014), covering 48 US states (Alaska and Hawaii are excluded). In view of the long duration of this panel, it is important to keep in mind that our estimators only assume parallel trends across pairs of consecutive years, not over the panel's entire duration. For each state×year cell (i,t), the data contains $Z_{i,t}$, the total (state plus federal) gasoline tax in cents per gallon, $D_{i,t}$, the log tax-inclusive price of gasoline, and $Y_{i,t}$, the log gasoline consumption per adult. Our goal is to estimate the effect of gasoline taxes on gasoline consumption and prices, and to estimate the price-elasticity of gasoline consumption, using taxes as an instrument. Instead, Li et al. (2014) jointly estimate the effect of gasoline taxes and tax-exclusive prices on consumption, using a TWFE regression with two treatments. Between each pair of consecutive periods, the tax-exclusive price changes in all states, so this treatment does not have stayers and its effect cannot be estimated using the estimators proposed in this paper. Thus, our estimates cannot be compared to those of Li et al. (2014).

Switching cells, and how they compare to the entire sample. Let $\mathcal{S}$ be the set of switching (i,t) cells such that $Z_{i,t} \neq Z_{i,t-1}$ but $Z_{i',t} = Z_{i',t-1}$ for some i'. The second condition drops from the estimation seven pairs of consecutive time periods between which the federal gasoline tax changed, thus implying that all states experienced a change of their tax. $\mathcal{S}$ includes 384 cells, so effects of taxes on gasoline prices and consumptions can be estimated for 19% of the 2,016 state×year cells for which $Z_{i,t} - Z_{i,t-1}$ can be computed. Table 1 below compares some observable characteristics of switchers and stayers. Switchers seem slightly over-represented in the later years of the panel: t is on average 2.5 years larger for switchers than for stayers, and the difference is significant. On the other hand, switchers are not more populated than stayers, and

their gasoline consumption and gasoline price in 1966 are not significantly different from that of stayers. Thus, there is no strong indication that the cells in $\mathcal{S}$ are a very selected subgroup.

Table 1: Comparing switchers and stayers

Dependent Variables:	t	Adult Population	$\log(\text{quantity})_{1966}$	$\log(\text{price})_{1966}$
Constant	1,986.7	3,691,608.0	-0.5161	3.471
	(0.2739)	(577,164.0)	(0.0210)	(0.0054)
$\mathbf{1}\{Z_{i,t} \neq Z_{i,t-1}\}$	2.481	39,588.0	-0.0099	0.0014
	(0.7519)	(320,342.1)	(0.0096)	(0.0029)
N	2,016	2,016	2,016	2,016

Notes: The table show the results of regressions of some dependent variables on a constant and an indicator for switching cells. The standard errors shown in parentheses are clustered at the state level.

Distribution of taxes. As an example, the top panel of Figure 1 below shows the distribution of $Z_{g,1987}$ for 1987-to-1988 stayers, while the bottom panel shows the distribution for 1987-to-1988 switchers. The figure shows that there are many values of $Z_{g,1987}$ such that only one or two states have that value, so $Z_{g,1987}$ is close to being continuously distributed. Moreover, all switchers g are such that

$$\min_{g':Z_{g',1988}=Z_{g',1987}} Z_{g',1987} \leq Z_{g,1987} \leq \max_{g':Z_{g',1988}=Z_{g',1988}} Z_{g',1987}.$$

Thus, Assumption 4 seems to hold for this pair of years. $(1987, 1988)$ is not atypical. While $Z_{i,t}$ varies less across states in the first years of the panel, there are many other years where $Z_{i,t}$ is close to being continuously distributed. Similarly, almost 95% of cells in $\mathcal{S}$ are such that $\min_{g':Z_{i',t}=Z_{i',t-1}} Z_{i',t-1} \leq Z_{i,t-1} \leq \max_{g':Z_{i',t}=Z_{i',t-1}} Z_{i',t-1}$. Dropping the few cells that do not satisfy this condition barely changes the results presented below.

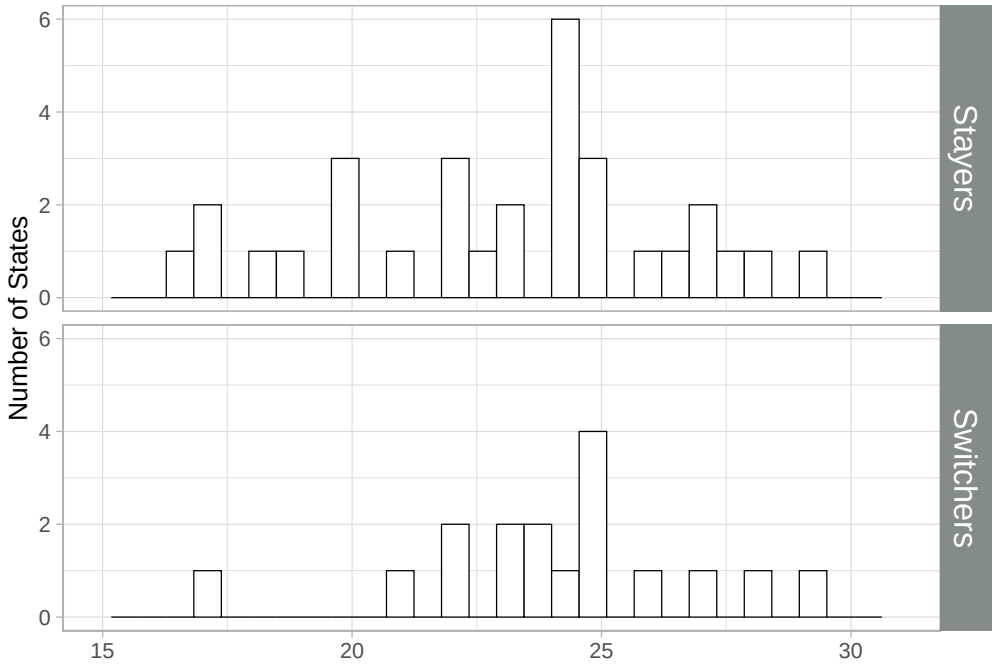

Figure 1: Gasoline tax in 1987 among 1987-to-1988 switchers and stayers

Distribution of tax changes. Figure 2 below shows the distribution of $Z_{i,t} - Z_{i,t-1}$ for the 384 cells in $\mathcal{S}$. The majority experience an increase in their taxes, but 38 cells experience a decrease. The average value of $|Z_{i,t} - Z_{i,t-1}|$ is equal to 1.61 cents, while prior to the tax change, switchers' average gasoline price is equal to 112 cents: our estimators leverage small changes in taxes relative to gasoline prices. Finally, $\min_{(i,t)\in\mathcal{S}} |Z_{i,t} - Z_{i,t-1}| = 0.05$: some switchers experience a very small change in their taxes.

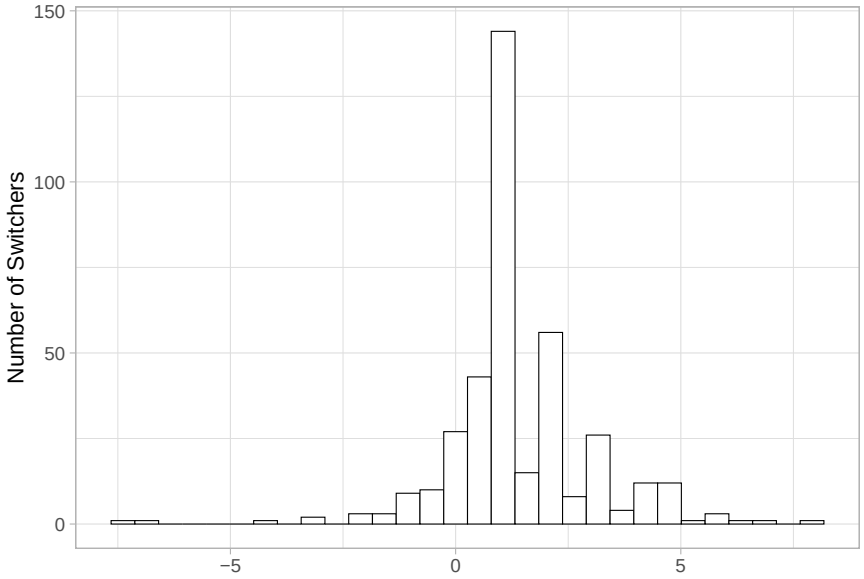

Figure 2: Distribution of tax changes between consecutive periods

Reduced-form and first-stage AS and WAS estimates. Table 2 below shows the AS and doubly-robust WAS estimates of the reduced-form (Panel A) and first-stage (Panel B) effects of taxes on quantities and prices. We follow results from Section 5, and estimate the reduced-form and the first-stage controlling for lagged prices D_{t-1}, to ensure that the resulting IV estimator is robust to heterogeneous effects over time. Reduced-form and first-stage estimators where D_{t-1} is not controlled for are not very different, but controlling for D_{t-1} reduces the standard error of the first-stage estimator. In Column (1), the estimators are computed using a polynomial of order 1 in (Z_{t-1}, D_{t-1}) to estimate $E(\Delta Y_t|Z_{t-1}, D_{t-1}, S_t = 0)$, $E(\Delta D_t|Z_{t-1}, D_{t-1}, S_t = 0)$, and the propensity scores $P(S_{+,t} = 1|Z_{t-1}, D_{t-1})$, $P(S_{-,t} = 1|Z_{t-1}, D_{t-1})$, and $P(S_t = 0|Z_{t-1}, D_{t-1})$. In Column (2), a polynomial of order 2 is used in those estimations. 10-folds cross-validation selects a polynomial of order two for $E(\Delta D_t|Z_{t-1}, D_{t-1}, S_t = 0)$, and a polynomial of order one for all the other conditional expectations. Thus, polynomials of order 1 and 2 are in line with those selected by cross validation. Standard errors clustered at the state level, computed following (18) and (19), are shown below the estimates, between parentheses. All estimations use 1632 (48$\times$ 35) first-difference observations: 7 periods have to be excluded as they do not have stayers. Finally, the last line of each panel shows the p-value of a test that the AS and WAS effects are equal. In Panel A Column (1), the AS estimate indicates that increasing gasoline tax by 1 cent decreases quantities consumed by 0.55 percent on average for the switchers. That effect is significant at the 5% level, but it becomes smaller and insignificant when one uses a quadratic model to estimate $E(\Delta Y_t|Z_{t-1}, D_{t-1}, S_t = 0)$. The WAS estimates are slightly lower than, but close to, the AS estimates, and they are significant irrespective of the polynomial order used in the estimation. As predicted by Proposition 1, the standard errors of the WAS estimators are almost 3 times smaller than that of the AS estimators. Equality tests that the AS and WAS effects are equal are not rejected. In Panel B, the AS estimates of the first-stage effect are insignificant. Importantly, this implies that an IV-AS estimator of the price elasticity of gasoline consumption cannot be used: this estimator does not have a significant first stage. The WAS estimates are significant, and they indicate that if gasoline tax increases by 1 cent on average, prices increase by around 0.5 percent on average for the switchers. Again, the differences between the AS and WAS effects of taxes on prices are insignificant.

Table 2: Effects of gasoline tax on quantities consumed and prices

	(1) Linear model	(2) Quadratic model
Panel A: Reduced-form effect of taxes on quantities consumed		
AS	-0.0055	-0.0034
	(0.0027)	(0.0032)
WAS	-0.0038	-0.0034
	(0.0010)	(0.0011)
Observations	1,632	1,632
P-value	0.4482	0.9974
Panel B: First-stage effect of taxes on prices		
AS	0.0042	0.0047
	(0.0024)	(0.0025)
WAS	0.0056	0.0056
	(0.0009)	(0.0008)
Observations	1,632	1,632
P-value	0.4729	0.6798

Notes: All estimators in the table are computed using the data of Li et al. (2014). Panel A (resp. B) shows the AS and doubly-robust WAS estimates of the reduced-form (resp. first-stage) effect of taxes on quantities (resp. prices). All estimates control for the lag of prices. In Column (1), estimates are computed using a polynomial of order 1 in (Z_{t-1}, D_{t-1}) to estimate $E(\Delta Y_t|Z_{t-1}, D_{t-1}, S_t = 0)$ and the propensity scores $P(S_{+,t} = 1|Z_{t-1}, D_{t-1})$, $P(S_{-,t} = 1|Z_{t-1}, D_{t-1})$, and $P(S_t = 0|Z_{t-1}, D_{t-1})$. In Column (2), estimates are computed using a polynomial of order 2 in those estimations. Standard errors clustered at the state level, computed following (18) and (19) are shown below the estimates, between parentheses. All estimations use 1632 ($48\times$ 35) first-difference observations: 7 periods have to be excluded as they do not have stayers. Finally, the last line of each panel shows the p-value of a test that the AS and WAS effects are equal.

Placebo analysis. Table 3 below shows placebo AS and doubly-robust WAS estimates of the reduced-form and first-stage effects. The placebo estimators are analogous to the actual estimators, but they replace ΔY_t by ΔY_{t-1}, and they restrict the sample, for each pair of consecutive time periods $(t-1, t)$, to states whose taxes did not change between $t-2$ and $t-1$. The placebo WAS estimates are small and insignificant, both for quantities and prices. The placebo AS estimates are larger for quantities, but they are insignificant, and less precisely estimated. This placebo analysis shows that before switchers change their gasoline taxes, switchers' and stayers' consumption of gasoline and gasoline prices do not follow detectably different evolutions. As a robustness check, we reestimate the AS and WAS in the placebo subsample, to ensure that effects are estimated in a subsample for which the identifying assumption can be placebo tested, and also because in that subsample estimators remain valid if the first lag of taxes affect current

gasoline prices and quantities. WAS reduced-form effects are very close to those in Table 2. WAS first-stage effects are 25 to 35% smaller, though they are still positive and highly significant.

Table 3: Placebo effects of gasoline tax on quantities consumed and prices

	(1) Linear model	(2) Quadratic model
Panel A: Reduced-form placebo effect of taxes on quantities consumed		
AS	0.0039	0.0055
	(0.0035)	(0.0036)
WAS	0.0001	0.0012
	(0.0017)	(0.0017)
Observations	1,059	1,059
Panel B: First-stage placebo effect of taxes on prices		
AS	0.0006	0.0009
	(0.0056)	(0.0053)
WAS	0.0014	0.0013
	(0.0017)	(0.0015)
Observations	1,059	1,059

Notes: The table shows the placebo AS and doubly-robust WAS estimates of the reduced-form and first-stage effects of taxes on quantities and prices. The estimators and their standard errors are computed as the actual estimators, replacing ΔY_t by ΔY_{t-1}, and restricting the sample, for each pair of consecutive time periods $(t-1,t)$, to states whose taxes did not change between $t-2$ and $t-1$.

IV-WAS estimate of the price-elasticity of gasoline consumption. Table 4 shows doubly-robust IV-WAS estimates of the price-elasticity of gasoline consumption. As the instrument's first stage is not very strong and the sample effectively only has 48 observations, asymptotic approximations may not be reliable for inference. In line with that conjecture, we find that the bootstrap distributions of the three estimators in Table 4 are non-normal, with some outliers. Therefore, we use the percentile bootstrap for inference, clustering the bootstrap at the state level. Reassuringly, these confidence intervals have nominal coverage in simulations tailored to our application.[7] The IV-WAS estimates are negative, significant, and larger than

[7]Here is the DGP used in our simulations. We estimate TWFE regressions of $Y_{i,t}$ on state and year fixed effects and $Z_{i,t}$, and of $D_{i,t}$ on state and year fixed effects and $Z_{i,t}$. We let $\hat{\gamma}_i^Y+\hat{\lambda}_t^Y+\hat{\beta}^Y Z_{i,t}+\epsilon_{i,t}^Y$ and $\hat{\gamma}_i^D+\hat{\lambda}_t^D+\hat{\beta}^D Z_{i,t}+\epsilon_{i,t}^D$ denote the resulting regression decompositions. In each simulation, the simulated instrument is just the actual instrument, while the simulated outcomes and treatments are respectively equal to $Y_{i,t}^s=\hat{\gamma}_i^Y+\hat{\lambda}_t^Y+\hat{\beta}^Y Z_{i,t}+\epsilon_{i,t}^{Y,s}$, and $D_{i,t}^s=\hat{\gamma}_i^D+\hat{\lambda}_t^D+\hat{\beta}^D Z_{i,t}+\epsilon_{i,t}^{D,s}$, where the vector of simulated residuals $(\epsilon_{g,1}^{Y,s},...,\epsilon_{g,T}^{Y,s},\epsilon_{g,1}^{D,s},...,\epsilon_{g,T}^{D,s})$ is drawn at random and with replacement from the estimated vectors of residuals $((\epsilon_{g',1}^Y,...,\epsilon_{g',T}^Y,\epsilon_{g',1}^D,...,\epsilon_{g',T}^D))_{g'\in\{1,...,G\}}$. Thus, the first-stage and reduced-form effects, the correlation between the reduced-form and first-stage residuals,

-1, though their confidence intervals contain -1. We compare those estimates to a 2SLS-TWFE estimator, computed via a 2SLS regression of $Y_{i,t}$ on $D_{i,t}$ and state and year fixed effects, using $Z_{i,t}$ as the instrument. The 2SLS-TWFE coefficient is equal to -1.0836 (bootstrap confidence interval=$[-2.1207, -0.4405]$), which is 60% larger in absolute value than the IV-WAS estimate in Column (1), and almost 80% larger than that in Column (2), though the 2SLS-TWFE coefficient does not significantly differ from the two IV-WAS estimates (P-value = 0.320 and 0.232 , respectively). Interestingly, the confidence interval of the 2SLS-TWFE coefficient is almost 80% wider than that of the IV-WAS coefficient in Column (1) and 27% wider than that in Column (2), thus showing that using a more robust estimator does not always come with a substantive precision cost.

Table 4: IV estimators of the price-elasticity of gasoline consumption

	(1) Linear model	(2) Quadratic model
IV-WAS	-0.6773	-0.6130
	[-1.2101,-0.2622]	[-1.3183,-0.0004]
Observations	1,632	1,632

Notes: The table shows doubly-robust IV-WAS estimates of the price-elasticity of gasoline consumption, computed using the data of Li et al. (2014). Bootstrap confidence intervals are shown below the estimates. They are computed with 500 bootstrap replications, clustered at the state level.

8 Conclusion

We propose new difference-in-difference (DID) estimators for continuous treatments. We assume that between pairs of consecutive periods, the treatment of some units, the switchers, changes, while the treatment of other units, the stayers, does not change. We propose a parallel trends assumption on the outcome evolution of switchers and stayers with the same baseline treatment. Under that assumption, two target parameters can be estimated. Our first target is the average slope of switchers' period-two potential outcome function, from their period-one to their period-two treatment, referred to as the AS. Our second target is a weighted average of switchers' slopes, where switchers receive a weight proportional to the absolute value of their treatment change, referred to as the WAS. Economically, the AS and WAS serve different purposes, so neither parameter dominates the other. On the other hand, when it comes to estimation, the WAS unambiguously dominates the AS. First, it can be estimated at the parametric rate even if units can experience an arbitrarily small treatment change. Second, under some conditions, its asymptotic variance is strictly lower than that of the AS estimator. Third, unlike the AS, it is

and the residuals' serial correlation are the same as in the sample.

amenable to doubly-robust estimation. In our application, we use US-state-level panel data to estimate the effect of gasoline taxes on gasoline consumption. The standard error of the WAS is almost three times smaller than that of the AS, and the two estimates are close.

We also consider the instrumental-variable case, as there are instances where units experiencing/not experiencing a treatment change are unlikely to be on parallel trends, but one has at hand an instrument such that units experiencing/not experiencing an instrument change are more likely to be on parallel trends. Then, we propose widely applicable IV-DID estimators, that are robust to heterogeneous effects over time but impose some restrictions on effects' heterogeneity across units.

References

Andrews, D. W. (1991). Asymptotic optimality of generalized cl, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics 47*(2-3), 359–377.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies 67*(3), 499–527.

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 (2024). Revisiting event-study designs: robust and efficient estimation.

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.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of econometrics 18*(1), 5–46.

de Chaisemartin, C. (2010). A note on instrumented difference in differences.

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 (2023a). Difference-in-differences estimators of intertemporal treatment effects. arXiv preprint arXiv:2007.04267.

de Chaisemartin, C. and X. D'Haultfœuille (2023b). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal 26*(3), C1–C30.

de Chaisemartin, C. and X. D'Haultfœuille (2024). Two-way fixed effects and differences-in-differences estimators in heterogeneous adoption designs. *arXiv preprint arXiv:2405.04465*.

de Chaisemartin, C., X. D'Haultfœuille, and G. Vazquez-Bare (2023). Difference-in-differences estimators with continuous treatments and no stayers.

D'Haultfœuille, X., S. Hoderlein, and Y. Sasaki (2023). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *Journal of Econometrics 234*(2), 664–690.

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.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

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.

Hausman, J. A. and W. K. Newey (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica: Journal of the Econometric Society*, 1445–1476.

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

Hudson, S., P. Hull, and J. Liebersohn (2017). Interpreting instrumented difference-in-differences. *Metrics Note, Sept*.

Imbens, G. and Y. Xu (2024). Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint arXiv:2406.00827*.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American economic review 91*(4), 778–794.

Li, S., J. Linn, and E. Muehlegger (2014). Gasoline taxes and consumer behavior. *American Economic Journal: Economic Policy 6*(4), 302–342.

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.

9 Proofs

Hereafter, $\text{Supp}(X)$ 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.

9.1 Theorem 1

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

9.2 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,

$$\lim_{\eta \downarrow 0} P(S_\eta = 1) = P(S = 1) > 0, \tag{20}$$

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),

$$\begin{aligned} E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_2) =& E(Y_2(d_1) - Y_1(d_1)|D_1 = d_1, D_2 = d_1) \\ =& E(\Delta Y|D_1 = d_1, S = 0), \end{aligned} \tag{21}$$

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,

$$\begin{aligned} & E\left(\frac{Y_2(D_2) - Y_2(D_1)}{\Delta D}\middle| S_\eta = 1\right) \\ =& E\left(\frac{E(Y_2(D_2) - Y_1(D_1)|D_1, D_2) - E(Y_2(D_1) - Y_1(D_1)|D_1, D_2)}{\Delta D}\middle| S_\eta = 1\right) \\ =& E\left(\frac{E(\Delta Y|D_1, D_2) - E(\Delta Y|D_1, S = 0)}{\Delta D}\middle| S_\eta = 1\right) \\ =& E\left(\frac{\Delta Y - E(\Delta Y|D_1, S = 0)}{\Delta D}\middle| S_\eta = 1\right), \end{aligned} \tag{22}$$

where the first equality follows from the law of iterated expectations, the second follows from (21), 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,

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

where the second inequality follows by Assumption 2. Now, by (20) 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

$$\delta_1=\lim_{\eta\downarrow 0} E\left[\frac{Y_2(D_2)-Y_2(D_1)}{\Delta D}\middle|S_\eta=1\right]. \tag{23}$$

The result follows by combining (22) and (23) □

9.3 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, $E[((1-S)\Delta Y-E[(1-S)\Delta Y|D_1])^2|D_1]\leq E\left[\Delta Y^2|D_1\right]<\infty$ 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 $\gamma = A_K^{-1'}\tilde{\gamma}$, $\gamma' P^k(d_1) = \tilde{\gamma}' A_K^{-1} P^K(d_1)$ 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 $\psi_1 = [g(Z_i, \mu_0) + \alpha(Z_i)]/E[S]$. □

9.4 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. Moreover, 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 (21) 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. □

9.5 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

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

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

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

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

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

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

$$\begin{aligned} E\left[\|U_i\|^2\right] \leq & E\left[\left(\frac{\Delta Y_i (1 - S_i)}{\mu_{2i}} - \frac{\mu_{Yi}}{\mu_{2i}} \right)^2 \left\| P^K(D_{1i}) \right\|^2 \right] \\ \leq & C E\left[\left\| P^K(D_{1i}) \right\|^2 \right] \\ = & C E\left[\text{trace}\{P^K(D_{1i})' P^K(D_{1i})\} \right] \\ = & C \times \text{trace}\left(E\left[P^K(D_{1i}) P^K(D_{1i})' \right] \right) \\ = & CK, \end{aligned} \tag{24}$$

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 (24). 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

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

and

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

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

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

Term 4. For the fourth term,

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

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

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

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 $\text{sgn}(\Delta D_i) = S_{i+} - S_{i-}$ and $|\Delta D_i| = \Delta D_i(S_{i+} - S_{i-})$, 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 □

9.6 Proposition 1

If $D_2 \geq D_1$ and $\Delta D \perp\!\!\!\perp D_1$,

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

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

If $(Y_2(D_2) - Y_2(D_1))/(D_2 - D_1) = \delta$, then $\delta_1 = \delta_2 = \delta$, and $\Delta Y = \Delta Y(D_1) + \Delta D \delta$, so after some algebra the previous display simplifies to

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

$$\psi_2 = \frac{1}{E(\Delta D)} \left(S - (1-S) \frac{E(S)}{1 - E(S)} \right) \times (\Delta Y(D_1) - E(\Delta Y(D_1) | D_1, S = 0)).$$

Then, under Assumption 1,

$$E(\psi_1|D_1, D_2) = E(\psi_2|D_1, D_2) = 0.$$

Then, using the law of total variance, the fact that $V(\Delta Y(D_1)|D_1, D_2) = \sigma^2$, and some algebra,

$$\begin{aligned} V(\psi_1) =& E(V(\psi_1|D_1, D_2)) \\ =& \sigma^2 E\left(\left[\frac{\frac{S}{\Delta D} - \frac{1-S}{1-E(S)}E\left(\frac{S}{\Delta D}\right)}{E(S)}\right]^2\right) \\ =& \sigma^2\left[\frac{E(1/(\Delta D)^2|S=1)}{P(S=1)} + \frac{\left(E(1/\Delta D|S=1)\right)^2}{P(S=0)}\right], \end{aligned}$$

and

$$\begin{aligned} V(\psi_2) =& E(V(\psi_2|D_1, D_2)) \\ =& \sigma^2 E\left(\left[\frac{S - (1-S)\frac{E(S)}{1-E(S)}}{E(\Delta D)}\right]^2\right) \\ =& \sigma^2 \frac{1}{\left(E(\Delta D|S=1)\right)^2}\left[\frac{1}{P(S=1)} + \frac{1}{P(S=0)}\right]. \end{aligned}$$

The inequality follows from the convexity of $x \mapsto x^2$, the convexity of $x \mapsto 1/x$ on $\mathbb{R}^+ \setminus \{0\}$ and $\Delta D|S=1 \in \mathbb{R}^+ \setminus \{0\}$, Jensen's inequality, and $x \mapsto x^2$ increasing on $\mathbb{R}^+$, which together imply that

$$E(1/(\Delta D)^2|S=1) \geq \left(E(1/\Delta D|S=1)\right)^2 \geq \frac{1}{\left(E(\Delta D|S=1)\right)^2}.$$

Finally, Jensen's inequality is strict for strictly convex functions, unless the random variable is actually constant. The last claim of the proposition follows.

9.7 Theorem 6

The parameter δ_{IV} can be written as:

$$\delta_{IV} = \frac{E[\text{sgn}(\Delta Z)\left(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))\right)|SC=1]}{E[|D_2(Z_2) - D_2(Z_1)|\,|SC=1]}. \tag{25}$$

The regression-based estimand is:

$$\frac{E\left[\text{sgn}(\Delta Z)\left(\Delta Y - E(\Delta Y|Z_1, S^I=0, D_1)\right)\right]}{E\left[\text{sgn}(\Delta Z)\left(\Delta D - E(\Delta D|Z_1, S^I=0, D_1)\right)\right]}.$$

Following previous arguments, the conditional expectations are well-defined under Assumption 13. For the denominator,

$$\begin{aligned}
&E\left[\text{sgn}(\Delta Z)\left(\Delta D - E(\Delta D|Z_1, S^I = 0, D_1)\right)\right] \\
=&E\left[\text{sgn}(\Delta Z)\left(D_2(Z_2) - D_2(Z_1)\right)\right] + E\left[\text{sgn}(\Delta Z)\left(D_2(Z_1) - D_1(Z_1) - E(\Delta D|Z_1, S^I = 0, D_1)\right)\right] \\
=&E\left[\text{sgn}(\Delta Z)\left(D_2(Z_2) - D_2(Z_1)\right)\right]
\end{aligned}$$

because

$$\begin{aligned}
&E\left[\text{sgn}(\Delta Z)\left(D_2(Z_1) - D_1(Z_1) - E(\Delta D|Z_1, S^I = 0, D_1)\right)\right] \\
=&E\left\{E\left[\text{sgn}(\Delta Z)\left(D_2(Z_1) - D_1(Z_1) - E(\Delta D|Z_1, S^I = 0, D_1)\right)|Z_1, Z_2, D_1\right]\right\} \\
=&E\left\{\text{sgn}(\Delta Z)\left(E\left(\Delta D(Z_1)|Z_1, Z_2, D_1\right) - E(\Delta D(Z_1)|Z_1, S^I = 0, D_1)\right)\right\} \\
=&0,
\end{aligned}$$

by Assumption 9. On the other hand,

$$\begin{aligned}
E\left[\text{sgn}(\Delta Z)\left(D_2(Z_2) - D_2(Z_1)\right)\right] =& E\left[\text{sgn}(\Delta Z)\left(D_2(Z_2) - D_2(Z_1)\right)|D_2(Z_2) \neq D_2(Z_1)\right] \\
&\times P\left(D_2(Z_2) \neq D_2(Z_1)\right) \\
=& E\left[|D_2(Z_2) - D_2(Z_1)|\,|SC = 1\right] P(SC = 1),
\end{aligned}$$

where the last equality follows from monotonicity (Assumption 10) and the definition of switchers-compliers. Next, the numerator is:

$$\begin{aligned}
&E\left[\text{sgn}(\Delta Z)\left(\Delta Y - E(\Delta Y|Z_1, S^I = 0, D_1 = 0)\right)\right] \\
=&E\left[\text{sgn}(\Delta Z)\left(Y_2(D_2(Z_2)) - Y_1(D_1(Z_1)) - E(\Delta Y|Z_1, S^I = 0, D_1 = 0)\right)\right] \\
=&E\left[\text{sgn}(\Delta Z)\left(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))\right)\right],
\end{aligned}$$

using the parallel trends assumption as before. Then,

$$\begin{aligned}
&E\left[\text{sgn}(\Delta Z)\left(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))\right)\right] \\
=&E\left[\text{sgn}(\Delta Z)\left(Y_2(D_2(Z_2)) - Y_2(D_2(Z_1))\right)|SC = 1\right] P(SC = 1),
\end{aligned}$$

and thus, in view of (25),

$$\frac{E\left[\text{sgn}(\Delta Z)\left(\Delta Y - E(\Delta Y|Z_1, S^I = 0, D_1)\right)\right]}{E\left[\text{sgn}(\Delta Z)\left(\Delta D - E(\Delta D|Z_1, S^I = 0, D_1)\right)\right]} = \delta_{IV}.$$

For the propensity-score estimand, notice that

$$\begin{aligned}
&\frac{E\left[\text{sgn}(\Delta Z)\left(\Delta Y - E(\Delta Y|Z_1, S^I = 0, D_1)\right)\right]}{E\left[\text{sgn}(\Delta Z)\left(\Delta D - E(\Delta D|Z_1, S^I = 0, D_1)\right)\right]} \\
=&\frac{E\left[\text{sgn}(\Delta Z)\Delta Y\right] - E\left[\text{sgn}(\Delta Z)E(\Delta Y|Z_1, S^I = 0, D_1)\right]}{E\left[\text{sgn}(\Delta Z)\Delta D\right] - E\left[\text{sgn}(\Delta Z)E(\Delta D|Z_1, S^I = 0, D_1)\right]}.
\end{aligned}$$

Then, using $\text{sgn}(\Delta Z) = S^I_+ - S^I_-$, the law of iterated expectations and Assumption 9,

$$\begin{aligned}
E\left[\text{sgn}(\Delta Z)E(\Delta D|Z_1, S^I = 0, D_1)\right] &= E\left[(S^I_+ - S^I_-)E\left(\frac{\Delta D(1-S^I)}{P(S^I=0|Z_1,D_1)}\middle| Z_1, D_1\right)\right] \\
&= E\left[E(S^I_+ - S^I_-|Z_1, D_1)E\left(\frac{\Delta D(1-S^I)}{P(S^I=0|Z_1,D_1)}\middle| Z_1, D_1\right)\right] \\
&= E\left[E\left(\frac{\Delta D(1-S^I)E(S^I_+ - S^I_-|Z_1,D_1)}{P(S^I=0|Z_1,D_1)}\middle| Z_1, D_1\right)\right] \\
&= E\left[\frac{\Delta D(1-S^I)E(S^I_+ - S^I_-|Z_1,D_1)}{P(S^I=0|Z_1,D_1)}\right] \\
&= E\left[\Delta D\frac{E(S^I_+ - S^I_-|Z_1,D_1)}{P(S^I=0|Z_1,D_1)}P(S^I=0)\middle| S^I=0\right] \\
&= E\left[\Delta D\frac{P(S^I_+=1|Z_1,D_1) - P(S^I_-=1|Z_1,D_1)}{P(S^I=0|Z_1,D_1)}\middle| S^I=0\right] \\
&\times P(S^I=0),
\end{aligned}$$

as required. The same argument replacing ΔD by ΔY completes the proof □

9.8 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 □

9.9 Theorem 8

The proof is similar to that of Theorem 7, and is therefore omitted.