Quantile Treatment Effects in Difference-in-Discontinuities Designs

Yingying Dong*

March 23, 2026

Abstract

This paper studies identification and doubly robust (DR) estimation of quantile treatment effects (QTEs) in difference-in-discontinuities (diff-in-disc) designs. We show that QTEs are point identified under a conditional stable-distributional-effect assumption for the confounding treatment, which is the diff-in-disc counterpart of the distributional parallel trends assumption in the recent difference-in-differences literature. We then propose a DR estimator and inference procedure that remain valid when either the outcome regression or the propensity score model is correctly specified, while avoiding high-dimensional nonparametric adjustment for covariates in the local estimation setting. We establish asymptotic normality of the proposed estimator, and Monte Carlo simulations illustrate its double-robustness and good finite-sample performance. In an application to Italian municipal fiscal data, extending the mean analysis of Grembi et al. (2016), the estimated effects suggest that relaxing fiscal constraints increases deficits mainly in the lower and middle parts of the distribution, moving municipalities near fiscal balance into moderate deficits—heterogeneity that is not visible in mean effects.

Key words: Quantile treatment effects, Diff-in-disc, Regression discontinuity design, Difference-in-differences, Doubly Robust, Fiscal policy

JEL classification: C13 C21 H71 H72

1 Introduction

The difference-in-discontinuities (diff-in-disc) design combines elements of regression discontinuity design (RDD) and difference-in-differences (DiD) to identify causal effects when a new treatment is introduced at a cutoff that is already associated with a discontinuity from a pre-existing policy or institutional rule. In an RDD, treatment assignment changes discretely at a known cutoff of a running variable, and the resulting discontinuity in outcomes identifies a local

*Yingying Dong: University of California Irvine, email: yyd@uci.edu

treatment effect at the threshold. In DiD, differences in pre–post changes between treated and control groups are used to identify treatment effects under a parallel-trends assumption. Diff-in-disc combines these ideas by contrasting RDD estimates before and after a policy intervention, thereby differencing out pre-existing discontinuities at the same threshold.

The diff-in-disc design was introduced by Grembi et al. (2016) in their study of fiscal rules, and subsequent work has refined identification and estimation strategies within this framework (e.g., Millán-Quijano, 2020; Galindo-Silva et al., 2021; Butts, 2023; Larsen and Valant, 2024; Picchetti et al., 2024). Existing studies, however, have focused almost exclusively on mean treatment effects. Mean effects can conceal substantial heterogeneity across the outcome distribution, especially when a policy shifts some parts of the distribution much more than others. This paper extends the diff-in-disc framework to quantile treatment effects (QTEs), enabling the analysis of distributional impacts that are not visible in mean effects.

Our first contribution is an identification result for QTEs in diff-in-disc designs. We show that QTEs are point identified under a conditional stable distributional effect assumption for the confounding treatment. This assumption is the diff-in-disc counterpart of the distributional parallel trends restriction in recent DiD research (Roth and Sant'Anna, 2023; Kim and Wooldridge, 2025): it requires the discontinuity generated by the confounding treatment to remain stable over time, conditional on observables. The conditional formulation allows time-varying confounding to operate through observed covariates and is especially useful when the composition of units near the cutoff changes over time. Because the identifying restriction is stated in terms of conditional distributions, it is invariant to strictly monotonic transformations of the outcome variable, making the framework applicable whether the outcome is measured in levels, logarithms, or another monotone scale.

Our second contribution is methodological. We propose a doubly robust (DR) estimator and inference procedure for the identified QTEs. The estimator remains consistent when either the outcome regression or the propensity score model is correctly specified, and it avoids high-dimensional nonparametric adjustment for covariates in the local estimation setting. We establish asymptotic normality of the proposed DR estimator and discuss the multiplier wild bootstrap as a convenient method for practical inference.

We study the finite-sample behavior of the estimator in a small-scale Monte Carlo exercise. The simulations illustrate its double-robustness and good finite-sample performance. We also apply the method to Italian municipal fiscal data, extending the mean analysis of Grembi et al. (2016). In this application, point estimates suggest that relaxing fiscal constraints increases deficits mainly in the lower and middle parts of the distribution, moving municipalities near fiscal balance into moderate deficits. These patterns suggest heterogeneity that may be masked by mean effects.

This paper contributes to two strands of literature. First, it advances the diff-in-disc literature by providing, to our knowledge, the first framework for identifying and estimating QTEs in this setting, together with DR inference.

Second, it contributes to the broader literature on distributional and quantile treatment effects, including conditional and unconditional QTEs under unconfoundedness (Koenker and Bassett, 1978; Firpo, 2007), regression discontinuity designs for quantiles (Frandsen et al., 2012), and extensions of DiD under distributional parallel trends (Athey and Imbens, 2006; Callaway and Sant'Anna, 2021; Roth and Sant'Anna, 2023). By combining these strands, the paper provides a new tool for studying heterogeneous policy effects in settings with both discontinuity-based and before–after variation.

The remainder of the paper is organized as follows. Section 2 presents the baseline identification result without covariates. Section 3 extends the analysis to allow for time-varying confounding through observables. Section 4 discusses estimation and inference. Section 5 presents Monte Carlo simulations. Section 6 contains the empirical application. Section 7 concludes.

2 Baseline Identification

Let $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ be the outcome of interest for unit i. We consider two time periods, or more generally, two groups. Let $T_i \in \{0,1\}$ be an indicator for the second period/group: $T_i = 0$ for the first period or group and $T_i = 1$ for the second. There are two binary treatments: a confounding treatment $C_i \in \{0,1\}$, present in both periods/groups; the treatment of interest $D_i \in \{0,1\}$, introduced only in period/group $T_i = 1$. Our goal is to evaluate the effects of D_i. For concreteness, we refer to T_i as a time indicator, but the framework also covers setups where T_i indexes groups rather than time. In such cases, the treated group experiences both C_i and D_i, while the control group receives only C_i. An example is Lalive (2008), where unemployment insurance benefits change discontinuously with age only in the treated region. Let R_i be the running variable. We assume a repeated cross-sectional setup, i.e., $(Y_i, R_i, C_i, D_i)\,|T_i = t \overset{iid}{\sim} F_{Y,R,C,D|T=t}$, $t = 0, 1$.

Both treatments follow a sharp RDD rule with a known cutoff, normalized to zero. Specifically:

$$C_i = 1\left(R_i \geq 0\right), \tag{1}$$

$$D_i = 1\left(R_i \geq 0\right) \cdot 1\left(T_i = 1\right). \tag{2}$$

The outcome is generated as:

$$Y_i = g\left(C_i, D_i, R_i, e_i\right), \tag{3}$$

where $g\left(\cdot\right)$ is an unknown function, and e_i captures all other observable or unobservable determinants of Y_i, in addition to C_i, D_i and R_i.

Potential outcomes are defined by fixing treatment values:

$$Y_i\left(c,d\right) := g\left(c,d,R_i,e_i\right),\ c,d \in \{0,1\}.$$

The observed outcome is then

$$Y_i = (1 - C_i)(1 - D_i) \cdot Y_i(0,0) + C_i(1 - D_i) \cdot Y_i(1,0) + (1 - C_i) D_i \cdot Y_i(0,1) + C_i D_i \cdot Y_i(1,1). \quad (4)$$

which simplifies to $Y_i = (1 - C_i) \cdot Y_i(0,0) + C_i \cdot Y_i(1,0)$ when $T_i = 0$ (since $D_i = 0$).

The causal parameter of interest is the quantile treatment effect (QTE) of D_i in the presence of C_i, at the threshold $R_i = 0$ for $T_i = 1$, i.e.,

$$\delta(\tau) := Q_{Y_i(1,1)|R_i=0,T_i=1}(\tau) - Q_{Y_i(1,0)|R_i=0,T_i=1}(\tau), \text{ for } \tau \in (0,1),$$

where $Q_{W_i|R_i=0,T_i=1}(\tau)$ is the conditional quantile function of a random variable W_i given $R_i = 0$ and $T_i = 1$. This parameter captures how D_i shifts the outcome distribution in the presence of C_i rather than individual treatment effects.

Notation rule: In the following we drop the i subscript. For conditional cumulative distribution functions (CDFs) and probability density functions (PDFs), conditioning on $R = r$ and $T = t$ appears in the subscript; additional conditioning variable (e.g., covariates X) appears in the argument. For instance, $F_{W|R=r,T=t}(w) := \Pr(W \leq w | R = r, T = t)$ denotes the conditional CDF of W given $R = r$ and $T = t$ (for $t = 0, 1$), whereas $F_{W|X,R=r,T=t}(w|x) := \Pr(W \leq w | X = x, R = r, T = t)$ denotes the conditional CDF of W given $X = x$, $R = r$, and $T = t$. Similarly, $f_{R|T=t}(r)$ and $f_{X|R=r,T=t}(x)$ denote the conditional PDFs of R given $T = t$ (for $t = 0, 1$) and of X given $R = r$ and $T = t$, respectively. Because R is continuously distributed, events like $\{R = 0\}$ have probabilities zero. Throughout, conditioning on $R = 0$ denotes the limiting distribution as $r \to 0$.

Assumption 1 (Smoothness): (i) For any $y \in \mathcal{Y}$, the CDFs $F_{Y(c,0)|R=r,T=0}(y)$, $c \in \{0,1\}$, and $F_{Y(c,d)|R=r,T=1}(y)$, $c, d \in \{0,1\}$, are continuous in r at $r = 0$. (ii) $f_{R|T=t}(r)$ is continuous and strictly positive in a neighborhood of $r = 0$ for $t = 0, 1$.

Assumption 2 (Stable Distributional Effect of C): For any $y \in \mathcal{Y}$,

$$F_{Y(1,0)|R=0,T=0}(y) - F_{Y(0,0)|R=0,T=0}(y) = F_{Y(1,0)|R=0,T=1}(y) - F_{Y(0,0)|R=0,T=1}(y).$$

Assumption 1 is the standard smoothness condition for identifying distributional effects in RDD (Frandsen et al., 2012). Assumption 2 states that the distributional effect of C, in the absence of D, is invariant across periods. In identifying mean effects, Millán-Quijano (2020), Larsen and Valant (2024), and Picchetti et al. (2024) impose the "stable mean effect" while Grembi et al. (2016) and Galindo-Silva et al. (2021) impose a stronger assumption that the individual treatment effect of C, in the absence of D, remains the same for $T = 0$ and $T = 1$, which implies that both the mean and the distributional effects remain the same.

Assumption 2 is the difference-in-discontinuities counterpart of the distributional parallel trends assumption in DiD (Roth and Sant'Anna, 2023; Kim and Wooldridge, 2025). In the absence of the confounding treatment C, the usual RDD logic implies that units just above and just below the cutoff are locally comparable, providing a design-based rationale for a stable above-below distributional contrast in the absence of D. With C present in both periods, Assumption 2 requires that the local distributional contrast generated by C at the cutoff be the same in $T = 0$ and $T = 1$. This allows the pre-period discontinuity due to C to be differenced out from the post-period discontinuity, thereby isolating the distributional effect of D. As with standard distributional identification, the assumption is invariant to strictly monotonic transformations of the outcome. This assumption is most plausible when C is a longstanding or institutionally stable policy, when its implementation does not change materially across periods, or when the pre- and post-periods are sufficiently close that any mild time variation in the effect of C is likely to be small.

Although Assumption 2 is not directly testable, its plausibility can be informally assessed when multiple pre-periods are available. A natural placebo exercise is to re-estimate the quantile effects exactly as in the main analysis using only pre-period data, treating an earlier pre-period as the base period and a later pre-period as a placebo post period. Since D is absent in both periods, placebo QTE estimates close to zero across quantiles would be consistent with the maintained identifying framework and, in particular, with stability of the discontinuity generated by C. Importantly, however, this placebo exercise is not a test of Assumption 2 alone, but a joint check of the full set of identifying assumptions underlying the placebo comparison. Thus, rejection may reflect instability in the effect of C or other departures from the maintained framework. Likewise, passing the placebo exercise is only suggestive, since stability across pre-periods need not extend to the actual post period. The exercise should therefore be interpreted as a plausibility check rather than a formal validation of Assumption 2.

Parallel trends in distributions are a stronger assumption than parallel trends in means; however, as discussed in Chen and Roth (2024), assuming parallel trends in distributions is necessary (and sufficient) for the parallel trends in means assumption to be invariant to any monotonic transformations of the outcome variable. Following their argument (Chen and Roth, 2024, Proposition 1), it is easy to show that Assumption 2 holds, if and only if for g: $\mathcal{Y} \to g(\mathcal{Y})$, a strictly monotonic bijection on $\mathcal{Y}$, the following holds

$$\begin{aligned} & F_{g\left(Y(1,0)\right)|R=0,T=0}(s) - F_{g\left(Y(0,0)\right)|R=0,T=0}(s) \\ = \; & F_{g\left(Y(1,0)\right)|R=0,T=1}(s) - F_{g\left(Y(0,0)\right)|R=0,T=1}(s) \text{ for any } s \in g(\mathcal{Y}). \end{aligned}$$

One example of such g function is $g(y) = \log(y)$ for $y \in (0, +\infty)$. It follows that the identification results presented in this paper are robust to monotonic transformation, including the popular log transformation of the outcome.

Given the above assumptions, one can identify the two distribution functions, $F_{Y(1,1)|R=0,T=1}(y)$ and $F_{Y(1,0)|R=0,T=1}(y)$. Since C and D follow sharp designs,

i.e., $C = 1(R \geq 0)$, and $D = 1(R \geq 0) \cdot 1(T = 1)$, we have $Y = Y(1,1)$ when $R \geq 0$ and $T = 1$. Then under Assumption 1,

$$F_{Y(1,1)|R=0,T=1}(y) = \lim_{r \downarrow 0} \mathbb{E}[I(y)|R = r, T = 1], \tag{5}$$

where $I(y) := 1(Y \leq y)$.

By Assumption 2, stable distributional effect of C, we have

$$\begin{aligned} F_{Y(1,0)|R=0,T=1}(y) &= F_{Y(0,0)|R=0,T=1}(y) \\ &\quad + F_{Y(1,0)|R=0,T=0}(y) - F_{Y(0,0)|R=0,T=0}(y) \\ &= \lim_{r \uparrow 0} \mathbb{E}[I(y)|R = r, T = 1] \\ &\quad + \lim_{r \downarrow 0} \mathbb{E}[I(y)|R = r, T = 0] \\ &\quad - \lim_{r \uparrow 0} \mathbb{E}[I(y)|R = r, T = 0], \end{aligned} \tag{6}$$

where the second equality follows from (1), (2), (4), and the smoothness conditions in Assumption 1.

The quantile treatment effect of interest is then given by

$$\delta(\tau) = F^{-1}_{Y(1,1)|R=0,T=1}(\tau) - F^{-1}_{Y(1,0)|R=0,T=1}(\tau),$$

where $F^{-1}_{Y(1,d)|R=0,T=1}(\tau) = \inf\{y \in \mathcal{Y}\colon F_{Y(1,d)|R=0,T=1}(y) \geq \tau\}$, $d = 0, 1$.

Under Assumptions 1 and 2, the right-hand side of eq. (6) is a proper CDF and therefore must be weakly increasing in y and take values in $[0, 1]$. This yields the testable restriction: for any $y', y \in \mathcal{Y}$, and $y' < y$,

$$\begin{aligned} &\lim_{r \uparrow 0} \mathbb{E}[I(y') - I(y)|R = r, T = 1] \\ &+ \lim_{r \downarrow 0} \mathbb{E}[I(y') - I(y)|R = r, T = 0] \\ &- \lim_{r \uparrow 0} \mathbb{E}[I(y') - I(y)|R = r, T = 0] \\ \leq\ & 0. \end{aligned} \tag{7}$$

A natural route to formal testing would be to estimate the implied counterfactual CDF on a finite grid of outcome values and test the resulting collection of one-sided moment inequalities—monotonicity across grid points, together with the $[0, 1]$ bounds—using a studentized Kolmogorov–Smirnov-type statistic with bootstrap critical values, in the spirit of Arai et al. (2022). An alternative, closer to Roth and Sant'Anna (2023), would be to test nonnegativity of the implied probability masses (or density) associated with the counterfactual CDF. Because a full treatment would need to account for boundary local-polynomial estimation in the diff-in-disc setting, we leave formal testing of eq. (7) for future research.

3 Identification with Time Varying Confounding Factors

The identification strategy in the preceding setup rests on the stable distributional effect of C (Assumption 2), which requires the discontinuity generated by C at the cutoff, in the absence of D, to be the same in the two periods. This restriction can be too strong if the composition of units near the cutoff changes over time, or if period-specific factors alter how C affects the outcome distribution.

We now extend the framework to allow such differential trends to operate through a set of observed covariates $X \in \mathbb{R}^p$. Throughout this section, X denotes observed covariates that may vary across periods but are not themselves affected by the cutoff-induced treatments C and D. This restriction is important because identification conditions on X and transports the pre-period conditional effect to the post-period covariate distribution. This generalized setting retains the previous diff-in-disc structure but replaces Assumption 2 with a conditional analogue, in which the stability restriction holds after conditioning on X. Unlike in standard RDD, where covariates are included mainly to improve efficiency, here conditioning on X is essential for identification. The approach also accommodates a doubly robust formulation that remains valid when either the outcome model or the relevant propensity score is correctly specified.

Let

$$\Delta_t(y,x) := F_{Y(1,0)|X,R=0,T=t}(y|x) - F_{Y(0,0)|X=x,R=0,T=t}(y|x),$$

for $t = 0, 1$. Further let $p(x) := \Pr(T = 1|R = 0, X = x)$.

Assumption 1G (Smoothness): (i) $F_{Y(c,0)|X,R=r,T=0}(y|x)$ for $c \in \{0,1\}$ and $F_{Y(c,d)|R=r,T=1}(y)$ for $c, d \in \{0,1\}$ are continuous in r at $r = 0$. (ii) $f_{X|R=r,T=t}(x)$ is continuous in r at $r = 0$ for $t \in \{0,1\}$. (iii) $f_{R|T=t}(r)$ is continuous and bounded away from zero at $r = 0$ for $t \in \{0,1\}$.

Assumption 2G (Conditional Stable Distributional Effect): $\Delta_0(y,x) = \Delta_1(y,x)$ for all (y,x).

Assumption 3G (One-sided local overlap): There exists $\varepsilon > 0$ such that $\Pr(p(X) \leq 1 - \varepsilon | R = 0, T = 1) = 1$.

The asymmetry in Assumption 1G(i) is intentional: conditional continuity in X is only imposed on $T = 0$, because $\Delta_0(y,x)$ is identified from pre-period one-sided limits and then transported to $T = 1$ via Assumption 2G; for $T = 1$, unconditional continuity is sufficient to identify $F_{Y(1,1)|R=0,T=1}(y)$ and $F_{Y(0,0)|R=0,T=1}(y)$. Assumption 1G(ii) should be understood as a continuity restriction on the distribution of the untreated covariates X at the cutoff; variables that respond to C and D are excluded from the conditioning set in the current framework. Assumption 1G strengthens Assumption 1 by requiring smoothness for $T = 0$ to hold conditional on X. In particular, parts (i)–(ii) imply the unconditional continuity in Assumption 1, making Assumption 1G

slightly stronger in theory. In practice, the difference is negligible, as smoothness in observables is commonly required in RDD applications—discontinuity in the distribution of observables at the cutoff is generally viewed as evidence invalidating the design.[1]

Assumption 2G relaxes Assumption 2 by allowing the distributional effect of C to vary across periods with observed covariates X. Conditional on X, however, the effect of C at the cutoff must remain stable across $T = 0$ and $T = 1$, so that the pre-period object $\Delta_0(y, x)$ can be transported to the post-period covariate distribution. Thus, Assumption 2G is a conditional transportability restriction: it allows cross-period compositional changes near the cutoff, provided they are captured by X. As with Assumption 2, it is invariant to strictly monotonic transformations of the outcome.

Assumption 2G can be informally probed using a placebo exercise based only on pre-period data when multiple pre-periods are available. One can re-estimate the same covariate-adjusted specification used in the main analysis, treating an earlier pre-period as the base period and a later pre-period as a placebo post period. Placebo QTE estimates close to zero across quantiles would be consistent with the maintained identifying framework and, in particular, with conditional stability of the discontinuity generated by C. Conversely, sizable placebo effects even after conditioning on X would suggest either that the effect of C changes over time in ways not captured by X, or that the covariates are not rich enough to absorb the relevant heterogeneity. Comparing unconditional and conditional placebo results can also be informative: rejection of the former but not the latter would suggest that observable compositional change is the main source of time variation. As before, however, this exercise is not a test of Assumption 2G alone, but a joint check of the full set of identifying assumptions underlying the placebo comparison. Thus, rejection may reflect failure of conditional stability, violations of the underlying RDD conditions, or other departures from the maintained framework. Likewise, passing the placebo exercise is only suggestive, since stability across pre-periods need not extend to the actual post period. The exercise should therefore be interpreted as a plausibility check rather than a formal validation of Assumption 2G.

Assumption 3G states that there exists $\varepsilon > 0$ such that $p(X) \leq 1 - \varepsilon$ a.s., with "a.s." taken under the $T = 1$ local distribution at the cutoff. It allows $p(X) = 0$ but rules out $p(X) = 1$ on covariate values relevant for $T = 1$. Let $\mathcal{S}_t{:=}Supp\,(X|R = 0, T = t)$. Assumption 3G implies support inclusion: $\mathcal{S}_1 \subseteq \mathcal{S}_0$, enabling extrapolation of the effect of C from $T = 0$ to $T = 1$. More importantly, Assumption 3G requires $p(X)$ to be bounded sufficiently away from 1, which ensures stability of inverse propensity score type of reweighting, which the proposed DR approach relies on.

[1] The current framework may be extended to accommodate discontinuities in the distribution of covariates at the RDD cutoff. This can be done following Frölich and Huber (2019), at the cost of more cumbersome notations. We do not pursue it here as it is not the main focus of the paper.

Under Assumption 1G,

$$F_{Y(1,1)|R=0,T=1}(y) = \lim_{r\downarrow 0} \mathbb{E}\left[I(y) \,|R = r, T = 1\right] \tag{8}$$

is identified as before. Let

$$\Delta(y) := F_{Y(1,0)|R=0,T=1}(y) - F_{Y(0,0)|R=0,T=1}(y).$$

Then

$$F_{Y(1,0)|R=0,T=1}(y) = F_{Y(0,0)|R=0,T=1}(y) + \Delta(y), \tag{9}$$

where the first term

$$F_{Y(0,0)|R=0,T=1}(y) = \lim_{r\uparrow 0} \mathbb{E}\left[I(y) \,|R = r, T = 1\right] \tag{10}$$

is identified from data with $T = 1$ and $R \leq 0$. To identify the second term $\Delta(y)$, note from Assumption 2G that $\Delta_1(y, x) = \Delta_0(y, x)$, and from Assumption 1G,

$$\Delta_0(y, x) = \lim_{r\downarrow 0} \mathbb{E}\left[I(y) \,|X = x, R = r, T = 0\right] - \lim_{r\uparrow 0} \mathbb{E}\left[I(y) \,|X = x, R = r, T = 0\right]. \tag{11}$$

By the law of iterated expectations and Assumption 3G,

$$\begin{aligned} \Delta(y) &= \mathbb{E}\left[\Delta_1(y, X) \,|R = 0, T = 1\right] \\ &= \mathbb{E}\left[\Delta_0(y, X) \,|R = 0, T = 1\right], \end{aligned} \tag{12}$$

where $\Delta_0(y, x)$ is given by eq. (11).

In contrast to the canonical RDD—where conditioning on covariates improves efficiency but is not needed for identification—conditioning on X is essential here to recover the counterfactual distribution $F_{Y(1,0)|R=0,T=1}(y)$.

If T is randomized or X is stable over time,
$f_{X|R=0,T=0}(x) = f_{X|R=0,T=1}(x)$, the above simplifies to

$$\Delta(y) = \lim_{r\downarrow 0} \mathbb{E}\left[I(y) \,|R = r, T = 0\right] - \lim_{r\uparrow 0} \mathbb{E}\left[I(y) \,|R = r, T = 0\right]. \tag{13}$$

The conditional approach in (12) – which requires estimating the conditional effect $\Delta_0(y, x)$ and then averaging it over the distribution of covariates X at $R = 0$ for $T = 1$ – can still yield efficiency gains over (13) (See Frölich and Huber, 2019).

Alternatively, $\Delta(y)$ can be identified from using inverse propensity score weighting. Let $p(x) := \Pr(T = 1 \mid X = x, R = 0)$ and $p := \Pr(T = 1 \mid R = 0) = \mathbb{E}[p(X) \mid R = 0]$. Define the odds $\pi(x) := \frac{p(x)}{1-p(x)}$ and the normalized weight

$$w(x) := \frac{\pi(x)}{\mathbb{E}[\pi(X) \mid R = 0, T = 0]} = \frac{(1-p)\, p(x)}{p\,(1-p(x))}. \tag{14}$$

The last equality follows from the standard identity $\mathbb{E}[\pi(X) \mid R = 0, T = 0] = \frac{p}{1-p}$. Assumption 3G ensures that $w(x)$ is finite almost surely under the $T = 1$

local distribution. This weight satisfies the identity, for any integrable function $G(X)$,

$$\mathbb{E}[G(X) \mid R=0, T=1] = \mathbb{E}[w(X)\, G(X) \mid R=0, T=0]. \tag{15}$$

Then

$$\Delta(y) = \lim_{r \downarrow 0} \mathbb{E}\left[w(X) \cdot I(y) \,|R=r, T=0\right] - \lim_{r \uparrow 0} \mathbb{E}\left[w(X) \cdot I(y) \,|R=r, T=0\right]. \tag{16}$$

A proof of Eqs. (15) and (16) is provided in the Appendix.

Eqs. (12) and (16) require estimating conditional mean or probability functions given R, T and X. Nonparametric estimation can suffer from the curse of dimensionality when X is high-dimensional and worse when involves many continuous covariates; parametric models risk misspecification. We therefore consider a doubly robust approach that mitigates the misspecification concern.

Let $F_0^+(y, X)$ and $F_0^-(y, X)$ be some chosen parametric models for the conditional mean functions

$$\lim_{r \downarrow 0} \mathbb{E}\big[I(y) \mid X, R=r, T=0\big] \quad \text{and} \quad \lim_{r \uparrow 0} \mathbb{E}\big[I(y) \mid X, R=r, T=0\big],$$

respectively. Let $\widetilde{w}(x)$ be the IPW weights in (14) when the propensity score $p(x)$ is estimated parametrically, such as by logit or probit. Define

$$\Delta^{DR}(y) := \Delta_1(y) + \Delta_2^+(y) - \Delta_2^-(y), \tag{17}$$

with

$$\begin{aligned}
\Delta_1(y) &:= \mathbb{E}\big[F_0^+(y, X) - F_0^-(y, X) \mid R=0, T=1\big], \\
\Delta_2^+(y) &:= \lim_{r \downarrow 0} \mathbb{E}\big[\widetilde{w}(X)\{I(y) - F_0^+(y, X)\} \\
&\qquad\qquad \mid R=r, T=0\big], \\
\Delta_2^-(y) &:= \lim_{r \uparrow 0} \mathbb{E}\big[\widetilde{w}(X)\{I(y) - F_0^-(y, X)\} \\
&\qquad\qquad \mid R=r, T=0\big].
\end{aligned}$$

Theorem 1 (QTE Identification) *Under Assumptions 1G - 3G,*

$$\begin{aligned}
\delta(\tau) = \inf\left\{y \in \mathcal{Y}\text{: } F_{Y(1,1)|R=0,T=1}(y) \geq \tau\right\} \\
- \inf\left\{y \in \mathcal{Y}\text{: } F_{Y(1,0)|R=0,T=1}(y) \geq \tau\right\},
\end{aligned}$$

where

$$\begin{aligned}
F_{Y(1,1)|R=0,T=1}(y) &= \lim_{r \downarrow 0} \mathbb{E}\left[I(y) \,|R=r, T=1\right], \\
F_{Y(1,0)|R=0,T=1}(y) &= \lim_{r \uparrow 0} \mathbb{E}\left[I(y) \,|R=r, T=1\right] + \Delta^{DR}(y),
\end{aligned}$$

and $\Delta^{DR}(y)$ is given by eq. (17). The estimand for $F_{Y(1,0)|R=0,T=1}(y)$ is valid if either (i) the conditional mean models F_0^+ and F_0^- are correctly specified, or (ii) the weights $\widetilde{w}$ are correctly specified.

The proposed doubly robust formulation in Theorem 1 ensures validity under correct specification of either the outcome regression or the propensity score model. This formulation forms the basis for the estimation procedures developed in the next section.

Note that we focus on the case in which both the confounding treatment C and the treatment of interest D follow a sharp design. If both instead follow a fuzzy design, the QTEs $\delta(\tau)$ are generally not point identified without additional restrictions.

4 Estimation and Inference

For clarity and to simplify notation, estimation and inference are presented using a single bandwidth h common to both periods ($t \in \{0,1\}$) and both sides of the cutoff ($s \in \{+,-\}$), i.e., $h_\tau^{t,s} = h$ for all τ. In practice, one may choose different bandwidths depending on the time period, the side of the cutoff, and the specific quantile. We discuss detailed bandwidth choices in Section 4.3.

4.1 Estimation

For the main exposition, assume we have an i.i.d. sample $\{Y_i, R_i, X_i, T_i\}_{i=1}^n$, with treatment $T_i \in \{0,1\}$, running variable $R_i \in \mathcal{R} \subseteq \mathbb{R}$, covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}$, and outcome $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$. Define $I_i(y) := 1\{Y_i \leq y\}$ and the kernel $K_h(u) := K(u/h)/h$ with bandwidth h. Denote by $\widehat{\theta}$ the estimator of any parameter θ.

Step 1: Estimate local CDFs for $T = 1$. Let $F_{c,d}(y) = F_{Y(c,d)|R=0,T=1}(y)$, $c, d \in \{0,1\}$. Estimate $F_{1,1}(y)$ by $\widehat{a}_0^+(y)$ from the following local linear regression (LLR):

$$\left\{\widehat{a}_0^+(y), \widehat{a}_1^+(y)\right\} = \arg\min_{a_0^+, a_1^+} \sum_{i: R_i \geq 0,\, T_i = 1} \left(I_i(y) - a_0^+ - a_1^+ R_i\right)^2 K_h(R_i), \quad (18)$$

Similarly, estimate $F_{0,0}(y) = \lim_{r \uparrow 0} \mathbb{E}[I(y) \mid R = r, T = 1]$ by $\widehat{a}_0^-(y)$ from

$$\left\{\widehat{a}_0^-(y), \widehat{a}_1^-(y)\right\} = \arg\min_{a_0^-, a_1^-} \sum_{i: R_i < 0,\, T_i = 1} \left(I_i(y) - a_0^- - a_1^- R_i\right)^2 K_h(R_i). \quad (19)$$

Step 2: Estimate DR correction $\widehat{\Delta}^{DR}(y)$. Fit side-specific parametric working models for $E[I_i(y) \mid X_i, R_i, T_i = 0]$ on the subsamples $\{T_i = 0, R_i \geq 0\}$ and $\{T_i = 0, R_i < 0\}$, and denote by $\widehat{F}_0^+(y, X_i)$ and $\widehat{F}_0^-(y, X_i)$ the corresponding fitted values evaluated at $R_i = 0$. Further estimate $\Delta_1(y)$ by $\widehat{a}(y)$ from LLR using the pseudo-outcome $\widehat{\Delta}(y, X_i) := \hat{F}_0^+(y, X_i) - \hat{F}_0^-(y, X_i)$ and the subsample $\{T_i = 1\}$, i.e.,

$$\left\{\widehat{a}(y), \widehat{b}(y)\right\} = \arg\min_{a,b} \sum_{i: T_i = 1} \left(\widehat{\Delta}(y, X_i) - a - bR_i\right)^2 K_h(R_i).$$

Fit $Pr(T=1 \mid X, R)$ by logit/probit and define $\widehat{p}(X_i) = \widehat{Pr}(T=1 \mid X_i, R_i=0)$. Set $\widehat{\pi}(X_i) = \hat{p}(X_i) / (1-\hat{p}(X_i))$. Further estimate $\widehat{\pi} = \mathbb{E}[\widehat{\pi}(X) | R=0, T=0]$ by a parametric intercept, regressing $\widehat{\pi}(X_i)$ on R_i using the subsample $\{T_i=0\}$. Alternatively, one may estimate π by local constant estimator i.e., $\widehat{\pi} = \frac{\sum_{i:T_i=0} \widehat{\pi}(X_i) K_{h_\pi}(R_i)}{\sum_{i:T_i=0} K_{h_\pi}(R_i)}$, choosing $h_\pi \gg h$ so that the estimation error of $\widehat{\pi}$ is first-order ignorable. Set

$$\hat{w}(X_i) = \frac{\widehat{\pi}(X_i)}{\widehat{\pi}}. \tag{20}$$

Estimate $\Delta_2^+(y)$ and $\Delta_2^-(y)$ by LLRs as in (18) and (19), replacing $I_i(y)$ with the pseudo-outcome $\hat{w}(X_i)\{I_i(y) - \hat{F}_0^+(y, X_i)\}$ for $T_i = 0, R_i \geq 0$ and with $\hat{w}(X_i)\{I_i(y) - \hat{F}_0^-(y, X_i)\}$ for $T_i = 0, R_i < 0$. Then

$$\hat{\Delta}^{DR}(y) = \hat{\Delta}_1(y) + \hat{\Delta}_2^+(y) - \hat{\Delta}_2^-(y), \tag{21}$$

$$\hat{F}_{1,0}(y) = \hat{F}_{0,0}(y) + \hat{\Delta}^{DR}(y). \tag{22}$$

Step 3: Estimate quantiles. Let $q_{1,d,\tau}$ be the τ-quantile of $F_{1,d}$ and $\widehat{q}_{1,d,\tau}$ its estimator obtained by inverting the monotone rearrangements of $\widehat{F}_{1,d}$; rearrangement is Hadamard directionally differentiable and does not affect first-order asymptotics (Chernozhukov et al., 2010). The estimated QTE is then

$$\widehat{\delta}(\tau) = \widehat{q}_{1,1,\tau} - \widehat{q}_{1,0,\tau}, \tau \in (0,1).$$

In practice, it might be helpful to adopt smoothing "in the y-direction" to estimate smoothed versions of CDFs, which in our setting amounts to replacing $\mathbf{1}\{Y_i \leq y\}$ with $\Phi((y - Y_i)/h_Y)$, where Φ is a smooth CDF (e.g., the standard normal), in all the estimation steps described above. Choose $h_Y \to 0$ with $h_Y/h \to 0$ so that y-smoothing is second order and does not affect the first-order limit of $\widehat{\delta}(\tau)$.

4.2 Inference

Define $n_t := \sum_{i=1}^n \mathbf{1}\{T_i = t\}$ for $t \in \{0,1\}$ as the period-specific sample sizes, and let $n := n_0 + n_1$. Let $n_t h$ denote the period-t effective local sample size at the boundary ($t \in \{0,1\}$), and nh the overall effective local sample size when pooling both periods (with $n := n_0 + n_1$).

We impose the following assumptions and regularity conditions for asymptotics.

Assumption A (Asymptotics): (i) Assumptions 1G–3G from the identification section; (ii) **Smoothness:** the conditional expectations entering (8), (10), and (17) are twice continuously differentiable in r near 0; for $t \in \{0,1\}$, there exists $\delta > 0$ such that $f_{R|T=t}(r)$ is continuous and strictly positive for $|r| < \delta$ (as in Assumption 1G(iii)); (iii) **Kernel:** K is bounded, symmetric,

nonnegative with compact support and $\int K = 1$; (iv) **Bandwidth:**

$$h \to 0, \quad nh \to \infty, \quad \sqrt{nh}\, h^2 \to \gamma \in [0, \infty), \quad \frac{n_t}{n} \to p_t \in (0,1),$$
$$\text{so} \quad \sqrt{n_t h}\, h^2 \to \sqrt{p_t}\, \gamma =: \gamma_t \text{ for } t \in \{0,1\};$$

(v) **Double robustness:** either the parametric outcome regressions $F_0^{\pm}(y,x)$ are correctly specified or the propensity score $p(x)$ is correctly specified; (vi) **Quantile regularity:** for $d \in \{0,1\}$ the cdf $F_{1,d}(y)$ is absolutely continuous at $q_{1,d,\tau}$ with density $f_{1,d}(q_{1,d,\tau})$ bounded away from 0 and ∞ on the index set of interest.

Following the standard asymptotic linear representation of local polynomial regression estimators applied to CDFs, with indicator outcomes $I(Y \le y)$, one can derive the influence functions for the CDF estimators $\widehat{F}_{1,d}$, $d \in \{0,1\}$. For $t \in \{0,1\}$ and $s \in \{+,-\}$ define the one-sided sample moments

$$S_{m,s}^{(t)}(h) := \mathbb{E}\left[\left(\frac{R}{h}\right)^m K_h(R)\, \mathbf{1}\{T = t,\ sR \ge 0\}\right],\ m = 0,1,2,$$

and further

$$D_s^{(t)}(h) := S_{0,s}^{(t)}(h)\, S_{2,s}^{(t)}(h) - \big(S_{1,s}^{(t)}(h)\big)^2,$$

as well as the associated local-linear terms

$$\begin{aligned} \ell_+^{(t)}(R;h) &: = S_{2,+}^{(t)}(h) - S_{1,+}^{(t)}(h)\,(R/h), \\ \ell_-^{(t)}(R;h) &: = S_{2,-}^{(t)}(h) - S_{1,-}^{(t)}(h)\,(R/h). \end{aligned}$$

Let $W_i = (Y_i, R_i, X_i, T_i)$. Define the side-specific local mean functions

$$\begin{aligned} m_{1,+}(r;y) &:= \mathbb{E}\big[I(y) \mid R = r, T = 1, R \ge 0\big], \\ m_{1,-}(r;y) &:= \mathbb{E}\big[I(y) \mid R = r, T = 1, R < 0\big], \end{aligned}$$

so that $F_{1,1}(y) = m_{1,+}(0;y)$ and $F_{0,0}(y) = m_{1,-}(0;y)$. Similarly define

$$M_{\Delta_1}(r;y) := \mathbb{E}\big[F_0^+(y,X) - F_0^-(y,X) \mid R = r, T = 1\big],$$

so that $\Delta_1(y) = M_{\Delta_1}(0;y)$. Further for the DR corrections, define

$$\begin{aligned} M_{\Delta_2^+}(r;y) &:= \mathbb{E}\big[\widetilde{w}(X)\big(I(y) - F_0^+(y,X)\big) \mid R = r, T = 0, R \ge 0\big], \\ M_{\Delta_2^-}(r;y) &:= \mathbb{E}\big[\widetilde{w}(X)\big(I(y) - F_0^-(y,X)\big) \mid R = r, T = 0, R < 0\big], \end{aligned}$$

so that $\Delta_2^+(y) = M_{\Delta_2^+}(0;y)$ and $\Delta_2^-(y) = M_{\Delta_2^-}(0;y)$. All the influence functions below are written as equivalent-kernel weights times residuals from these local mean functions.

The boundary local linear estimators for $F_{1,1}(y)$ and $F_{0,0}(y)$ admit influence functions

$$\psi_{1,1,y}(W_i) = \frac{K_h(R_i)\,\mathbf{1}\{T_i = 1, R_i \geq 0\}\ell_+^{(1)}(R_i;h)}{D_+^{(1)}(h)}\left(I_i(y) - m_{1,+}(R_i;y)\right), \quad (23)$$

$$\psi_{0,0,y}(W_i) = \frac{K_h(R_i)\,\mathbf{1}\{T_i = 1, R_i < 0\}\ell_-^{(1)}(R_i;h)}{D_-^{(1)}(h)}\left(I_i(y) - m_{1,-}(R_i;y)\right). \quad (24)$$

For the components of the DR estimator for $F_{1,0}(y) = F_{0,0}(y)+\Delta_1(y)+\Delta_2^+(y)-\Delta_2^-(y)$, the associated influence functions are

$$\begin{aligned}\psi_{\Delta_1,y}(W_i) = {} & \frac{K_h(R_i)\mathbf{1}\{T_i = 1\}}{\mathbb{E}\big[K_h(R_i)\mathbf{1}\{T_i = 1\}\big]} \\ & \times \left(F_0^+(y,X_i) - F_0^-(y,X_i) - M_{\Delta_1}(R_i;y)\right), \quad (25)\\ \psi_{\Delta_2^+,y}(W_i) = {} & \frac{K_h(R_i)\,\mathbf{1}\{T_i = 0, R_i \geq 0\}\,\ell_+^{(0)}(R_i;h)}{D_+^{(0)}(h)}\big(\widetilde{w}(X_i)\big(I_i(y) - F_0^+(y,X_i)\big) \\ & - M_{\Delta_2^+}(R_i;y)\big), \quad (26)\\ \psi_{\Delta_2^-,y}(W_i) = {} & \frac{K_h(R_i)\,\mathbf{1}\{T_i = 0, R_i < 0\}\,\ell_-^{(0)}(R_i;h)}{D_-^{(0)}(h)}\big(\widetilde{w}(X_i)\big(I_i(y) - F_0^-(y,X_i)\big) \\ & - M_{\Delta_2^-}(R_i;y)\big). \quad (27)\end{aligned}$$

The expressions for $\psi_{1,1,y}$ and $\psi_{0,0,y}$ in eq.s (23) and (24) correspond to the canonical form given in Frandsen et al. (2012, hereafter FFM; Section 4 and Appendix C), adapted to our setup. The DR components influence functions $\psi_{\Delta_1,y}$, $\psi_{\Delta_2^+,y}$, and $\psi_{\Delta_2^-,y}$ extend these results to incorporate parametric adjustments and inverse probability weighting, in the spirit of doubly robust estimation (see, e.g., Robins et al., 1994; Chernozhukov et al., 2018). For $\Delta_1(y)$, the derivation in Appendix B.2(iii) shows that, under a symmetric kernel and evaluation at the interior point $R = 0$, the local linear weights admit the equivalent–kernel representation used in (25).

By linearity, reflecting the additive construction of $F_{1,0}(y)$ from its identified components, we can define the overall influence function for the DR estimator of $F_{1,0}(y)$ as

$$\begin{aligned}\psi_{1,0,y}(W_i) := {} & \mathbf{1}\{T_i = 1\}\frac{1}{\sqrt{p_1}}\big\{\psi_{0,0,y}(W_i) + \psi_{\Delta_1,y}(W_i)\big\} \\ & + \mathbf{1}\{T_i = 0\}\frac{1}{\sqrt{p_0}}\big\{\psi_{\Delta_2^+,y}(W_i) - \psi_{\Delta_2^-,y}(W_i)\big\}. \quad (28)\end{aligned}$$

The nuisance models are low-dimensional parametric, estimated at $n^{-1/2}$ rate, so their contribution is second order relative to $(nh)^{-1/2}$, and sample splitting/cross-fitting is therefore unnecessary here.

Theorem 2 (Asymptotic linearity of CDF estimators) *Under Assumption A, for each fixed y,*

$$\sqrt{n_1 h}\left(\widehat{F}_{1,1}(y) - F_{1,1}(y)\right) = \frac{1}{\sqrt{n_1 h}} \sum_{i:T_i=1} \psi_{1,1,y}(W_i) + \gamma_1\, b_{1,1}(y) + o_p(1), \quad (29)$$

$$\sqrt{nh}\left(\widehat{F}_{1,0}(y) - F_{1,0}(y)\right) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \psi_{1,0,y}(W_i) + \gamma\, b_{1,0}(y) + o_p(1), \quad (30)$$

where $b_{1,\cdot}(y)$ are the usual $O(h^2)$ boundary bias constants for LLR. The exact expressions of $b_{1,\cdot}(y)$ are provided in the Appendix.

Under Assumption A(vi), the quantile map is Hadamard differentiable, hence

$$\sqrt{n_1 h}\left(\widehat{q}_{1,1,\tau} - q_{1,1,\tau}\right) = -\frac{1}{\sqrt{n_1 h}} \sum_{i=1}^{n} \frac{\psi_{1,1,\,q_{1,1,\tau}}(W_i)}{f_{1,1}(q_{1,1,\tau})} + \gamma_1\, \beta_{1,1}(\tau) + o_p(1).$$

$$\sqrt{nh}\left(\widehat{q}_{1,0,\tau} - q_{1,0,\tau}\right) = -\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \frac{\psi_{1,0,\,q_{1,0,\tau}}(W_i)}{f_{1,0}(q_{1,0,\tau})} + \gamma\, \beta_{1,0}(\tau) + o_p(1).$$

with $\beta_{1,d}(\tau)$ for $d \in \{0,1\}$ the density-weighted drift induced by the $O(h^2)$ cdf bias. Therefore, for $\widehat{\delta}(\tau) = \widehat{q}_{1,1,\tau} - \widehat{q}_{1,0,\tau}$,

$$\sqrt{nh}\left(\widehat{\delta}(\tau) - \delta(\tau)\right) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \phi_\tau(W_i) + \gamma\, \mu(\tau) + o_p(1), \quad (31)$$

where

$$\phi_\tau(W_i) := -\frac{1}{\sqrt{p_1}} \frac{\psi_{1,1,\,q_{1,1,\tau}}(W_i)}{f_{1,1}(q_{1,1,\tau})} + \frac{\psi_{1,0,\,q_{1,0,\tau}}(W_i)}{f_{1,0}(q_{1,0,\tau})}$$

and $\mu(\tau) := \beta_{1,1}(\tau) - \beta_{1,0}(\tau)$.

Corollary 3 (Asymptotic normality of QTE) *Under Assumption A, for fixed $\tau \in (0,1)$,*

$$\sqrt{nh}\big(\hat{\delta}(\tau) - \delta(\tau)\big) \xrightarrow{d} \mathcal{N}\big(\gamma\, \mu(\tau),\ \sigma^2(\tau)\big),$$

where $\mu(\tau) \coloneqq \beta_{1,1}(\tau) - \beta_{1,0}(\tau)$ and $\sigma^2(\tau) = \mathrm{Var}\big(\phi_\tau(W)\big)$. The exact expressions of $\beta_{1,d}(\tau)$, $d = 0, 1$, are presented in the Appendix.

The above provides a basis for two standard approaches for inference: I) undersmoothing, so that $\gamma = 0$, and the leading bias vanishes; (ii) robust bias correction (RBC), where one subtracts off the estimated leading bias from the original estimator $\hat{\delta}(\tau)$, i.e. the RBC estimator is $\hat{\delta}^{rbc}(\tau) := \widehat{\delta}(\tau) - \hat{\gamma}\hat{\mu}(\tau)/\sqrt{nh}$, where $\hat{\gamma} = \sqrt{nh}h^2$. The corresponding variance then needs to take into account the added variability of the bias correction. In the following we separately discuss practical approaches of CI construction under undersmoothing and RBC.

Assuming $\gamma = 0$, i.e., undersmoothing, one can have the following plug-in estimator for $\widehat{\sigma}^2(\tau)$, which uses empirical influence functions at $\hat{q}_{1,d,\tau}$ and consistent density estimates $\hat{f}_{1,d}(\hat{q}_{1,d,\tau})$:

$$\widehat{\phi}_\tau(W_i) := -\sqrt{\frac{n}{n_1}}\frac{\widehat{\psi}_{1,1,\widehat{q}_{1,1,\tau}}(W_i)}{\widehat{f}_{1,1}(\widehat{q}_{1,1,\tau})} + \frac{\widehat{\psi}_{1,0,\widehat{q}_{1,0,\tau}}(W_i)}{\widehat{f}_{1,0}(\widehat{q}_{1,0,\tau})}, \tag{32}$$

where $\sqrt{n/n_1}$ is the plug-in counterpart of $1/\sqrt{p_1}$ in the population influence function ϕ_τ.

$$\widehat{\sigma}^2(\tau) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\phi}_\tau(W_i) - \overline{\widehat{\phi}}_\tau\right)^2, \qquad \overline{\widehat{\phi}}_\tau = \frac{1}{n}\sum_{i=1}^{n}\widehat{\phi}_\tau(W_i). \tag{33}$$

For example, $f_{1,d}(y)$, $d \in \{0,1\}$, can be estimated by numerically differentiating a y-smoothed version of the CDF, $\widehat{F}^{\text{sm}}_{1,d}(y)$, obtained from the corresponding estimators of $\widehat{F}_{1,d}(y)$ after replacing $\mathbf{1}\{Y_i \le y\}$ with $\Phi\big((y - Y_i)/h_Y\big)$, where Φ is a smooth CDF (e.g., the standard normal).[2] The pointwise $(1-\alpha)$ Wald CI is

$$\hat{\delta}(\tau) \pm z_{1-\alpha/2}\,\widehat{\sigma}(\tau)/\sqrt{nh},$$

where $z_{1-\alpha/2}$ is the $100 \cdot (1-\alpha/2) - th$ percentile of a standard normal distribution.

In practice, we recommend bootstrap, as it provides a better finite sample approximation of the distribution of $\hat{\delta}(\tau)$, particularly at tails. Bootstrap by resampling with replacement is valid under the usual regularity conditions (in particular, positive densities at the target quantiles). Alternatively, for the case of undersmoothing, one may leverage the influence function in (32) and implement influence function based multiplier bootstrap as follows.

1. Compute unit-level influence functions $\widehat{\phi}_\tau(W_i)$ by eq. (32).
2. Draw i.i.d. multipliers $\{\xi_i\}_{i=1}^n$ with $\mathbb{E}[\xi_i] = 0$ and $Var(\xi_i) = 1$ (e.g., Rademacher or normal $N(0,1)$).
3. For each draw, $b = 1, \ldots, B$, compute the bootstrap score

$$T_b(\tau) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i^{(b)}\big(\widehat{\phi}_\tau(W_i) - \overline{\widehat{\phi}}_\tau\big).$$

[2] An equivalent route is to *differentiate inside* the local polynomial, i.e., estimate $f_{1,1}(y)$ by an LLR of the y-smoothed score $\frac{1}{h_Y}\phi\big((y - Y_i)/h_Y\big)$ on R_i using $\{T_i = 1, R_i \ge 0\}$ (and analogously for the left side), where ϕ is a smooth pdf. For $f_{1,0}(y)$, one may (i) apply this to each term in the DR decomposition $F_{1,0}(y) = F_{0,0}(y) + \Delta_1(y) + \Delta_2^+(y) - \Delta_2^-(y)$, so that $f_{1,0}(y) = f_{0,0}(y) + f_{\Delta_1}(y) + f_{\Delta_2^+}(y) - f_{\Delta_2^-}(y)$; or (ii) form $\widehat{F}^{\text{sm}}_{1,0}(y)$ first and then differentiate it as above. Because the RDD local linear fit is linear in the outcome and its weights do not depend on y, differentiation in y commutes with the R-fit; hence these approaches are asymptotically equivalent. As in the main text, take $h_Y \to 0$ with $h_Y/h \to 0$.

Because $\mathrm{Var}\big(\phi_\tau(W_i)\big) = O(1/h)$, the distribution of $T_b(\tau)$ approximates that of $\frac{1}{\sqrt{n}}\sum_i \phi_\tau(W_i)$, which, after dividing by $\sqrt{nh}$ in step 4, yields a valid approximation to the law of $\sqrt{nh}\big(\hat{\delta}(\tau) - \delta(\tau)\big)$.

4. Construct the $(1-\alpha)$ CI as $\left[\hat{\delta}(\tau) - q^*_{1-\alpha/2}(\tau)/\sqrt{nh}, \hat{\delta}(\tau) - q^*_{\alpha/2}(\tau)/\sqrt{nh}\right]$, where $q^*_u(\tau)$ is the u quantile of $\{T_b(\tau)\}_{b=1}^B$.

Remark 4 (Clustered sampling) *For notational simplicity, the asymptotic results above are stated for i.i.d. observations. In applications with repeated observations for the same unit over time, let the data be indexed by clusters $g = 1, \ldots, G$ and within-cluster observations $j = 1, \ldots, m_g$, with $W_{gj} = (Y_{gj}, R_{gj}, X_{gj}, T_{gj})$ and $n = \sum_{g=1}^G m_g$. Suppose clusters are independent, while observations within a cluster may be arbitrarily dependent, and assume cluster sizes are uniformly bounded, i.e.,*

$$\max_{1\le g\le G} m_g \le \bar{m} < \infty.$$

Then $n \asymp G$, so the rate conditions in Assumption A continue to govern the first-order behavior up to constants, and the asymptotic linear representation in (31) carries over after grouping the influence-function contributions at the cluster level.

In particular, for the multiplier bootstrap under undersmoothing, define

$$\widehat{\Phi}_{g,\tau} \coloneqq \sum_{j=1}^{m_g}\big(\widehat{\phi}_\tau(W_{gj}) - \overline{\widehat{\phi}}_\tau\big), \qquad \overline{\widehat{\phi}}_\tau \coloneqq \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{m_g}\widehat{\phi}_\tau(W_{gj}),$$

draw i.i.d. cluster-level multipliers $\{\xi_g^{(b)}\}_{g=1}^G$ with $\mathbb{E}[\xi_g^{(b)}] = 0$ and $Var(\xi_g^{(b)}) = 1$, and compute

$$T_b^{\mathrm{cl}}(\tau) \coloneqq \frac{1}{\sqrt{n}}\sum_{g=1}^{G}\xi_g^{(b)}\,\widehat{\Phi}_{g,\tau}.$$

Likewise, a clustered plug-in variance estimator can be formed as

$$\widehat{\sigma}^2_{\mathrm{cl}}(\tau) \coloneqq \frac{1}{n}\sum_{g=1}^{G}\widehat{\Phi}^2_{g,\tau}.$$

Accordingly, whenever the same unit is observed in multiple periods, inference should be clustered at the unit level.

For a fixed finite grid of quantiles, the same influence-function-based multiplier bootstrap (using cluster-level multipliers when appropriate) can also be used to construct simultaneous confidence bands by using the bootstrap distribution of the maximum absolute t-statistic over the grid. We use this as a supplementary empirical diagnostic in the application. Formal uniform validity over a continuum of quantiles would require additional empirical-process arguments and is beyond the scope of the present paper.

The analytical and multiplier-bootstrap CIs above rely on undersmoothing, which may be undesirable when one uses a large bandwidth, such as an asymptotic mean squared error (AMSE)-optimal bandwidth. In this case, one may estimate the RBC estimator $\widehat{\delta}^{rbc}(\tau)$ by replacing all the boundary or interior LLR estimators involved in estimating $\hat{F}_{1,1}(y)$ and $\hat{F}_{1,0}(y)$ by their RBC alternatives before inverting to obtain $\widehat{\delta}^{rbc}(\tau)$. One can further construct CIs by the usual resampling bootstrap; when observations repeat across periods, resampling should be done at the cluster level. The RBC estimators for local polynomial regression at both boundary and interior points can be conveniently implemented in Stata by *lprobust* or in R by *nprobust* (Calonico et al., 2019).

4.3 Bandwidth Selection

Estimation of the conditional distribution functions in our framework requires choosing bandwidths for local polynomial fits on either side of the cutoff in each period. While most data-driven selectors in the RDD literature – such as Imbens and Kalyanaraman (2012) or Calonico et al. (2014) – are designed for mean treatment effects, our target is the CDF (and its inverse). We therefore follow the approach of FFM, who adapt mean-optimal bandwidths to quantile estimation by first selecting a reference bandwidth for the mean on each side of the cutoff and then scaling it to each quantile τ via a closed-form factor. Specifically, letting $h^{t,s}_{mean}$ denote the plug-in bandwidth for the mean in period $t \in \{0,1\}$ and side $s \in \{+,-\}$ the quantile specific bandwidth is

$$h^{t,s}_{\tau} = h^{t,s}_{mean} \left[\frac{\tau(1-\tau)}{\phi\left(\Phi^{-1}(\tau)\right)^2} \right]^{1/5}, \tag{34}$$

where ϕ and Φ are the standard normal PDF and CDF, respectively.[3]

This rule is valid in our setting because the local CDF estimators we use have the same boundary bias–variance trade-off as in FFM's RDD-QTE framework: they are local-linear fits of binary indicators $1(Y \leq y)$ on each side of the cutoff, and the doubly robust correction terms are likewise functions of such boundary fits. The derivation of the scaling factor in FFM does not depend on whether the design is a standard RDD or our difference-in-discontinuities extension; it only requires the usual smoothness and kernel regularity conditions already imposed in our identification arguments.

In practice, one may choose symmetric bandwidths ($h^{t,+}_{\tau} = h^{t,-}_{\tau}$) within a period if curvature and variance appear similar on both sides, or allow side-specific choices if smoothness or effective sample sizes differ—both options are accommodated by the FFM procedure. Following their recommendation, we use plug-in (mean-optimal) $h^{t,s}_{mean}$ as the reference rather than cross-validation, given their better performance at boundaries.

[3] For $\tau \in [0.05, 0.95]$, the scaling factor $\left[\frac{\tau(1-\tau)}{\phi\left(\Phi^{-1}(\tau)\right)^2}\right]^{1/5}$ is between 1.095 and 1.345.

For the LLR interior estimator of $\Delta_1(y)$, use a center bandwidth

$$h_\tau^1 = h_{mean}^1 \left[\frac{\tau(1-\tau)}{\phi\left(\Phi^{-1}(\tau)\right)^2} \right]^{1/5}, \tag{35}$$

where h_{mean}^1 can be the harmonic mean of $h_{mean}^{1,+}$ and $h_{mean}^{1,-}$ – any consistent choice of the same rate works.

The CDF estimation uses binary outcomes $I(Y \leq y)$, while the quantile level τ associated with y is unknown a priori. We recommend a plug-in rule that maps $y \mapsto \hat{\tau}(y)$ via a pilot CDF and then applies the FFM scaling shown above. We provide details in the Appendix.

5 Monte Carlo Simulations

This section reports a small-scale Monte Carlo study that evaluates the finite-sample performance of the doubly robust (DR) estimator developed in Sections 3–4. The goals are threefold: (i) verify that double robustness holds in finite samples, (ii) compare the DR estimator with single-robust alternatives (OR-only and IPW-only), and (iii) document how bias and dispersion evolve with sample size.

5.1 Simulation Design

Data-generating process. For $i = 1, \ldots, n$ we generate (X_i, T_i, R_i, Y_i) as follows. Draw independently two covariates:

$$X_1 \sim N(0,\, 0.7), \qquad X_2 \sim \text{Bernoulli}(1/2).$$

The time-period indicator is

$$T \mid X \sim \text{Bernoulli}\big(p(X)\big), \qquad p(X) := \Lambda(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2),$$

where $\Lambda(u) = \{1 + \exp(-u)\}^{-1}$ is the logistic CDF. Because $p(X)$ depends on X, the covariate distribution differs across $T = 0$ and $T = 1$.

The running variable is

$$R = 2U - 1, \qquad U \sim \text{Beta}(5, 5),$$

drawn independently of (X, T). Define the confounding policy and the treatment of interest:

$$C = \mathbf{1}\{R \geq 0\}, \qquad D = \mathbf{1}\{R \geq 0\} \cdot \mathbf{1}\{T = 1\}.$$

The baseline outcome index is

$$m(R, T, X) := \alpha_0 + \alpha_1 T + \rho_1 R + \rho_2 R^2 + \beta_1 X_1 + \beta_2 X_2.$$

The confounding-policy effect is heterogeneous in X_1:

$$\Delta_C(X) := \kappa_0 + \kappa_1 X_1,$$

and the treatment of interest has both a location and a scale component:

$$\Delta_D(X) := \delta_0 + \delta_1 X_2.$$

Let $\varepsilon \sim N(0,1)$ be drawn independently of (R, T, X). The potential outcomes are

$$\begin{aligned} Y(0,0) &:= m(R,T,X) + \varepsilon, \\ Y(1,0) &:= m(R,T,X) + \Delta_C(X) + \varepsilon, \\ Y(1,1) &:= m(R,T,X) + \Delta_C(X) + \Delta_D(X) + \sigma_D\,\varepsilon, \end{aligned}$$

with $\sigma_D := 1 + \delta_2 > 0$. The observed outcome is therefore

$$Y = m(R,T,X) + C \cdot \Delta_C(X) + D \cdot \Delta_D(X) + (\sigma_D - 1) \cdot D\,\varepsilon.$$

Parameter values. Table 1 collects all scalar DGP parameters.

Table 1: DGP parameter values

Group	Parameter	Value
Propensity score	$(\gamma_0,\ \gamma_1,\ \gamma_2)$	$(0,\ 0.7,\ -0.4)$
Baseline outcome	$(\alpha_0,\ \alpha_1)$	$(0,\ 0)$
	$(\rho_1,\ \rho_2)$	$(0.6,\ 0.3)$
	$(\beta_1,\ \beta_2)$	$(0.5,\ -0.3)$
Confounding effect	$(\kappa_0,\ \kappa_1)$	$(0.6,\ 0.8)$
Treatment effect	$(\delta_0,\ \delta_1,\ \delta_2)$	$(0.5,\ 0.1,\ -0.4)$
	$\sigma_D = 1 + \delta_2$	0.6

This DGP has three noteworthy features. First, conditional distributional stability holds: $Y(1,0) - Y(0,0) = \Delta_C(X)$ does not depend on T, satisfying our identifying assumption conditional on X. Second, unconditional stability can fail, because $p(X)$ and $\Delta_C(X)$ both depend on X_1, so the marginal distributional effect of C may differ across $T = 0$ and $T = 1$—motivating covariate adjustment. Third, quantile treatment effects are non-constant: because $\sigma_D \neq 1$, the QTE $\delta(\tau)$ varies with τ.

True quantile treatment effects. Because the estimand is defined at the limit $R \to 0$, we compute the "true" $\delta(\tau)$ by generating a large auxiliary sample ($n = 100{,}000$) at $(R = 0,\ T = 1)$ under $D = 1$ and under $D = 0$ (holding the same $X \mid T = 1$ distribution) and taking sample quantiles. At the five quantile indices used in the simulation, the true QTEs are:

$$\begin{gathered} \delta(0.10) = 0.892, \quad \delta(0.25) = 0.728, \quad \delta(0.50) = 0.542, \\ \delta(0.75) = 0.370, \quad \delta(0.90) = 0.192. \end{gathered}$$

The treatment effect declines monotonically across quantiles, reflecting the scale compression ($\sigma_D = 0.6 < 1$).

Estimators and specification cases. We evaluate the DR estimator of Section 4 under four specification cases, each combining a propensity-score (PS) model and an outcome-regression (OR) model:

Case 1. Both correct: OR includes (X_1, X_2, R, R^2); PS is logit of T on (X_1, X_2).

Case 2. OR wrong, PS correct: OR omits R^2 and X_1; PS is correctly specified as in Case 1.

Case 3. OR correct, PS wrong: OR is correctly specified as in Case 1; PS omits X_1 and X_2.

Case 4. Both wrong: OR omits R^2 and X_1; PS omits X_1 and X_2.

Under double robustness, Cases 1–3 yield consistent estimation of $\delta(\tau)$ because at least one nuisance component is correct; Case 4, where both are wrong, should exhibit persistent bias.

In addition, we report two single-robust benchmarks: 1) *OR-only*: the DR estimator with propensity-score weights set to unity ($\hat{w}(X_i) \equiv 1$), relying solely on the outcome regression; evaluated under correct and wrong OR specifications. 2) *IPW-only*: the DR estimator with $\hat{F}_0^{\pm} \equiv 0$, relying solely on inverse-probability weighting; evaluated under correct and wrong PS specifications.

Design choices. Sample sizes are $n \in \{1000, 2000, 5000, 10{,}000\}$, with $n_{\text{mc}} = 300$ Monte Carlo replications per design point. Bandwidths are $h = c \cdot n^{-1/5}$, with tuning constants $c_{\text{bw}} = 3.27$ for boundary CDF estimation and $c_\pi = 3.82$ for the propensity-score ratio $\bar{\pi}$, capped at $h_{\max} = 0.80$ and 0.90, respectively. CDFs are evaluated on a grid of $n_{\text{grid}} = 200$ equally spaced y-values, and quantiles are obtained by inverting the monotonized CDF estimates. Across Monte Carlo replications we report bias, root mean squared error (RMSE), and standard deviation (SD) for each $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$.[4]

5.2 Results

For brevity, we present the simulation results at sample size $n = 10{,}000$ in the main text. Table 2 reports bias, RMSE, and standard deviation across 300 Monte Carlo replications for each estimator and specification case. Tables A1–A3 in the Appendix report analogous results for $n \in \{1,000, 2,000, 5,000\}$, showing that standard deviations decline with sample size, while bias remains modest in the DR-valid cases. Several findings emerge.

1. Double robustness in finite samples. The DR estimator in Cases 1–3—where at least one of the OR or PS models is correctly specified—exhibits small bias across all reported quantile indices and sample sizes. At the largest sample

[4] Results using local quadratic estimation are qualitatively similar but exhibit higher RMSE due to increased variance; they are available upon request.

Table 2: Monte Carlo results: $n = 10{,}000$, $n_{\text{mc}} = 300$. True QTEs: $\delta(0.10) = 0.892$, $\delta(0.25) = 0.728$, $\delta(0.50) = 0.542$, $\delta(0.75) = 0.370$, $\delta(0.90) = 0.192$.

	$\tau =$ 0.10	$\tau =$ 0.25	$\tau =$ 0.50	$\tau =$ 0.75	$\tau =$ 0.90
Panel A: Bias					
DR (OR ✓, PS ✓)	0.002	−0.001	0.013	0.000	0.007
DR (OR ×, PS ✓)	0.001	−0.002	0.018	0.007	0.017
DR (OR ✓, PS ×)	0.000	−0.007	0.014	0.003	0.011
DR (OR ×, PS ×)	0.171	0.175	0.245	0.313	0.386
OR-only (✓)	−0.001	−0.001	0.008	−0.004	0.004
OR-only (×)	−0.202	0.016	0.268	0.499	0.730
IPW-only (✓)	−0.001	−0.005	0.012	−0.009	−0.036
IPW-only (×)	0.151	0.166	0.250	0.323	0.395
Panel B: RMSE					
DR (OR ✓, PS ✓)	0.210	0.148	0.117	0.117	0.152
DR (OR ×, PS ✓)	0.212	0.154	0.126	0.130	0.173
DR (OR ✓, PS ×)	0.210	0.142	0.111	0.105	0.122
DR (OR ×, PS ×)	0.266	0.225	0.267	0.331	0.410
OR-only (✓)	0.182	0.134	0.101	0.087	0.103
OR-only (×)	0.255	0.107	0.281	0.506	0.738
IPW-only (✓)	0.212	0.152	0.124	0.151	0.287
IPW-only (×)	0.254	0.218	0.272	0.340	0.419
Panel C: Standard Deviation					
DR (OR ✓, PS ✓)	0.210	0.148	0.117	0.117	0.152
DR (OR ×, PS ✓)	0.213	0.154	0.125	0.130	0.172
DR (OR ✓, PS ×)	0.210	0.142	0.110	0.105	0.122
DR (OR ×, PS ×)	0.204	0.141	0.107	0.107	0.139
OR-only (✓)	0.182	0.135	0.101	0.087	0.103
OR-only (×)	0.157	0.106	0.084	0.084	0.103
IPW-only (✓)	0.213	0.152	0.123	0.151	0.285
IPW-only (×)	0.205	0.141	0.108	0.106	0.139

Notes: Bold labels indicate DR-valid cases (at least one nuisance component correctly specified). ✓ = correctly specified; × = misspecified. Bias = $n_{\text{mc}}^{-1} \sum_m [\hat{\delta}^{(m)}(\tau) - \delta(\tau)]$; RMSE = $\{n_{\text{mc}}^{-1} \sum_m [\hat{\delta}^{(m)}(\tau) - \delta(\tau)]^2\}^{1/2}$; SD is the Monte Carlo standard deviation of $\hat{\delta}^{(m)}(\tau)$.

size, $n = 10{,}000$, the absolute bias of Cases 1–3 is uniformly below 0.025 at every reported quantile $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, consistent with the double-robustness property that the estimator is consistent whenever either nuisance component is correctly specified.

2. Inconsistency when both models are wrong. Case 4, where both the OR and PS are misspecified, displays large and persistent bias. At $n = 10{,}000$, the bias ranges from 0.171 at $\tau = 0.10$ to 0.386 at $\tau = 0.90$, and these magnitudes barely shrink from $n = 5{,}000$ to $n = 10{,}000$—consistent with asymptotic bias. The bias is economically meaningful: relative to the true QTE, it exceeds 100% at $\tau = 0.90$ (bias = 0.386 vs. $\delta(0.90) = 0.192$).

3. Convergence and variance reduction. For the DR-valid cases, the Monte Carlo standard deviation shrinks monotonically with n. For example, SD for Case 1 at the median falls from 0.337 ($n = 1{,}000$) to 0.117 ($n = 10{,}000$), roughly proportional to $n^{-2/5}$ as expected for local polynomial estimators at an interior point of the running-variable density. Meanwhile, bias remains small and stable, so RMSE declines at the same rate.

4. DR versus OR-only. The correctly specified OR-only estimator has small bias across all quantiles and sample sizes. At $n = 10{,}000$, its absolute bias does not exceed 0.008 at any reported quantile index. Moreover, OR-only achieves lower RMSE than the DR estimator at most quantiles—for example, 0.087 versus 0.117 at $\tau = 0.75$ and 0.103 versus 0.152 at $\tau = 0.90$ ($n = 10{,}000$)—reflecting the efficiency gain from imposing the parametric structure without the additional noise introduced by the IPW correction. However, the OR-only estimator lacks protection against misspecification: when the outcome model is wrong, its bias is large and persistent (e.g., 0.499 at $\tau = 0.75$), whereas the DR estimator with a correctly specified PS (Case 2) remains consistent, with absolute bias below 0.020 across all reported quantiles. This illustrates a bias–efficiency trade-off: the OR-only estimator is more efficient when correct, but the DR estimator provides insurance against model misspecification.

5. DR versus IPW-only. The correctly specified IPW-only estimator has small bias at most quantiles but suffers from extremely high variance in the tails. The variance inflation is concentrated at the extreme upper tail: at $\tau = 0.75$, the IPW-only SD is within 30% of DR Case 1 (0.151 vs. 0.117), but at $\tau = 0.90$ it nearly doubles to 0.285 compared with 0.152 for DR Case 1. This excess variance arises because propensity-score reweighting amplifies noise when $p(X)$ is close to zero or one, and no outcome model is present to stabilize the estimates. The DR estimator combines the unbiasedness of IPW-only with the variance reduction from the OR component, yielding uniformly lower RMSE.

6. Efficiency–robustness trade-off. The three single- and doubly-robust estimators illustrate a clear efficiency–robustness trade-off. At $n = 10{,}000$, the DR Case 1 achieves RMSE values

$$(0.210,\ 0.148,\ 0.117,\ 0.117,\ 0.152) \quad \text{across} \quad \tau = (0.10,\ 0.25,\ 0.50,\ 0.75,\ 0.90),$$

while the OR-only correct estimator achieves

$$(0.182,\ 0.134,\ 0.101,\ 0.087,\ 0.103).$$

The OR-only estimator is more efficient when the outcome model is correctly specified, because it avoids the additional variance from propensity-score estimation. The IPW-only correct estimator, by contrast, has inflated RMSE in the tails (0.287 at $\tau = 0.90$), driven by high variance. The DR estimator occupies a middle ground: it is less efficient than OR-only when the outcome model is

correct, but unlike OR-only it remains consistent when the outcome model is misspecified. This robustness property is the key practical advantage of the DR approach, as the researcher typically cannot verify whether the outcome model is correctly specified.

In summary, the Monte Carlo evidence confirms the double robustness property of the proposed estimator in finite samples. When at least one nuisance component is correctly specified, the DR estimator delivers small bias and well-behaved convergence; indeed, bias^2 accounts for less than 5% of MSE at every quantile index, confirming that estimation error is variance-dominated. The OR-only estimator can be more efficient when correctly specified, but the DR estimator provides a safeguard against outcome-model misspecification while the OR correction stabilizes the IPW estimator in the tails—making the DR approach the most reliable choice when model correctness is uncertain.

6 Empirical Application

In this section, we revisit the quasi-experimental setting studied by Grembi et al. (2016) to estimate the distributional impacts of relaxing fiscal constraints on municipal deficits in Italy. Before 2001, all Italian municipalities were subject to the Domestic Stability Pact (DSP), which imposed binding numerical constraints on local fiscal discipline. In 2001, the central government exempted municipalities with fewer than 5,000 residents from the DSP. At the same 5,000-resident threshold, the mayor's wage also changes discontinuously. Following Grembi et al. (2016), we exploit the before/after policy change and the shared population threshold to isolate the causal effect of relaxing fiscal rules while differencing out the confounding wage discontinuity.

Our sample and variables follow Grembi et al. (2016). We focus on the period 1999–2004, defining 1999–2000 as the pre-policy period and 2001–2004 as the post-policy period. The overall sample size is 6,656, with 2,251 from the pre-period and 4,405 from the post-period. The running variable is municipal population measured in the most recent census available for each period (1991 census for the pre period; 2001 census for the post period), consistent with the institutional timing in Grembi et al. (2016). We use the municipal budget deficit as the primary outcome, following Grembi et al.'s (2016) argument that deficit is the relevant "policy variable"—transfers are not locally raised and interest payments reflect past borrowing, so mayors are less directly accountable for those components. Additional results for the DSP target variable (the fiscal gap) are reported in the Appendix. Following Grembi et al. (2016), we adopt the same covariate set as in their analysis, mainly for comparability with the benchmark application and to capture the main dimensions of heterogeneity emphasized there. The covariate set includes an indicator for a binding mayoral term limit, the number of parties in the city council, the percentage of young voters, the speed of public good provision, average taxable income, the mayor's years of schooling, and an indicator for northern municipalities.

We implement the DR estimator described in Section 4. Implementaion de-

Table 3: Counterfactual quantiles and quantile treatment effects of relaxing fiscal constraints on deficit

τ	$q_{1,1,\tau}$		$q_{1,0,\tau}$		QTE	
0.10	-17.5	(4.0)***	-30.1	(7.5)***	12.6	(8.9)
0.15	-11.2	(3.6)***	-24.9	(6.6)***	13.6	(8.0)*
0.20	-6.5	(3.3)**	-21.0	(6.7)***	14.5	(7.9)*
0.25	-2.4	(3.1)	-17.7	(6.8)***	15.3	(7.8)**
0.30	1.3	(3.1)	-15.0	(7.0)**	16.2	(8.0)**
0.35	4.7	(3.1)	-12.4	(7.4)*	17.1	(8.4)**
0.40	8.2	(3.3)**	-10.1	(7.8)	18.3	(8.9)**
0.45	11.7	(3.5)***	-7.9	(8.5)	19.5	(9.5)**
0.50	15.3	(3.6)***	-5.7	(9.2)	21.0	(10.2)**
0.55	19.0	(3.8)***	-3.5	(10.1)	22.5	(11.2)**
0.60	22.9	(3.8)***	-1.3	(11.1)	24.1	(12.2)**
0.65	26.9	(3.8)***	1.1	(12.5)	25.8	(13.5)*
0.70	31.0	(3.8)***	3.6	(14.5)	27.4	(15.4)*
0.75	35.5	(4.0)***	6.4	(17.6)	29.1	(18.4)
0.80	40.6	(4.2)***	9.7	(22.7)	30.9	(23.5)
0.85	46.9	(4.7)***	14.1	(35.2)	32.8	(36.0)

Notes: Covariates include an indicator for a binding mayoral term limit, the number of parties in the city council, the percentage of young voters, the speed of public good provision, average taxable income, the mayor's years of schooling, and an indicator for northern cities. Standard errors (in parentheses) are computed using a multiplier wild bootstrap with 9,999 draws and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

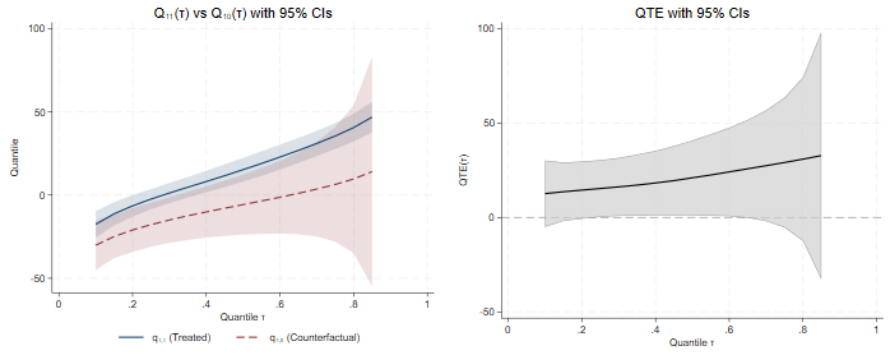

Figure 1: Counterfactual quantile curves and QTEs: deficit (full sample)

tails, including kernel, bandwidth choices and specifications for the parametric functions are provided in the Appendix. Although the theory is presented for repeated cross-sections, the empirical application uses a short panel of municipalities observed over multiple years. We therefore treat municipalities as independent clusters, allow arbitrary within-municipality dependence over time, and base inference on a municipality-clustered multiplier bootstrap described in Section 4.2, with $B = 9{,}999$ draws of i.i.d. Rademacher multipliers at the municipality level. When the unit of observation is municipality-year, each municipality contributes at most six observations over 1999–2004, so cluster sizes are uniformly bounded in this application. Reliable nonparametric inversion of the counterfactual CDF at extreme quantiles requires a sufficient number of observations in the upper tail of the local sample. We restrict reported quantiles to $\tau \in [0.10, 0.85]$ in our analyses, where the local sample is large enough for stable quantile inversion.[5]

Table 3 reports estimated treated quantiles $\widehat{q}_{1,1,\tau}$, counterfactual quantiles $\widehat{q}_{1,0,\tau}$, and the quantile treatment effects $\hat{\delta}(\tau) = \widehat{q}_{1,1,\tau} - \widehat{q}_{1,0,\tau}$ for the full sample. Figure 1 visualizes these quantile curves (left panel) and the QTE function (right panel) with pointwise 95% confidence intervals. Appendix Figure B2 reports simultaneous confidence bands over the reported quantile grid based on the clustered multiplier bootstrap. Although the point estimates display a clear pattern, these bands include zero at every reported quantile, so we do not reject the joint null that $\delta(\tau) = 0$ across the grid. Accordingly, the discussion below should be read as pointwise rather than uniform evidence that the QTE function differs from zero.

The estimates in Table 3 and Figure 1 indicate that exempting small municipalities from the DSP shifts the deficit distribution upward over most of the support. The estimated QTEs are positive over all the reported quantiles, and they are statistically significant at conventional levels from $\tau = 0.15$ through $\tau = 0.70$ (with the tightest precision in the middle of the distribution). The magnitudes are economically meaningful: at low quantiles the exemption reduces fiscal surpluses, and around the center of the distribution it moves municipalities from near balance into clear deficits. For example, at $\tau = 0.10$ the treated quantile is -17.5 while the counterfactual is -30.1, implying a QTE of 12.6; at the median ($\tau = 0.50$) the counterfactual is -5.7 while the treated quantile is 15.3, implying a 21.0 increase; and at $\tau = 0.65$ the treated quantile is 26.9 versus a counterfactual of 1.1 (QTE 25.8). Overall, the results suggest that relaxing the DSP mainly increases the incidence and size of *moderate* deficits, rather than shifting the entire distribution by a constant amount.

In the upper tail, point estimates remain sizeable (e.g., 29.1 at $\tau = 0.75$ and 32.8 at $\tau = 0.85$) but standard errors grow quickly, and none of these upper-tail QTEs is statistically distinguishable from zero. The widening confidence bands

[5] On the right side of the cutoff within the RD bandwidth, the overall effective local sample contains 414 observations in the main analysis (years 2001–2004). Due to the outcome sparsity in the upper tail, the estimated counterfactual quantile sequence $\widehat{q}_{1,0,\tau}$ display a big jump from $\tau = 0.85$ to $\tau = 0.90$, so we deem the estimated QTE at $\tau = 0.90$ unreliable for our sample.

in Figure 1 are consistent with greater sampling variability in counterfactual quantiles obtained by inversion of an estimated counterfactual CDF: when the counterfactual distribution is thin in the right tail, small CDF estimation errors translate into large quantile errors. Overall, the strongest pointwise evidence is for positive effects through the middle of the distribution, with less conclusive evidence about the extreme upper tail.

Table 4: Quantile Treatment Effects for Deficit: Subgroup Analysis

	Term limit		Young voters		Public good provision	
τ	Non-binding	Binding	Below median	Above median	Below median	Above median
0.10	5.4 (14.0)	58.7 (15.4)***	37.1 (12.9)***	-6.9 (10.1)	14.6 (12.6)	10.7 (21.7)
0.15	7.4 (14.5)	26.0 (17.8)	30.2 (10.9)***	-4.8 (9.0)	16.4 (10.6)	9.7 (20.9)
0.20	9.2 (14.9)	23.8 (13.6)*	29.4 (9.8)***	-2.7 (8.9)	17.4 (10.2)*	10.1 (16.2)
0.25	10.7 (15.9)	23.1 (12.8)*	29.2 (9.9)***	-1.2 (9.2)	18.7 (10.1)*	11.2 (14.8)
0.30	12.1 (16.9)	25.6 (15.0)*	29.3 (10.4)***	0.2 (9.4)	19.6 (10.6)*	12.4 (15.0)
0.35	13.8 (18.3)	26.5 (17.1)	29.5 (11.0)***	1.7 (10.1)	20.6 (11.1)*	13.6 (15.7)
0.40	15.5 (19.5)	27.3 (20.1)	29.8 (11.9)**	3.1 (10.8)	21.5 (11.5)*	15.0 (17.1)
0.45	17.5 (21.2)	27.5 (26.0)	30.5 (12.8)**	4.6 (11.9)	22.5 (12.2)*	16.4 (19.2)
0.50	19.6 (23.4)	26.5 (37.5)	31.1 (14.0)**	6.2 (13.3)	23.8 (12.9)*	17.7 (23.1)
0.55	21.6 (26.1)	11.9 (41.8)	32.1 (15.2)**	7.7 (15.0)	25.0 (13.6)*	18.7 (32.0)
0.60	23.4 (30.3)	7.0 (22.5)	33.1 (16.7)**	9.3 (16.7)	26.5 (14.4)*	16.7 (75.2)
0.65	24.8 (37.6)	6.4 (18.9)	34.4 (18.2)*	10.2 (18.2)	28.3 (15.3)*	1.2 (29.2)
0.70	25.9 (47.6)	6.8 (18.5)	36.2 (20.0)*	10.6 (19.1)	30.6 (16.1)*	0.7 (19.4)
0.75	26.6 (76.7)	7.4 (41.2)	38.3 (22.0)*	11.0 (20.4)	33.3 (16.7)**	1.0 (15.8)
0.80	-0.1 (28.8)	8.4 (40.9)	41.5 (24.3)*	11.4 (20.2)	37.3 (18.0)**	1.2 (14.9)
0.85	2.2 (26.8)	8.9 (36.1)	46.0 (27.4)*	12.8 (22.6)	43.8 (20.7)**	0.0 (18.1)

Note: QTEs for budget deficit. Covariates in Columns (1)–(2) include the number of parties in the city council, the percentage of young voters, the speed of public good provision, average taxable income, the mayor's years of schooling, and an indicator for whether the city is located in the North. Columns (3)–(6) additionally control for an indicator for a binding mayoral term limit. Standard errors (in parentheses) are computed using a multiplier wild bootstrap with 9,999 draws and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We next explore heterogeneity along the key political-economy dimensions highlighted by Grembi et al. (2016): term limits (reelection incentives), voter age composition, and the speed of public good provision. Table 4 reports subgroup QTEs. Figures 2–4 plot treated and counterfactual quantile curves by subgroup (the corresponding QTE plots are reported in Appendix Figures B3–B5).

Term limits. The heterogeneity by mayoral term limits is suggestive but not conclusive. For municipalities with non-binding mayoral term limits, the estimated QTEs are positive but fairly modest and imprecise throughout the distribution. For municipalities with binding mayoral term limits, the response is much stronger at the bottom of the distribution and then fades toward the upper quantiles. Taken together, the estimates suggest that when reelection incentives are absent, relaxing fiscal constraints mainly shifts municipalities in the lower part of the deficit distribution upward—eroding surpluses or pushing near-balance municipalities toward deficits—whereas with non-binding term limits the response is flatter and more diffuse across quantiles. At the same time,

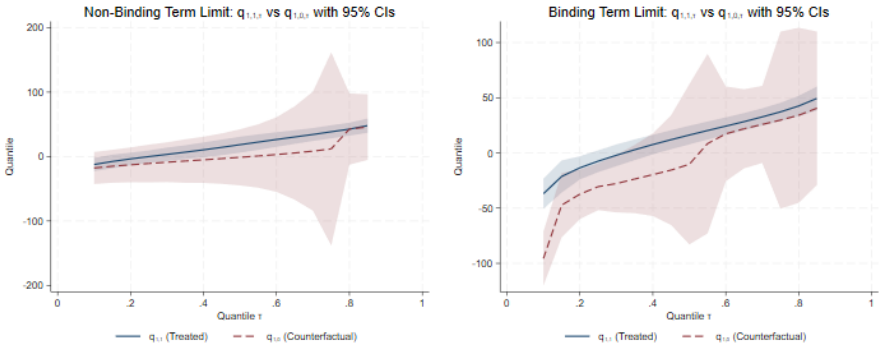

Figure 2: Counterfactual quantile curves for deficit: term limit non-binding (left) and binding (right)

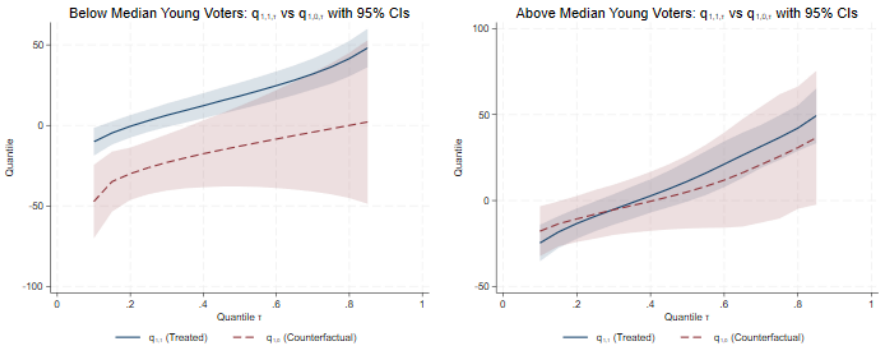

Figure 3: Counterfactual quantile curves for deficit: share of young voters below median (left) and above median (right)

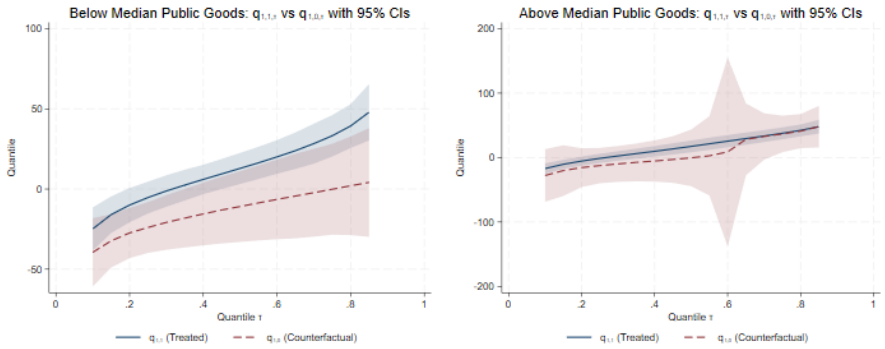

Figure 4: Counterfactual quantile curves for deficit: speed of public good provision below-median (left) and above-median (right)

the standard errors are wide at many quantiles, so these subgroup differences should be interpreted cautiously.

Voter age composition. The deficit response differs sharply by the share of young voters. For municipalities with a below-median share of young voters (older electorates), the QTEs are large, positive and statistically significant throughout the distribution. For municipalities with an above-median share of young voters, estimated QTEs are small—slightly negative at the bottom and modestly positive at higher quantiles—and they are not statistically significant. This pattern is consistent with the political-economy mechanism emphasized in Grembi et al. (2016): younger electorates may discipline fiscal policy and limit deficit expansion when formal constraints are relaxed, whereas older electorates exhibit a much stronger deficit response.

Public good provision. The speed of public good provision is defined as the ratio of provided to promised public goods in the provisional budget. A below-median value corresponds to slower provision and greater under-delivery relative to promises. For municipalities below the median speed, QTEs are positive across the distribution and become larger toward the upper tail; they are statistically significant from $\tau = 0.20$ onward, reaching 43.8 at $\tau = 0.85$. For above-median speed municipalities, estimated effects are much smaller and imprecisely estimated across quantiles. These results reinforce the interpretation in Grembi et al. (2016) that formal fiscal constraints bind more strongly where political distortions are more pronounced, while they matter less where governance performance is stronger.

Lastly, to probe the plausibility of the maintained identifying framework—including, in particular, Assumption 2G, the conditional stable distributional effect of the confounding policy—we conduct a falsification exercise using the two-year pre-policy sample (1999–2000). Specifically, we re-estimate the same covariate-adjusted specification as in the main analysis, treating 1999 as the base period and 2000 as a placebo post period, even though the fiscal-policy

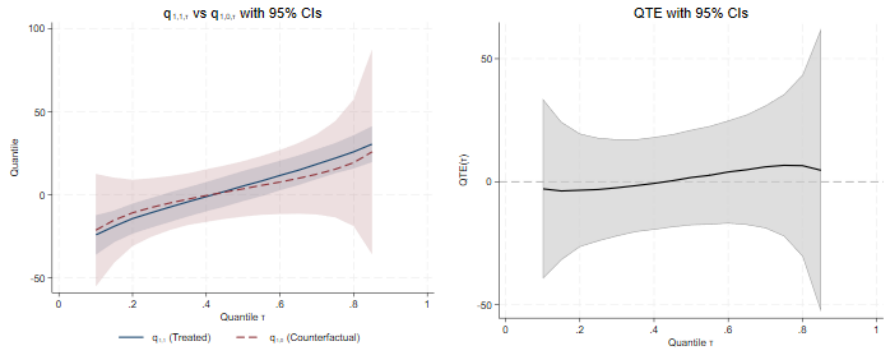

Figure 5: Counterfactual quantile curves and QTEs for deficit: falsification test

relaxation was not introduced until 2001. Figure 5 plots the resulting counterfactual quantile curves together with the placebo QTEs for deficit. The placebo point estimates remain close to zero, and the pointwise 95% confidence intervals contain zero at all reported quantiles, which is consistent with the maintained identifying framework in this empirical setting. As discussed above, however, this exercise should be interpreted as a plausibility check rather than as a formal validation of the assumptions.

Other standard RDD design diagnostics, including running-variable density and covariate balancing checks around the 5,000-resident threshold, are reported in Grembi et al. (2016), whose sample construction and variable definitions we follow, so we do not reproduce them here.

The mean diff-in-disc estimates in Grembi et al. (2016) establish that relaxing the DSP raises deficits on average and that the average response is larger in politically more distorted environments. The distributional results here refine that conclusion in several ways. First, in the full sample the deficit increase is concentrated in the lower and middle parts of the distribution: QTEs are positive and significant from $\tau = 0.15$ to $\tau = 0.70$ (Table 3 and Figure 1), indicating that the exemption mainly shifts municipalities from surplus or near-balance into moderate deficits. Second, subgroup comparisons (Table 4 and Figures 2–4) provide further evidence of strikingly different distributional shifts across the comparison groups. The clearest differences arise by electorate age and governance performance: municipalities with older electorates and those with slow public-good provision exhibit positive QTEs across most or all of the distribution, whereas municipalities with younger electorates and faster provision show little evidence of deficit expansion. Together with Grembi et al.'s (2016) evidence that adjustment operates largely through lower taxes, these patterns highlight that relaxing numerical rules can change not only average deficits but also which part of the distribution shifts, and that complementary monitoring/enforcement may be most valuable where political-economy frictions generate broad-based distributional increases in deficits.

7 Conclusion

This paper studies identification and estimation of quantile treatment effects (QTEs) in difference-in-discontinuities (diff-in-disc) designs when a new treatment is introduced in the post period and a confounding discontinuous policy is present in both periods. Our main identification result shows that QTEs are point identified under a conditional stable distributional effect assumption for the confounding treatment, stated directly in terms of outcome distributions at the cutoff. This assumption is the distributional diff-in-disc counterpart of the distributional parallel trends restriction in DiD. It delivers identification of the post-period counterfactual distribution and, through inversion, the corresponding QTEs at the threshold.

Our second contribution is methodological. We propose a doubly robust (DR) estimator and inference procedure for the identified QTEs. The estimator remains valid when either the parametric outcome regression or the propensity score model is correctly specified, and it avoids high-dimensional nonparametric adjustment for covariates in the local estimation environment of RDD and diff-in-disc designs. The framework is also invariant to strictly monotonic transformations of the outcome, making it applicable whether the outcome is measured in levels, logarithms, or another monotone scale. On the inference side, we establish asymptotic normality of the proposed estimator and show how influence-function-based bootstrap methods can be used to conduct inference in practice.

Our Monte Carlo results illustrate the practical value of the DR approach. The proposed estimator exhibits the expected double-robustness property, with small bias whenever at least one of the two working models is correctly specified, while the singly robust alternatives are sensitive to misspecification of their respective nuisance component. Overall, the simulations suggest that the DR estimator offers a favorable robustness-efficiency trade-off in finite samples.

In an empirical application revisiting Grembi et al. (2016), we apply the method to study the distributional effects of exempting small Italian municipalities from the Domestic Stability Pact. The estimated quantile effects suggest that the exemption increases deficits mainly in the lower and middle parts of the distribution, moving municipalities near fiscal balance into moderate deficits, while effects in the upper tail are considerably more uncertain. Subgroup analyses further suggest that the response may vary across municipal characteristics in ways that are not visible in mean effects. Although the simultaneous confidence bands indicate nontrivial uncertainty, the application illustrates the type of distributional heterogeneity that can be missed by average-effect analysis.

Taken together, the paper extends the diff-in-disc framework beyond mean effects and provides a new tool for studying heterogeneous policy impacts in settings with both discontinuity-based and before–after variation. More broadly, the results show that distributional diff-in-disc methods can uncover where in the outcome distribution policy responses occur, thereby complementing standard mean-based evaluations.

References

Arai, Y., Y.-C. Hsu, T. Kitagawa, I. Mourifié, and Y. Wan (2022). Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics 13*(1), 1–28.

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

Butts, K. (2023). Geographic difference-in-discontinuities. *Applied Economics Letters 30*(5), 615–619.

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

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. arXiv:1906.00198.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Chen, J. and J. Roth (2024). Logs with zeros? Some problems and solutions. *The Quarterly Journal of Economics 139*(2), 891–936.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*(1), 259–276.

Frandsen, B., M. Frölich, and B. Melly (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics 168*(2), 382–395.

Frölich, M. and M. Huber (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics 37*(4), 736–748.

Galindo-Silva, H., N. H. Some, and G. Tchuente (2021). Fuzzy difference-in-discontinuities: Identification theory and application to the Affordable Care Act. arXiv:1812.06537 [econ].

Grembi, V., T. Nannicini, and U. Troiano (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics 8*(3), 1–30.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies 79*(3), 933–959.

Kim, D. and J. M. Wooldridge (2025). Difference-in-differences estimator of quantile treatment effect on the treated. *Journal of Business & Economic Statistics 43*(2), 401–412.

Koenker, R. and G. Bassett, Jr. (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics 142*(2), 785–806.

Larsen, M. F. and J. Valant (2024). The long-term effects of grade retention: Evidence on persistence through high school and college. *Journal of Research on Educational Effectiveness 17*(4), 615–646.

Millán-Quijano, J. (2020). Fuzzy difference in discontinuities. *Applied Economics Letters 27*(19), 1552–1555.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*(2), 233–253.

Picchetti, P., C. C. X. Pinto, and S. T. Shinoki (2024). Difference-in-discontinuities: Estimation, inference and validity tests. arXiv:2405.18531 [econ.EM].

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

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

8 Appendix A Supplemental Monte Carlo results

Table A1: Monte Carlo results: $n = 1{,}000$, $n_{\text{mc}} = 300$. True QTEs: $\delta(0.10) = 0.892$, $\delta(0.25) = 0.728$, $\delta(0.50) = 0.542$, $\delta(0.75) = 0.370$, $\delta(0.90) = 0.192$.

	$\tau =$ 0.10	$\tau =$ 0.25	$\tau =$ 0.50	$\tau =$ 0.75	$\tau =$ 0.90
Panel A: Bias					
DR (OR ✓, PS ✓)	−0.011	0.047	0.057	0.055	0.059
DR (OR ×, PS ✓)	0.002	0.030	0.044	0.041	0.039
DR (OR ✓, PS ×)	0.013	0.057	0.069	0.062	0.062
DR (OR ×, PS ×)	0.201	0.231	0.319	0.356	0.434
OR-only (✓)	0.003	0.074	0.091	0.053	0.039
OR-only (×)	−0.143	0.082	0.326	0.538	0.757
IPW-only (✓)	0.000	0.043	0.044	−0.001	−0.319
IPW-only (×)	0.185	0.221	0.318	0.360	0.439
Panel B: RMSE					
DR (OR ✓, PS ✓)	0.493	0.403	0.342	0.359	0.424
DR (OR ×, PS ✓)	0.494	0.405	0.373	0.402	0.486
DR (OR ✓, PS ×)	0.506	0.399	0.325	0.322	0.342
DR (OR ×, PS ×)	0.549	0.448	0.451	0.490	0.574
OR-only (✓)	0.481	0.405	0.325	0.299	0.310
OR-only (×)	0.434	0.327	0.417	0.601	0.817
IPW-only (✓)	0.490	0.407	0.357	0.539	1.307
IPW-only (×)	0.540	0.439	0.447	0.492	0.576
Panel C: Standard Deviation					
DR (OR ✓, PS ✓)	0.494	0.401	0.337	0.356	0.421
DR (OR ×, PS ✓)	0.495	0.404	0.371	0.401	0.485
DR (OR ✓, PS ×)	0.507	0.395	0.318	0.316	0.337
DR (OR ×, PS ×)	0.512	0.385	0.320	0.338	0.376
OR-only (✓)	0.482	0.399	0.312	0.295	0.308
OR-only (×)	0.410	0.318	0.261	0.269	0.309
IPW-only (✓)	0.491	0.405	0.355	0.540	1.270
IPW-only (×)	0.508	0.380	0.314	0.336	0.373

Notes: Bold labels indicate DR-valid cases (at least one nuisance component correctly specified). ✓ = correctly specified; × = misspecified. Bias = $n_{\text{mc}}^{-1} \sum_m [\hat{\delta}^{(m)}(\tau) - \delta(\tau)]$; RMSE = $\{n_{\text{mc}}^{-1} \sum_m [\hat{\delta}^{(m)}(\tau) - \delta(\tau)]^2\}^{1/2}$; SD is the Monte Carlo standard deviation of $\hat{\delta}^{(m)}(\tau)$.

Table A2: Monte Carlo results: $n = 2{,}000$, $n_{\text{mc}} = 300$. True QTEs: $\delta(0.10) = 0.892$, $\delta(0.25) = 0.728$, $\delta(0.50) = 0.542$, $\delta(0.75) = 0.370$, $\delta(0.90) = 0.192$.

	$\tau =$ 0.10	$\tau =$ 0.25	$\tau =$ 0.50	$\tau =$ 0.75	$\tau =$ 0.90
Panel A: Bias					
DR (OR ✓, PS ✓)	0.060	0.037	0.028	−0.010	−0.006
DR (OR ×, PS ✓)	0.043	0.028	0.012	−0.015	−0.004
DR (OR ✓, PS ×)	0.063	0.038	0.033	0.004	−0.009
DR (OR ×, PS ×)	0.216	0.222	0.260	0.304	0.353
OR-only (✓)	0.038	0.055	0.043	0.003	−0.019
OR-only (×)	−0.152	0.048	0.279	0.496	0.700
IPW-only (✓)	0.053	0.034	0.027	−0.004	−0.153
IPW-only (×)	0.195	0.206	0.261	0.313	0.360
Panel B: RMSE					
DR (OR ✓, PS ✓)	0.392	0.312	0.247	0.238	0.306
DR (OR ×, PS ✓)	0.403	0.318	0.251	0.281	0.359
DR (OR ✓, PS ×)	0.397	0.299	0.236	0.207	0.234
DR (OR ×, PS ×)	0.443	0.374	0.351	0.372	0.437
OR-only (✓)	0.369	0.289	0.243	0.203	0.206
OR-only (×)	0.347	0.230	0.338	0.529	0.728
IPW-only (✓)	0.386	0.312	0.264	0.306	0.871
IPW-only (×)	0.432	0.363	0.349	0.380	0.443
Panel C: Standard Deviation					
DR (OR ✓, PS ✓)	0.388	0.310	0.246	0.238	0.306
DR (OR ×, PS ✓)	0.401	0.318	0.251	0.281	0.360
DR (OR ✓, PS ×)	0.392	0.297	0.234	0.207	0.235
DR (OR ×, PS ×)	0.387	0.302	0.236	0.215	0.258
OR-only (✓)	0.368	0.284	0.240	0.203	0.205
OR-only (×)	0.313	0.225	0.190	0.185	0.201
IPW-only (✓)	0.383	0.311	0.263	0.306	0.859
IPW-only (×)	0.387	0.299	0.232	0.215	0.259

Notes: See Table A1.

Table A3: Monte Carlo results: $n = 5{,}000$, $n_{\text{mc}} = 300$. True QTEs: $\delta(0.10) = 0.892$, $\delta(0.25) = 0.728$, $\delta(0.50) = 0.542$, $\delta(0.75) = 0.370$, $\delta(0.90) = 0.192$.

	$\tau =$ 0.10	$\tau =$ 0.25	$\tau =$ 0.50	$\tau =$ 0.75	$\tau =$ 0.90
Panel A: Bias					
DR (OR ✓, PS ✓)	0.011	−0.013	−0.012	−0.011	0.008
DR (OR ×, PS ✓)	0.013	−0.007	−0.003	0.004	0.012
DR (OR ✓, PS ×)	0.015	−0.014	−0.009	−0.006	0.007
DR (OR ×, PS ×)	0.192	0.181	0.229	0.308	0.391
OR-only (✓)	0.002	−0.006	−0.007	−0.015	−0.008
OR-only (×)	−0.183	0.018	0.261	0.496	0.733
IPW-only (✓)	0.009	−0.013	−0.016	−0.027	−0.100
IPW-only (×)	0.174	0.172	0.232	0.316	0.400
Panel B: RMSE					
DR (OR ✓, PS ✓)	0.267	0.205	0.166	0.168	0.194
DR (OR ×, PS ✓)	0.276	0.225	0.185	0.191	0.221
DR (OR ✓, PS ×)	0.269	0.199	0.151	0.146	0.165
DR (OR ×, PS ×)	0.332	0.271	0.279	0.345	0.431
OR-only (✓)	0.230	0.184	0.146	0.137	0.149
OR-only (×)	0.277	0.156	0.290	0.512	0.749
IPW-only (✓)	0.270	0.214	0.181	0.208	0.524
IPW-only (×)	0.323	0.262	0.280	0.353	0.440
Panel C: Standard Deviation					
DR (OR ✓, PS ✓)	0.267	0.205	0.166	0.168	0.194
DR (OR ×, PS ✓)	0.276	0.225	0.185	0.192	0.221
DR (OR ✓, PS ×)	0.269	0.199	0.151	0.146	0.165
DR (OR ×, PS ×)	0.272	0.203	0.159	0.156	0.182
OR-only (✓)	0.230	0.184	0.146	0.137	0.149
OR-only (×)	0.208	0.155	0.126	0.127	0.153
IPW-only (✓)	0.270	0.214	0.181	0.207	0.516
IPW-only (×)	0.272	0.199	0.157	0.156	0.184

Notes: See Table A1.

9 Appendix B Supplemental empirical discussion and results

9.1 Empirical implementation details

In the empirical analysis, all local linear regressions use a uniform (rectangular) kernel $K(u) = \frac{1}{2}\mathbf{1}\{|u| \leq 1\}$. The reference bandwidth for mean estimation, h_{mean}, is selected by the plug-in rule of Calonico et al. (2014) applied to $T = 1$, $R \geq 0$. We then undersmooth by setting $h_{base} = h_{mean} \cdot n^{-\eta}$ with $\eta = 0.055$; the CDF estimation bandwidth is $h = 1.095 \cdot h_{base}$, where 1.095 is the minimum of the FFM scaling factor in eq. (34) at $\tau = 0.50$. Density estimation for the sparsity function uses the full τ-specific scaling.

The propensity score $\hat{p}(X_i)$ is estimated by a global logit of T_i on R_i, R_i^2, and the full covariate vector, evaluated at $R_i = 0$. The outcome regression $\hat{F}_0^{\pm}(y, X_i)$ is estimated by OLS of the smoothed indicator $\Phi((y - Y_i)/h_Y)$ on R_i, R_i^2, and the covariates, separately for $T_i = 0$, $R_i \geq 0$ and $T_i = 0$, $R_i < 0$, and evaluated at $R_i = 0$. Both nuisance models are estimated on the full support of R on each side; in contrast, the LLR steps for $\hat{F}_{1,1}(y)$, $\hat{\Delta}_1(y)$, $\hat{\Delta}_2^{\pm}(y)$, and the sparsity estimates are all local to the cutoff, using observations within $|R_i| \leq h$. Following the discussion in Section 4.1, we replace the sharp indicator $\mathbf{1}\{Y_i \leq y\}$ with the smoothed version $\Phi((y - Y_i)/h_Y)$ throughout, using $h_Y = 1.06 \cdot \hat{\sigma}_Y \cdot n^{-1/5}$ (Silverman's rule of thumb).

9.2 Supplemental empirical results

Table B1 presents the results for fiscal gap. Figure B1 further visualizes the estimated counterfactual quantile curves (left) and QTEs (right). Figure B2. shows QTEs along with the 95% simultaneous CIs for deficit (left) and fiscal gap (right) for the full sample. Figures B3-B5 visualize the estimated counterfactual quantile curves of deficit for the subgroups defined by non-binding vs. binding term limit, below- vs. above-median share of young voters, and below- vs. above-median speed of public good provision, respectively.

Table B1: Counterfactual quantiles and quantile treatment effects of relaxing fiscal constraints on fiscal gap

τ	$q_{1,1,\tau}$		$q_{1,0,\tau}$		QTE	
0.10	75.7	(9.5)***	35.4	(18.1)**	40.2	(21.5)*
0.15	93.9	(12.4)***	44.0	(22.1)**	50.0	(26.8)*
0.20	106.6	(14.6)***	51.0	(27.3)*	55.6	(32.6)*
0.25	116.0	(15.2)***	59.0	(34.6)*	57.0	(39.2)
0.30	123.9	(14.0)***	68.9	(39.7)*	55.0	(43.1)
0.35	131.1	(14.2)***	87.8	(41.9)**	43.3	(45.3)
0.40	138.0	(15.0)***	107.8	(43.0)**	30.2	(46.7)
0.45	145.0	(16.3)***	118.1	(38.3)***	26.9	(43.1)
0.50	152.7	(18.6)***	125.9	(35.8)***	26.8	(42.3)
0.55	161.6	(22.2)***	133.0	(32.3)***	28.6	(42.2)
0.60	172.9	(23.4)***	140.0	(32.8)***	32.9	(43.5)
0.65	187.6	(23.5)***	147.4	(35.4)***	40.2	(45.6)
0.70	204.1	(22.8)***	155.8	(40.4)***	48.3	(49.1)
0.75	219.1	(21.6)***	166.0	(49.1)***	53.1	(55.8)
0.80	233.7	(20.0)***	179.2	(50.0)***	54.5	(55.8)
0.85	252.7	(18.0)***	195.4	(48.9)***	57.3	(53.7)

Notes: Covariates include an indicator for a binding mayoral term limit, the number of parties in the city council, the percentage of young voters, the speed of public good provision, average taxable income, the mayor's years of schooling, and an indicator for northern cities. Standard errors (in parentheses) are computed using a multiplier wild bootstrap with 9,999 draws and are clustered at the municipality level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

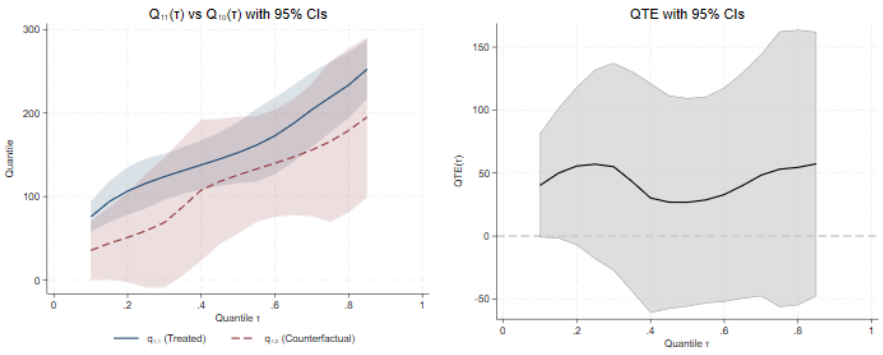

Figure B1: Counterfactual quantile curves and QTEs: fiscal gap (full sample)

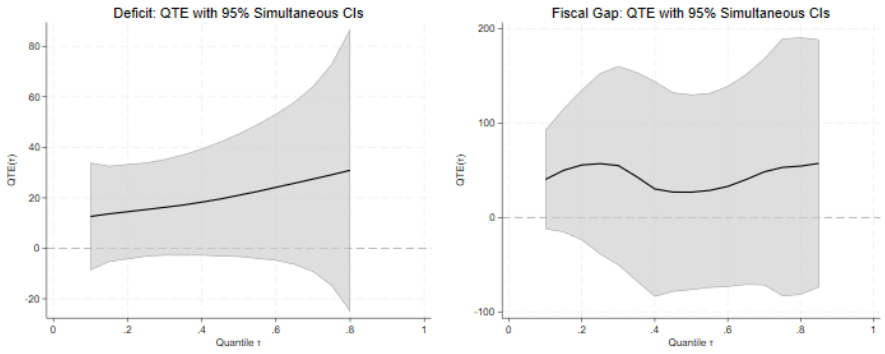

Figure B2: simultaneous CIs: deficit (left) and fiscal gap (right)

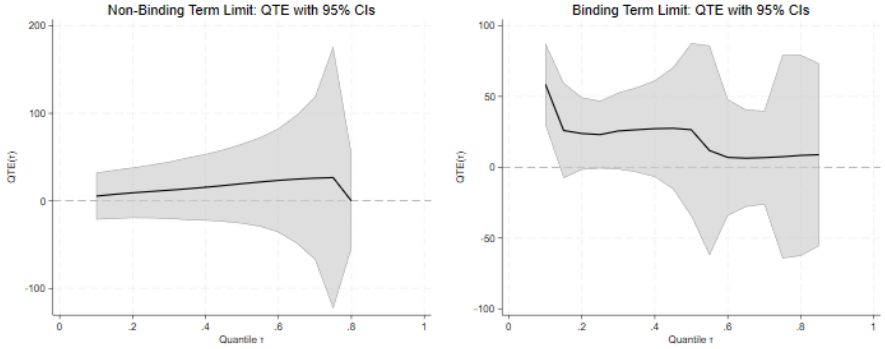

Figure B3: QTEs for deficit: term limit non-binding (left) and binding (right)

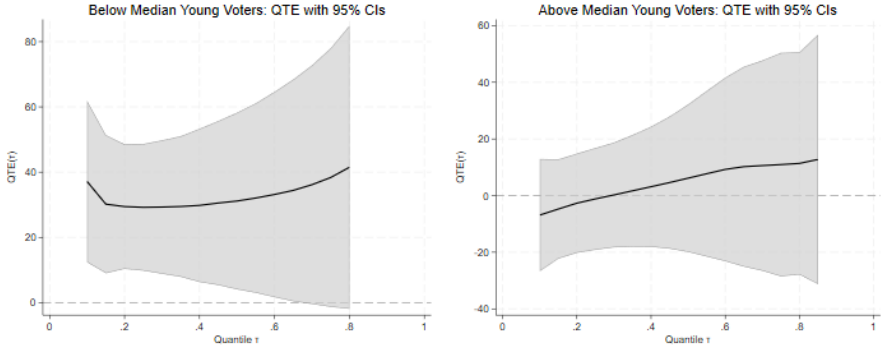

Figure B4: QTEs for deficit: share of young voters below median (left) and above median (right)

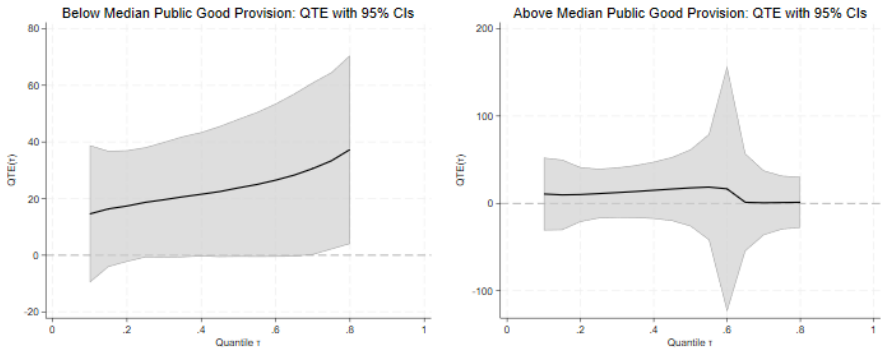

Figure B5: QTEs for deficit: speed of public good provision below median (left) and above median (right)

10 Appendix C Proofs

10.1 Proofs for identification

Proof of eq.s (15) and (16): Fix $y \in \mathcal{Y}$ and write $I_y{:=}\mathbf{1}\{Y \leq y\}$. For $c, d \in \{0,1\}$ define $I_{cd}(y){:=}\mathbf{1}\{Y(c,d) \leq y\}$.

For clarity, define one-sided limits of conditional expectations:

$$\mathbb{E}_t^+[W]{:=}\lim_{r\downarrow 0}\mathbb{E}[W \mid R=r, T=t], \qquad \mathbb{E}_t^-[W]{:=}\lim_{r\uparrow 0}\mathbb{E}[W \mid R=r, T=t].$$

Recall $p(x){:=}\Pr(T=1 \mid X=x, R=0)$ and $p{:=}\Pr(T=1 \mid R=0) = \mathbb{E}[p(X) \mid R=0]$. Further the odds $\pi(x){:=}\frac{p(x)}{1-p(x)}$ and the normalized weight

$$w(x){:=}\frac{\pi(x)}{\mathbb{E}[\pi(X) \mid R=0, T=0]} = \frac{(1-p)\,p(x)}{p\,(1-p(x))}.$$

To see the last equality,

$$\begin{aligned}
& \mathbb{E}\left[\pi\left(X\right)|R=0,T=0\right]\\
= & \mathbb{E}\left[\frac{p\left(X\right)}{1-p\left(X\right)}|R=0,T=0\right]\\
= & \frac{1}{1-p}\mathbb{E}\left[\frac{\left(1-T\right)p\left(X\right)}{1-p\left(X\right)}|R=0\right]\\
= & \frac{1}{1-p}\mathbb{E}\left[\mathbb{E}\left[\frac{\left(1-T\right)p\left(X\right)}{1-p\left(X\right)}|X,R=0\right]|R=0\right]\\
= & \frac{1}{1-p}\mathbb{E}\left[\frac{p\left(X\right)}{1-p\left(X\right)}\mathbb{E}\left[\left(1-T\right)|X,R=0\right]|R=0\right]\\
= & \frac{1}{1-p}\mathbb{E}\left[p\left(X\right)|R=0\right]\\
= & \frac{p}{1-p}.
\end{aligned}$$

Then eq. (15) states for any integrable function $G(X)$,

$$\mathbb{E}[G(X) \mid R=0,T=1] = \mathbb{E}[w(X)\,G(X) \mid R=0,T=0]. \tag{36}$$

To verify the above, note that

$$\begin{aligned}
\mathbb{E}[w(X)\,G(X) \mid R{=}0,T{=}0] &= \tfrac{1}{p}\,\mathbb{E}\big[p(X)\,G(X) \mid R{=}0\big]\\
&= \mathbb{E}[G(X) \mid R{=}0,T{=}1],
\end{aligned}$$

where the first equality uses Bayes' rule and the definition of $w(\cdot)$.

Now consider the two one-sided limits in eq. (16). Under the sharp rule $C = \mathbf{1}\{R \geq 0\}$ and because $D = 0$ when $T = 0$, we have: for $r \geq 0$ (right side), $Y = Y(1,0)$, hence $I_y = I_{10}(y)$; for $r < 0$ (left side), $Y = Y(0,0)$, hence $I_y = I_{00}(y)$. By the smoothness conditions in Assumption 1G,

$$\begin{aligned}
\mathbb{E}_0^+\big[w(X)I_y\big] &= \mathbb{E}\big[w(X)I_{10}(y) \mid R=0,T=0\big],\\
\mathbb{E}_0^-\big[w(X)I_y\big] &= \mathbb{E}\big[w(X)I_{00}(y) \mid R=0,T=0\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}_0^+[w(X)I_y] - \mathbb{E}_0^-[w(X)I_y] &= \mathbb{E}\big[w(X)\{I_{10}(y) - I_{00}(y)\} \mid R{=}0,T{=}0\big]\\
&= \mathbb{E}\big[w(X)\,\mathbb{E}[I_{10}(y) - I_{00}(y) \mid X,R{=}0,T{=}0]\\
&\qquad \big|\, R{=}0,T{=}0\big]\\
&= \mathbb{E}\big[\mathbb{E}[I_{10}(y) - I_{00}(y) \mid X,R{=}0,T{=}0]\\
&\qquad \big|\, R{=}0,T{=}1\big],
\end{aligned}$$

where the last step uses the reweighting identity (36) with $g(X) = \mathbb{E}[I_{10}(y) - I_{00}(y) \mid X,R{=}0,T{=}0]$.

By Assumption 2G (conditional stable distributional effect),

$$\mathbb{E}[I_{10}(y) - I_{00}(y) \mid X, R=0, T=0] = \mathbb{E}[I_{10}(y) - I_{00}(y) \mid X, R=0, T=1]\,.$$

Hence,

$$\begin{aligned}\mathbb{E}_0^+[w(X)I_y] - \mathbb{E}_0^-[w(X)I_y] &= \mathbb{E}\big[\mathbb{E}\big[I_{10}(y) - I_{00}(y) \mid X, R{=}0, T{=}1\big] \mid R{=}0, T{=}1\big] \\ &= \mathbb{E}[I_{10}(y) - I_{00}(y) \mid R=0, T=1] \\ &= F_{Y(1,0)|R=0,T=1}(y) - F_{Y(0,0)|R=0,T=1}(y),\end{aligned}$$

which is exactly eq.(16).

Proof of Theorem 1 (double robustness of $\Delta^{DR}(y)$): Fix $y \in \mathcal{Y}$ and write $I_y = \mathbf{1}\{Y \le y\}$. Use the one-sided expectation operators

$$\begin{aligned}\mathbb{E}_t^+[\cdot] &\coloneqq \lim_{r \downarrow 0} \mathbb{E}[\cdot \mid R = r, T = t], \\ \mathbb{E}_t^-[\cdot] &\coloneqq \lim_{r \uparrow 0} \mathbb{E}[\cdot \mid R = r, T = t].\end{aligned}$$

Let the true one-sided conditional CDFs in the $T = 0$ sample be

$$\begin{aligned}\mu_0^+(y, x) &\coloneqq \lim_{r \downarrow 0} \mathbb{E}[I_y \mid X = x, R = r, T = 0], \\ \mu_0^-(y, x) &\coloneqq \lim_{r \uparrow 0} \mathbb{E}[I_y \mid X = x, R = r, T = 0].\end{aligned}$$

Let $F_0^+(y, x)$ and $F_0^-(y, x)$ be (possibly misspecified) working models for $\mu_0^+(y, x)$ and $\mu_0^-(y, x)$. Let $w(X)$ denote the (normalized) IPW weight constructed from the true propensity score $p(X) = \Pr(T = 1 \mid X, R = 0)$ as in eq. (14), and $\widetilde{w}(X)$ the (possibly misspecified) working models for $w(X)$.

Proof: Define the population DR correction

$$\begin{aligned}\Delta^{DR}(y) \coloneqq & \underbrace{\mathbb{E}\big[F_0^+(y, X) - F_0^-(y, X) \mid R = 0, T = 1\big]}_{\Delta_1(y)} \\ & + \underbrace{\mathbb{E}_0^+\big[\widetilde{w}(X)\{I_y - F_0^+(y, X)\}\big]}_{\Delta_2^+(y)} \\ & - \underbrace{\mathbb{E}_0^-\big[\widetilde{w}(X)\{I_y - F_0^-(y, X)\}\big]}_{\Delta_2^-(y)}.\end{aligned}$$

We show that $\Delta^{DR}(y) = \Delta(y)$, where

$$\Delta(y) \coloneqq F_{Y(1,0)|R=0,T=1}(y) - F_{Y(0,0)|R=0,T=1}(y),$$

if either (i) $F_0^\pm = \mu_0^\pm$ or (ii) $\widetilde{w} = w$.

Case (i): outcome models correct. Suppose $F_0^+(y, x) = \mu_0^+(y, x)$ and $F_0^-(y, x) = \mu_0^-(y, x)$. Then, by iterated expectations and the definition of μ_0^+,

$$\Delta_2^+(y) = \mathbb{E}_0^+\big[\widetilde{w}(X)\{\mathbb{E}[I_y \mid X, R, T = 0] - \mu_0^+(y, X)\}\big] = 0,$$

and similarly $\Delta_2^-(y) = 0$ (under the same regularity conditions used to exchange limits and expectations). Therefore,

$$\Delta^{DR}(y) = \Delta_1(y) = \mathbb{E}\big[\mu_0^+(y,X) - \mu_0^-(y,X) \mid R=0, T=1\big].$$

Because the design is sharp in $C = \mathbf{1}\{R \geq 0\}$ when $T = 0$, $\mu_0^+(y,x) = F_{Y(1,0)|X,R=0,T=0}(y \mid x)$ and $\mu_0^-(y,x) = F_{Y(0,0)|X,R=0,T=0}(y \mid x)$. Thus $\mu_0^+(y,x) - \mu_0^-(y,x) = \Delta_0(y,x)$, and by Assumption 2G, $\Delta_0(y,x) = \Delta_1(y,x)$ for all (y,x). Hence $\Delta^{DR}(y) = \mathbb{E}[\Delta_1(y,X) \mid R=0, T=1] = \Delta(y)$.

Case (ii): weights correct. Now suppose $\widetilde{w}(\cdot)$ is correctly specified, while $F_0^{\pm}$ may be misspecified. Using the standard reweighting identity (proved in the proof of eq. (16)), for any integrable $G(X)$,

$$\mathbb{E}[G(X) \mid R=0, T=1] = \mathbb{E}[w(X)\, G(X) \mid R=0, T=0],$$

and by smoothness we may apply it to the one-sided limits as $r \to 0$. In particular,

$$\begin{aligned}
\mathbb{E}_0^+\big[w(X)F_0^+(y,X)\big] &= \mathbb{E}\big[F_0^+(y,X) \mid R=0, T=1\big], \\
\mathbb{E}_0^-\big[w(X)F_0^-(y,X)\big] &= \mathbb{E}\big[F_0^-(y,X) \mid R=0, T=1\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Delta^{DR}(y) &= \mathbb{E}\big[F_0^+(y,X) - F_0^-(y,X) \mid R=0, T=1\big] \\
&\quad + \mathbb{E}_0^+\big[w(X)I_y\big] - \mathbb{E}_0^+\big[w(X)F_0^+(y,X)\big] \\
&\quad - \Big(\mathbb{E}_0^-\big[w(X)I_y\big] - \mathbb{E}_0^-\big[w(X)F_0^-(y,X)\big]\Big) \\
&= \mathbb{E}_0^+[w(X)I_y] - \mathbb{E}_0^-[w(X)I_y].
\end{aligned}$$

By equation 16, the last line equals $F_{Y(1,0)|R=0,T=1}(y) - F_{Y(0,0)|R=0,T=1}(y) = \Delta(y)$. Hence $\Delta^{DR}(y) = \Delta(y)$ under correct weights.

Combining the two cases proves the claimed double robustness for $\Delta^{DR}(y)$, and therefore the identification statement in Theorem 1 follows.

10.2 Derivation of the influence functions

Remark 5 *We summarize the standard linearization for the local-linear (LL) intercept at a boundary (Frandsen et al., 2012, Sec. 4 & App. C) and adapt it to our DR components.*

Let $I_i(y) := 1\{Y_i \leq y\}$ and $K_h(u) := K(u/h)/h$. For $t \in \{0,1\}$ and $s \in \{+,-\}$ recall the one-sided moment functionals introduced in Section 4.2:

$$S_{m,s}^{(t)}(h) := \mathbb{E}\big[(R/h)^m K_h(R)\, 1\{T=t,\ sR \geq 0\}\big], \qquad m = 0,1,2,$$

together with

$$\begin{aligned}
D_s^{(t)}(h) &:= S_{0,s}^{(t)}(h)\, S_{2,s}^{(t)}(h) - \big(S_{1,s}^{(t)}(h)\big)^2, \\
\ell_s^{(t)}(r;h) &:= S_{2,s}^{(t)}(h) - S_{1,s}^{(t)}(h)\,(r/h).
\end{aligned}$$

When $t = 1$ we drop the superscript and write $S_{m,s}(h) := S^{(1)}_{m,s}(h)$, $D_s(h) := D^{(1)}_s(h)$ and $\ell_s(r;h) := \ell^{(1)}_s(r;h)$. The (population) effective local sample sizes on the right/left of the cutoff for $T = 1$ are

$$n_{+,h} \coloneqq S_{0,+}(h) = \mathbb{E}\big[K_h(R)\,1\{T = 1, R \geq 0\}\big],$$
$$n_{-,h} \coloneqq S_{0,-}(h) = \mathbb{E}\big[K_h(R)\,1\{T = 1, R < 0\}\big],$$

and, analogously for $T = 0$, we define

$$n_{0,+,h} := \mathbb{E}\big[K_h(R)\,1\{T = 0, R \geq 0\}\big], \qquad n_{0,-,h} := \mathbb{E}\big[K_h(R)\,1\{T = 0, R < 0\}\big].$$

(i) Right-limit CDF in $T = 1$. Consider the T=1, $R \geq 0$ local linear (LL) problem

$$(\hat{a}^+_0(y), \hat{a}^+_1(y)) = \arg\min_{a_0,a_1} \sum_i \big(I_i(y) - a_0 - a_1 R_i\big)^2 K_h(R_i)\,1\{T_i = 1, R_i \geq 0\}.$$

With $Z^+_i = (1, R_i)^\top K_h(R_i) 1\{T_i = 1, R_i \geq 0\}$, we have

$$\hat{a}^+_0(y) = e_1^\top \big(\textstyle\sum Z^+_i Z^{+\top}_i\big)^{-1} \sum Z^+_i I_i(y).$$

A first-order expansion of the normal equations at the population values yields

$$\sqrt{n_{+,h}}\big(\hat{a}^+_0(y) - F_{1,1}(y)\big) = \frac{1}{\sqrt{n_{+,h}}} \sum_i \frac{K_h(R_i)\,\mathbf{1}\{T_i = 1, R_i \geq 0\}\,\ell^{(1)}_+(R_i;h)}{\mathbb{E}\big[K_h(R)\,\mathbf{1}\{T = 1, R \geq 0\}\,\big(\ell^{(1)}_+(R;h)\big)^2\big]} \times \big(I_i(y) - F_{1,1}(y)\big) + o_p(1),$$

which gives $\psi_{1,1,y}(W_i)$ in (23).[6]

(ii) Left-limit CDF in $T = 1$. Analogously, for the T=1, $R < 0$ LL fit,

$$(\hat{a}^-_0(y), \hat{a}^-_1(y)) = \arg\min_{a_0,a_1} \sum_i \big(I_i(y) - a_0 - a_1 R_i\big)^2 \times K_h(R_i)\,\mathbf{1}\{T_i = 1, R_i < 0\},$$

we obtain

$$\sqrt{n_{-,h}}\big(\hat{a}^-_0(y) - F_{0,0}(y)\big) = \frac{1}{\sqrt{n_{-,h}}} \sum_i \frac{K_h(R_i)\,\mathbf{1}\{T_i = 1, R_i < 0\}\,\ell^{(1)}_-(R_i;h)}{\mathbb{E}\big[K_h(R)\,\mathbf{1}\{T = 1, R < 0\}\,\big(\ell^{(1)}_-(R;h)\big)^2\big]} \times \big(I_i(y) - F_{0,0}(y)\big) + o_p(1),$$

which is $\psi_{0,0,y}(W_i)$ in (24).

[6] See Frandsen et al. (2012, Sec. 4 and App. C); the same algebra for boundary LL appears in Calonico et al. (2014, App. B).

(iii) Interior local-linear regression for $\Delta_1(y)$. With $Z_i(y) := F_0^+(y; X_i) - F_0^-(y; X_i)$, consider the $T = 1$ LL problem

$$(\hat{a}_1(y), \hat{b}_1(y)) = \arg\min_{a_0, a_1} \sum_i \big(Z_i(y) - a_0 - a_1 R_i\big)^2 K_{h_0}(R_i) \mathbf{1}\{T_i = 1\},$$

and define the estimator

$$\hat{\Delta}_1(y) := \hat{a}_1(y).$$

Let

$$S_m^{(1)}(h_0) := \mathbb{E}\left[\left(\tfrac{R}{h_0}\right)^m K_{h_0}(R)\, \mathbf{1}\{T = 1\}\right], \quad m = 0, 1, 2,$$

$$D^{(1)}(h_0) := S_0^{(1)}(h_0)\, S_2^{(1)}(h_0) - \big(S_1^{(1)}(h_0)\big)^2,$$

and the interior design adjustment

$$\ell^{(1)}(r; h_0) := S_2^{(1)}(h_0) - S_1^{(1)}(h_0)\, \frac{r}{h_0}.$$

Let

$$M_{\Delta_1}(r, y) := \mathbb{E}\big[Z(y, X) \mid R = r, T = 1\big], \quad \text{so that} \quad \Delta_1(y) = M_{\Delta_1}(0, y).$$

A first-order expansion of the LL normal equations around the population values (Newey, 1994) yields

$$\begin{aligned}\sqrt{n_{1;h_0}}\big(\hat{\Delta}_1(y) - \Delta_1(y)\big) = \frac{1}{\sqrt{n_{1;h_0}}} \sum_i K_{h_0}(R_i)\, \mathbf{1}\{T_i = 1\} \\ \times \frac{\ell^{(1)}(R_i; h_0)}{D^{(1)}(h_0)} \Big\{ Z_i(y) - M_{\Delta_1}(R_i, y) \Big\} + o_p(1),\end{aligned}$$

so the influence function contribution of $\Delta_1(y)$ is

$$\begin{aligned}\phi_{\Delta_1; y}(W_i) = K_{h_0}(R_i)\, \mathbf{1}\{T_i = 1\}\, \frac{\ell^{(1)}(R_i; h_0)}{D^{(1)}(h_0)} \\ \times \Big\{ F_0^+(y; X_i) - F_0^-(y; X_i) - M_{\Delta_1}(R_i, y) \Big\}.\end{aligned}$$

For a symmetric kernel K and an interior evaluation point $R = 0$, we have $S_1^{(1)}(h_0) = 0$ at the population level, so that

$$\ell^{(1)}(r; h_0) = S_2^{(1)}(h_0), \qquad D^{(1)}(h_0) = S_0^{(1)}(h_0)\, S_2^{(1)}(h_0),$$

and hence

$$\frac{\ell^{(1)}(r; h_0)}{D^{(1)}(h_0)} = \frac{1}{S_0^{(1)}(h_0)} = \frac{1}{\mathbb{E}\big[K_{h_0}(R)\, \mathbf{1}\{T = 1\}\big]}.$$

In this case the influence function simplifies to

$$\phi_{\Delta_1; y}(W_i) = \frac{K_{h_0}(R_i)\, \mathbf{1}\{T_i = 1\}}{\mathbb{E}\big[K_{h_0}(R)\, \mathbf{1}\{T = 1\}\big]} \Big\{ F_0^+(y; X_i) - F_0^-(y; X_i) - M_{\Delta_1}(R_i, y) \Big\},$$

which coincides with the expression in (25) in the main text.

Remark 6 *(iv) Residual LL fits for $\Delta_2^{\pm}(y)$ in $T = 0$. Define the residualized outcomes on T=0,*

$$\begin{aligned} U_i^+(y) &: = \widetilde{w}_i\big(I_i(y) - F_0^+(y, X_i)\big) \text{ on } R \geq 0, \\ U_i^-(y) &: = \widetilde{w}_i\big(I_i(y) - F_0^-(y, X_i)\big) \text{ on } R < 0. \end{aligned}$$

Run the same boundary LL problems as in (i)-(ii) with T=0 and outcomes $U_i^{\pm}(y)$. The identical expansion gives

$$\sqrt{n_{0,+,h}}\big(\widehat{\Delta}_2^+(y) - \Delta_2^+(y)\big) = \frac{1}{\sqrt{n_{0,+,h}}} \sum_i \frac{K_h(R_i)\,\mathbf{1}\{T_i{=}0,\, R_i{\geq}0\}\,\ell_+^{(0)}(R_i; h)}{\mathbb{E}\big[K_h(R)\,\mathbf{1}\{T{=}0,\, R{\geq}0\}\,(\ell_+^{(0)}(R; h))^2\big]} \times \big(U_i^+(y) - \Delta_2^+(y)\big) + o_p(1),$$

and the analogous left-side expression for $\widehat{\Delta}_2^-(y)$, which are (26)-(27)*.*

(v) DR assembly for $F_{1,0}(y)$. Since $\hat{F}_{1,0}(y) = \hat{F}_{0,0}(y) + \hat{\Delta}_1(y) + \hat{\Delta}_2^+(y) - \hat{\Delta}_2^-(y)$, linearity gives the unnormalized influence function

$$\widetilde{\psi}_{1,0,y}(W_i) = \widetilde{\psi}_{0,0,y}(W_i) + \widetilde{\psi}_{\Delta_1,y}(W_i) + \widetilde{\psi}_{\Delta_2^+,y}(W_i) - \widetilde{\psi}_{\Delta_2^-,y}(W_i),$$

which is (28)*. Under the pooled $\sqrt{nh}$ scaling and $n_t/n \to p_t$, the normalized influence function in the main text is*

$$\begin{aligned} \psi_{1,0,y}(W_i) \coloneqq\, & \mathbf{1}\{T_i = 1\}\frac{1}{\sqrt{p_1}}\big\{\psi_{0,0,y}(W_i) + \psi_{\Delta_1,y}(W_i)\big\} \\ & + \mathbf{1}\{T_i = 0\}\frac{1}{\sqrt{p_0}}\big\{\psi_{\Delta_2^+,y}(W_i) - \psi_{\Delta_2^-,y}(W_i)\big\}, \end{aligned}$$

as stated in equation (28).

(vi) From CDF to quantiles. Let $q_{1,d,\tau} {:=} Q_{Y(1,d)|R=0,T=1}(\tau)$ with $f_{1,d}(q_{1,d,\tau}) > 0$. The quantile map is Hadamard differentiable with derivative $-\psi_{1,d,q}/f_{1,d}(q)$ for each $d \in \{0, 1\}$, so the influence function for the QTE estimator under the $\sqrt{nh}$ scaling is

$$\phi_\tau(W_i) {:=} -\frac{1}{\sqrt{p_1}}\frac{\psi_{1,1,\,q_{1,1,\tau}}(W_i)}{f_{1,1}(q_{1,1,\tau})} + \frac{\psi_{1,0,\,q_{1,0,\tau}}(W_i)}{f_{1,0}(q_{1,0,\tau})},$$

as stated in equation (31) *of the main text. Rearrangement does not alter first-order asymptotics (Chernozhukov et al., 2010).*

10.3 Leading bias of the QTE estimator $\mu(\tau)$

The QTE estimator has the following asymptotic linear representation:

$$\sqrt{nh}\big(\hat{\delta}(\tau) - \delta(\tau)\big) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \phi_\tau(W_i) \;+\; \gamma\,\mu(\tau) \;+\; o_p(1),$$

where

$$\mu(\tau) := \beta_{1,1}(\tau) - \beta_{1,0}(\tau),$$

with $\beta_{1,d}(\tau) := -\, b_{1,d}\big(q_{1,d,\tau}\big) \,/\, f_{1,d}\big(q_{1,d,\tau}\big)$, $d \in \{0,1\}$ and $b_{1,d}(y)$ is the leading h^2 bias constant of the *CDF* estimator used for $F_{1,d}(y)$.

Let

$$\begin{aligned}
c_b(K) &:= \tfrac{1}{2}\, e_0^\top \Gamma_{1,+}^{-1} \Lambda_{1,+}, \\
\Gamma_{1,+} &= \int_{u \geq 0} K(u) \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 & u \end{pmatrix} du, \\
\Lambda_{1,+} &= \int_{u \geq 0} K(u) \begin{pmatrix} u^2 \\ u^3 \end{pmatrix} du, \\
c_i(K) &:= \tfrac{1}{2}\, \mu_2(K), \\
\mu_2(K) &= \int_{-1}^{1} u^2 K(u)\, du,
\end{aligned}$$

where $e_0 = (1,0)^\top$. For any symmetric kernel K, $c_b(K)$ is side-invariant: the same constant applies to left and right one-sided LLR fits.

Let m'' for any function m be the second derivative with respect to r at 0.

- Right-side period 2 CDF $F_{1,1}(y) = \lim_{r \downarrow 0} g_{1,+}(r;y)$, where $g_{1,+}(r;y) := \mathbb{E}[\mathbf{1}\{Y \leq y\} \mid R = r,\, T = 1]$, $r \geq 0$. The leading h^2 bias constant is

$$b_{1,1}(y) = c_b(K)\, g''_{1,+}(0;y).$$

- Composite $F_{1,0}(y) = F_{0,0}(y) + \Delta_1(y) + \Delta_2^+(y) - \Delta_2^-(y)$. The bias constants combine linearly:

$$b_{1,0}(y) = b_{0,0}(y) + b_{\Delta_1}(y) + b_{\Delta_2^+}(y) - b_{\Delta_2^-}(y).$$

  - Left-side period 2 CDF $F_{0,0} = \lim_{r \uparrow 0} g_{1,-}(r;y)$, where $g_{1,-}(r;y) := \mathbb{E}[\mathbf{1}\{Y \leq y\} \mid R = r,\, T = 1]$, $r < 0$. The bias constant is

$$b_{0,0}(y) = c_b(K)\, g''_{1,-}(0;y).$$

  - Interior period 2 $\Delta_1(y) = \lim_{r \to 0} M_{\Delta_1}(r;y)$, where $M_{\Delta_1}(r;y) := \mathbb{E}\big[F_0^+(y,X) - F_0^-(y,X) \mid R = r,\, T = 1\big]$. Then the LLR interior bias is

$$b_{\Delta_1}(y) = c_i(K)\, M''_{\Delta_1}(0;y),$$

    since for triangular K the interior LLR constant is $\frac{1}{2}\mu_2(K) = \frac{1}{12}$.

  - $\Delta_2^+(y) = \lim_{r \downarrow 0} m_{\Delta_2^+}(r;y)$ and $\Delta_2^-(y) = \lim_{r \uparrow 0} m_{\Delta_2^-}(r;y)$, where

$$m_{\Delta_2^s}(r;y) := \mathbb{E}\big[\widetilde{w}(X_i)\{I_i(y) - F_0^s(y,X_i)\} \mid R = r,\, T = 0\big],$$

    for $s \in \{+,-\}$ and $sr > 0$. The general bias constant is

$$b_{\Delta_2^s}(y) = c_b(K)\, m''_{\Delta_2^s}(0;y)\,.$$

Together, we have

$$b_{1,0}(y) = c_b(K)\, g''_{1,-}(0;y) + c_i(K)\, M''_{\Delta_1}(0;y) + c_b(K)\, m''_{\Delta_2^+}(0;y) - c_b(K)\, m''_{\Delta_2^-}(0;y)\,,$$

and further

$$\begin{aligned}\mu(\tau) \;=\; & -c_b(K)\,\frac{g''_{1,+}(0;q_{1,1,\tau})}{f_{1,1}(q_{1,1,\tau})} \\ & + \frac{1}{f_{1,0}(q_{1,0,\tau})}\left[\begin{array}{l} c_b(K)\, g''_{1,-}(0;q_{1,0,\tau}) + c_i(K)\, M''_{\Delta_1}(0;q_{1,0,\tau}) \\ + c_b(K)\, m''_{\Delta_2^+,+}(0;q_{1,0,\tau}) \\ - c_b(K)\, m''_{\Delta_2^-,-}(0;q_{1,0,\tau}) \end{array}\right].\end{aligned}$$

- **Triangular** $K(u) = (1-|u|)\mathbf{1}\{|u| \le 1\}$: $c_b(K) = -\frac{1}{20}$, $c_i(K) = \frac{1}{12}$.
- **Uniform** $K(u) = \frac{1}{2}\mathbf{1}\{|u| \le 1\}$: $c_b(K) = -\frac{1}{12}$, $c_i(K) = \frac{1}{6}$.
- **Epanechnikov** $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}\{|u| \le 1\}$: $c_b(K) = -\frac{11}{190}$, $c_i(K) = \frac{1}{10}$.

10.4 Pilot bandwidth

Because CDF estimation uses binary outcomes $I(Y \le y)$ or their smoothed counterparts $\Phi\big((y-Y_i)/h_Y\big)$, the quantile level associated with y is unknown a priori, making the bandwidth choices in (34) - (35) not immediately usable. We therefore adopt a plug-in rule that maps $y \mapsto \hat{\tau}(y)$ via a pilot CDF and then applies the FFM (2012) scaling.

Step 0 (reference mean bandwidths). For each period $t \in \{0,1\}$ and side $s \in \{-,+\}$, compute a plug-in bandwidth $h^{t,s}_{\text{mean}}$ for the boundary mean (e.g., IK or CCT). For the centered local-linear fit in $\Delta_1(y)$, compute a centered mean bandwidth h^1_{mean} (e.g., the harmonic mean of $h^{1,+}_{\text{mean}}$ and $h^{1,-}_{\text{mean}}$).

Step 1 (pilot CDF at y). Using the reference mean bandwidths (no τ-scaling), estimate the relevant pilot CDFs at y:

- $\tilde{F}_{1,1}(y)$ from the T=1, $R \ge 0$ fit;
- $\tilde{F}_{0,0}(y)$ from the T=1, $R < 0$ fit;
- $\tilde{\Delta}_2^+(y)$ and $\tilde{\Delta}_2^-(y)$ from the T=0 residual fits;
- $\widetilde{\Delta}_1(y)$ from the centered local-linear fit in $T = 1$.

Define $\hat{\tau}_{1,1}(y) := \tilde{F}_{1,1}(y)$ and $\hat{\tau}_{0,0}(y) := \tilde{F}_{0,0}(y)$; for the residual components, use the same $\hat{\tau}_{0,0}(y)$ on the corresponding side for stability. Clip $\hat{\tau}(\cdot)$ to $[\underline{\tau}, 1-\underline{\tau}]$ for a small $\underline{\tau} \in (0, 0.1)$ to avoid tail instability.

Step 2 (FFM scaling with $\hat{\tau}(y)$). For ϕ and Φ the standard normal pdf and cdf, set

$$h^{t,s}_{\hat{\tau}(y)} = h^{t,s}_{\text{mean}} \left[\frac{\hat{\tau}(y)\big(1-\hat{\tau}(y)\big)}{\phi\big(\Phi^{-1}(\hat{\tau}(y))\big)^2} \right]^{1/5}, \qquad t \in \{0,1\},\ s \in \{-,+\}, \tag{37}$$

$$h^{1}_{\hat{\tau}(y)} = h^{1}_{\text{mean}} \left[\frac{\hat{\tau}(y)\big(1-\hat{\tau}(y)\big)}{\phi\big(\Phi^{-1}(\hat{\tau}(y))\big)^2} \right]^{1/5}. \tag{38}$$

If diagnostics suggest similar smoothness on both sides, impose $h^{t,+}_{\hat{\tau}(y)} = h^{t,-}_{\hat{\tau}(y)}$ within period t; otherwise keep them side-specific.

Step 3 (final CDFs at y). Re-estimate each component at y with the scaled bandwidths (37)–(38):

$$(\hat{a}^{1,+}_{0,\hat{\tau}}(y), \hat{a}^{1,+}_{1,\hat{\tau}}(y)) = \arg\min_{a_0,a_1} \sum_{i:T_i=1,\, R_i\geq 0} \big(I_i(y) - a_0 - a_1 R_i\big)^2 K_{h^{1,+}_{\hat{\tau}(y)}}(R_i),$$
$$\hat{F}_{1,1}(y) = \hat{a}^{1,+}_{0,\hat{\tau}}(y), \tag{39}$$

$$(\hat{a}^{1,-}_{0,\hat{\tau}}(y), \hat{a}^{1,-}_{1,\hat{\tau}}(y)) = \arg\min_{a_0,a_1} \sum_{i:T_i=1,\, R_i<0} \big(I_i(y) - a_0 - a_1 R_i\big)^2 K_{h^{1,-}_{\hat{\tau}(y)}}(R_i),$$
$$\hat{F}_{0,0}(y) = \hat{a}^{1,-}_{0,\hat{\tau}}(y), \tag{40}$$

$$\Big(\widehat{d}_{0,\widehat{\tau}}(y), \widehat{d}_{1,\widehat{\tau}}(y)\Big) = \arg\min_{d_0,d_1} \sum_{i:T_i=1} \Big(\widehat{F}^{+}_{0}(y,X_i) - \widehat{F}^{-}_{0}(y,X_i) - d_0 - d_1 R_i\Big)^2$$
$$\times K_{b^1_{\hat{\tau}(y)}}(R_i), \quad \widehat{\Delta}_1(y) = \widehat{d}_{0,\widehat{\tau}}(y), \tag{41}$$

$$(\hat{c}^{0,+}_{0,\hat{\tau}}(y), \hat{c}^{0,+}_{1,\hat{\tau}}(y)) = \arg\min_{c_0,c_1} \sum_{i:T_i=0,\, R_i\geq 0} \big(\hat{w}_i\{I_i(y) - \hat{F}^{+}_{0}(y,X_i)\} - c_0 - c_1 R_i\big)^2$$
$$\times K_{h^{0,+}_{\hat{\tau}(y)}}(R_i), \quad \widehat{\Delta}^{+}_2(y) = \hat{c}^{0,+}_{0,\hat{\tau}}(y), \tag{42}$$

$$(\hat{c}^{0,-}_{0,\hat{\tau}}(y), \hat{c}^{0,-}_{1,\hat{\tau}}(y)) = \arg\min_{c_0,c_1} \sum_{i:T_i=0,\, R_i<0} \big(\hat{w}_i\{I_i(y) - \hat{F}^{-}_{0}(y,X_i)\} - c_0 - c_1 R_i\big)^2$$
$$\times K_{h^{0,-}_{\hat{\tau}(y)}}(R_i), \quad \widehat{\Delta}^{-}_2(y) = \hat{c}^{0,-}_{0,\hat{\tau}}(y). \tag{43}$$

Finally, $\widehat{\Delta}^{DR}(y) = \widehat{\Delta}_1(y) + \widehat{\Delta}^{+}_2(y) - \widehat{\Delta}^{-}_2(y)$ and $\hat{F}_{1,0}(y) = \hat{F}_{0,0}(y) + \widehat{\Delta}^{DR}(y)$. One pilot iteration is typically sufficient; a second pass rarely changes results materially.

Remarks. (i) Treating bandwidths as fixed for first-order inference is standard; randomness from the pilot is $o_p(1)$ under usual conditions. (ii) The mapping $y \mapsto \hat{\tau}(y)$ introduces smooth variation in the bandwidth across y, which is compatible with local polynomial theory. (iii) Near extreme quantiles, enforce

a minimum bandwidth to stabilize density estimates used in the quantile delta method.