Regression Discontinuity Designs with a Continuous Treatment

Yingying Dong, Ying-Ying Lee, Michael Gou*

First version: April 2017; this version: April 2021

Abstract

The standard regression discontinuity (RD) design deals with a binary treatment. Many empirical applications of RD designs involve continuous treatments. This paper establishes identification and robust bias-corrected inference for such RD designs. Causal identification is achieved by utilizing any changes in the distribution of the continuous treatment at the RD threshold (including the usual mean change as a special case). We discuss a double-robust identification approach and propose an estimand that incorporates the standard fuzzy RD estimand as a special case. Applying the proposed approach, we estimate the impacts of bank capital on bank failure in the pre-Great Depression era in the United States. Our RD design takes advantage of the minimum capital requirements, which change discontinuously with town size.

JEL codes: C21, C26, E58

Keywords: Continuous treatment, Treatment quantiles, Rank invariance, Rank similarity, Double-robust identification

*Yingying Dong and Ying-Ying Lee, Department of Economics, University of California Irvine, yyd@uci.edu and yingying.lee@uci.edu; Michael Gou, PricewaterhouseCoopers, michaelgou@gmail.com.

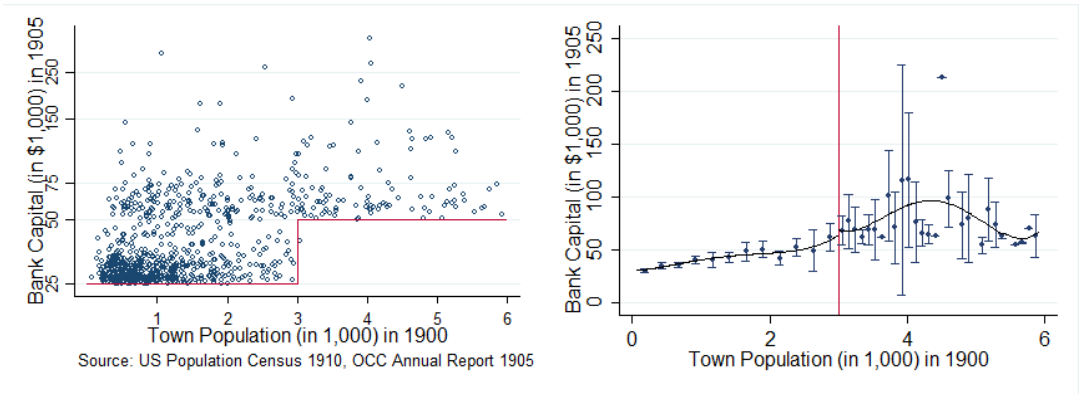

Figure 1: Scatter plot (left) and the RD mean plot (right) of bank capital against town population

1 Introduction

RD designs have been widely used for causal analysis in many disciplines, including economics, political science, education, epidemiology, public health, and medicine. The standard RD design assumes a binary treatment. In practice, many empirical applications of RD designs involve continuous treatments, e.g., alcohol consumption around the minimum legal drinking age, air pollution across neighboring geographical regions, or medical expenditure around the low birth weight cutoff (Almond et al., 2010, Litschig and Morrison, 2010, Chen et al., 2013, Ebenstein, 2017, Giuntella and Mazzonna, 2019, and Fan et al., 2020). In this paper we consider nonparametric identification and inference for fuzzy RD designs with a continuous treatment, where the distribution of the continuous treatment variable changes at the RD threshold.

Consider our empirical question for concreteness - are banks less likely to fail when they hold more capital? To provide a credible estimate of the causal effect of bank capital on bank failure, one needs some quasi-experimental variation in bank capital. As seen in Figure 1 (left), one potential source of variation is the relationship between the minimum capital requirement and town size in the early 20th century of the United States – as town size crosses a certain threshold, the minimum capital requirement (marked by the solid line) jumps up and the bottom of the capital distribution shifts up correspondingly. Given this relationship, one may be tempted to apply the standard RD estimand, i.e., the RD local Wald ratio that associates a mean change in the outcome (bank failure) with a mean change in the treatment (bank capital) at the RD threshold.

Hahn, Todd, and van der Klaauw (2001) show that under proper conditions, the RD local Wald ratio with a binary treatment identifies a local average treatment effect (LATE) for compliers at the RD threshold. In RD designs with a continuous treatment, empirical researchers typically apply this RD local Wald ratio to estimate the causal effect of the continuous treatment. Causal identification and inference rely solely on the mean shift of the continuous treatment variable. A few issues arise with this practice. The first issue is interpretation – we show in Section 3.1 that the LATE interpretation no longer holds with a continuous treatment. Intuitively, there are an infinite number of potential outcomes, and compliers are not immediately defined.

The second issue is potential weak identification or identification failure, when there is little or no mean change in the treatment variable. In our empirical scenario, the discontinuous relationship between the minimum capital requirement (the policy instrument) and town size generates only a weak first-stage discontinuity in the relationship between the average bank capital and town size. Figure 1 (right) plots the mean capital against town size along with the 95% confidence intervals. No significant changes are found in the mean capital at the threshold. The standard fuzzy RD estimation does not directly apply.

The third issue is policy relevance. The average level of treatment may not always be the appropriate measure to look at from a policy perspective. In practice, many polices target some parts (e.g., top or bottom) of the treatment distribution or aim to change some features of the distribution (e.g., reducing the variance). The minimum capital requirement, the policy instrument here, targets banks at the bottom of the capital distribution. Similarly, many other treatment guidelines or policies frequently target one or two tails of the treatment distribution. Examples include the minimum or maximum recommended medication dosage, minimum wages, maximum welfare benefits, government transfers that are capped at certain levels, and the pollution ceiling set by the environmental protection agency. Focusing on the mean treatment may miss the true sources of identification, i.e., where the true changes are in the treatment distribution.

In this paper, we show that causal identification can be achieved by utilizing any changes in the distribution of the continuous treatment variable at the RD threshold. These include not only the

usual mean change, but also changes at various points of the treatment distribution. By focusing on where the true exogenous changes are in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects.

We identify and provide inference for two types of causal effects. The first is the LATE at a particular treatment quantile. We refer to this quantile specific LATE as Q-LATE. Q-LATE captures treatment effect heterogeneity at different treatment intensities, which the standard RD design fails to capture by solely focusing on the average treatment change in the first stage. For example, Q-LATE can be useful if one is interested in examining diminishing returns to treatment. The second is a weighted average of Q-LATEs averaging over the treatment distribution at the RD threshold (WQ-LATE). Importantly, we discuss a double-robust approach and provide a WQ-LATE estimand that incorporates the standard RD estimand, the RD local Wald ratio, as a special case. When the standard RD estimand is valid, the proposed estimand reduces to the standard RD estimand; when the standard RD estimand is not valid, the proposed estimand continues to be valid under our alternative assumptions. In addition, we develop robust bias-correct inference and the asymptotic mean squared error (AMSE) optimal bandwidths for estimating either effect.

Our empirical application demonstrates the usefulness of the proposed approach. The minimum capital requirement shifts up the bottom of the capital distribution, but leads to no mean change in bank capital. We cannot apply the standard fuzzy RD estimation. However, taking advantage of lower quantile changes in the capital distribution, we are able to quantify the causal impacts of increased capital on banks' short-run responses and long-run failure rates particularly among those banks targeted by the minimum capital policy.

Our paper adds to the growing literature of RD designs, which focuses on binary treatments. See, Imbens and Lemieux (2008) for an early review of the RD literature. For more recent reviews, see Cattaneo, Idrobo and Titiunik (2019) and Cattaneo, Idrobo and Titiunik (2020a and b). Note that our model is different than the RD quantile treatment effect (RD QTE) model discussed by Frandsen, Frölich, and Melly (2012). The RD QTE model still requires a binary treatment along with a continuous outcome. In contrast, our model requires a continuous treatment with either a

discrete or continuous outcome. RD QTE captures treatment effect heterogeneity at different points of the outcome distribution, while our Q-LATE parameter captures treatment effect heterogeneity at different points of the treatment distribution. Caetano, Caetano, and Escanciano (2020) discuss identification and estimation of RD designs with a multi-valued treatment variable. A continuous treatment has been considered in the literature of regression kink (RK) designs (Card et al., 2015). In RK designs, identification relies on treatment assignment as a kinked function of the running variable.

Our paper is related to the non-separable instrumental variable (IV) literature with continuous endogenous covariates. Identification in this literature typically requires a scalar unobservable (rank invariance) in either the first stage or the outcome equation or both (see, e.g., discussion in Torgovitsky, 2015, and D'haultfoeuille and Février, 2015). In contrast, we allow for rank similarity (instead of just rank invariance) in the first stage and unrestricted multidimensional unobservables in the outcome equation.

The rest of the paper proceeds as follows. Section 2 discusses causal identification and the parameters of interest. Section 3 proposes a causal estimand that incorporates the standard fuzzy RD estimand as a special case. Section 4 describes estimation and specification testing. Section 5 provides robust bias-corrected inference and the AMSE optimal bandwidths for the Q-LATE and WQ-LATE estimators. Section 6 presents the empirical analysis. Concluding remarks are provided in Section 7. All proofs, alternative inference based on undersmoothing, details of estimating the biases, variances and AMSE optimal bandwidths of the proposed estimators, as well as additional empirical results are gathered in the Appendix.

2 Identification

In this section, we discuss nonparametric identification of RD designs with a continuous treatment. To fix the idea, ignore the running variable for now. Consider a continuous treatment T and a binary "IV" Z. For an observation i, let $T_i = a_i + b_i Z_i$, where a_i and b_i are random coefficients. Typically

one would estimate a constant coefficient regression in the first stage of the linear IV model, where the constant coefficient of the binary Z captures the exogenous change in the mean treatment. Here we show that under proper conditions, the random coefficient b_i captures exogenous changes in the distribution of the treatment, which can be used for identification.

2.1 Basic setup

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be the outcome of interest, and $T \in \mathcal{T} \subset \mathbb{R}$ be the treatment. Let $R \in \mathcal{R} \subset \mathbb{R}$ be the continuous running variable that partly determines the treatment. Assume $Y = G(T, R, \varepsilon)$, where $\varepsilon \in \mathcal{E} \subset \mathbb{R}^{d_\varepsilon}$ is allowed to be of arbitrary dimension. Further assume that T has a reduced-form equation $T = q(R, U)$ with a reduced-form disturbance U.

Define $Z \equiv \mathbf{1}(R \geq r_0)$ for some known threshold value r_0, where $\mathbf{1}(\cdot)$ is an indicator function equal to 1 if the expression in the parentheses is true and 0 otherwise. Given that Z is binary and is a deterministic function of R, without loss of generality, one can write $T = q_1(R, U_1)Z + q_0(R, U_0)(1 - Z)$, where $U_z \in \mathcal{U}_z \subset \mathbb{R}$, $z = 0, 1$. Let $T_z \equiv q_z(R, U_z)$, $z = 0, 1$ be the potential treatment when Z is exogenously set at z. One can then write $T = T_1 Z + T_0(1 - Z)$ and correspondingly $U = U_1 Z + U_0(1 - Z)$.

In the following we establish identification of the conditional RD LATE given $U = u$, i.e., $\mathbb{E}\left[\frac{Y_{t_1(u)} - Y_{t_0(u)}}{t_1(u) - t_0(u)} | U = u, R = r_0\right]$, where the potential outcome $Y_t \equiv G(t, R, \varepsilon)$, $t_0(u) \equiv q_0(r_0, u)$, and $t_1(u) \equiv q_1(r_0, u)$. It will be shown that the potential treatment value change $t_1(u) - t_0(u)$ captures the exogenous change in the u quantile of the continuous treatment under our identifying assumptions. We refer to this conditional RD LATE given $U = u$ as quantile specific LATE or Q-LATE. We further discuss identification of some weighted average of Q-LATE averaging over the distribution of U at $R = r_0$, which we refer to as WQ-LATE.

Denote the conditional cumulative distribution function (CDF) as $F_{\cdot|\cdot}(\cdot, \cdot)$, the conditional probability density function (PDF) as $f_{\cdot|\cdot}(\cdot, \cdot)$ and the unconditional PDF as $f_{\cdot}(\cdot)$.

Assumption 1 (Quantile representation)**.** *$q_z(r, u)$, $z = 0, 1$, is strictly monotonic in u for any $r \in \mathcal{R}$, where $\mathcal{R}$ is an arbitrarily small closed interval around r_0. The conditional distribution of*

T_z given $R = r$ is continuous with a strictly increasing CDF $F_{T_z|R}(t, r)$.

Assumption 1 imposes monotonicity on the unobserved heterogeneity in the first stage. Given Assumption 1, one can normalize U_z to be $F_{T_z|R}(T_z, R)$, so $U_z \sim Unif\,(0, 1)$. That is, U_z is the conditional rank of T_z given R, and $q_z(r, u)$ is the conditional u quantile of T_z given $R = r$.

Assumption 2 (Smoothness)**.** *$q_z\,(r, u)$, $z = 0, 1$, is continuous in $r \in \mathcal{R}$ for any $u \in [0, 1]$. Either $G\,(t, r, e)$ is continuous in all its arguments, or it is a.e. continuous and bounded. $f_{\varepsilon|U_z R}\,(e, u, r)$ is continuous in $r \in \mathcal{R}$ for any $u \in [0, 1]$ and $e \in \mathcal{E}$, where $\mathcal{E}$ is compact. $f_R\,(r)$ is continuous and strictly positive around r_0.*

Assumption 3 (Local treatment rank invariance or similarity)**.** *Conditional on $R = r_0$, 1. $U_0 = U_1$; or more generally, 2. $U_0|\varepsilon \sim U_1|\varepsilon$.*

Assumption 4 (First-stage)**.** *$t_1(u) \neq t_0(u)$ for at least some $u \in [0, 1]$.*

Assumption 2 assumes that the running variable has only smooth effects on potential treatments and that the treatment, running variable, and unobservables all impose smooth impacts on the outcome. It further assumes that at a given rank of the potential treatment, the distribution of the unobservables in the outcome model is smooth near the RD threshold. The last condition, the running variable is continuous with a positive density around the RD threshold, is standard and is typically required for RD designs (see, e.g., Hahn, Todd, and van der Klaauw, 2001).

Note that R, U_z, and ε are required to have compact support, which serves as a regularity condition. The continuity conditions in Assumption 2 along with compact support ensures interchangeability of limit and integral (expectation). It follows that $\mathbb{E}\left[G\,(q_z(r, u), r, \varepsilon)\,|U_z = u, R = r\right] = \int_{\mathcal{E}} G\,(q_z\,(r, u)\,, r, \varepsilon)\; f_{\varepsilon|U_z R}\,(e, u, r)\,de,\; z = 0, 1$, is continuous in r, which is the key to causal identification in our setup. Without compact support, other alternative regularity conditions need to be imposed under which one can interchange limit and integral.

Assumption 3 imposes local treatment rank restrictions. That is, treatment rank invariance or similarity is required to hold only at the RD cutoff. Assumption 3.1 requires units to stay at the same rank of the potential treatment distribution right above or below the RD threshold.

Assumption 3.2 assumes rank similarity, a weaker condition than Assumption 3.1. Without conditioning on ε, U_0 and U_1 given $R = r_0$ both follow a uniform distribution over the unit interval, i.e., $U_0|\,(R = r_0) \sim U_1|\,(R = r_0)$ by construction. Local rank similarity permits random "slippages" from the common rank level in the treatment distribution just above or just below the RD cutoff. Rank similarity has been proposed to identify quantile treatment effects (QTEs) in IV models (Chernozhukov and Hansen, 2005). Unlike the IV QTE model, we impose the similarity assumption on the ranks of potential treatments, instead of the ranks of potential outcomes. In our empirical analysis, Assumption 3.2 requires that the probability for a bank to stay at the certain rank of the capital distribution stays the same regardless of whether it is in a town with a population just above or just below 3,000.

Assumption 4 requires that the distribution of treatment changes at $R = r_0$. This is strictly weaker than the standard RD design first-stage assumption that requires a mean change in treatment, i.e., $\mathbb{E}[T_1|R = r_0] \neq \mathbb{E}[T_0|R = r_0]$.

The above identifying assumptions have potentially testable implications. Under either 3.1 or 3.2, $U_0|\,(\varepsilon, R = r_0) \sim U_1|\,(\varepsilon, R = r_0)$. By Bayes' theorem, $U_0|\,(\varepsilon, R = r_0) \sim U_1|\,(\varepsilon, R = r_0)$ if and only if $\varepsilon|\,(U_0 = u, R = r_0) \sim \varepsilon|\,(U_1 = u, R = r_0)$. Let X be some observable component of ε, assuming such X exists. Then $X|\,(U_0 = u, R = r_0) \sim X|\,(U_1 = u, R = r_0)$. Further by Assumption 2, $F_{X|U_z R}(x, u, r)$, $z = 0, 1$, is continuous at $r = r_0$. One can then test the condition $\lim_{r \to r_0^+} F_{X|UR}(x, u, r) - \lim_{r \to r_0^-} F_{X|UR}(x, u, r) = 0$. Later in Section 4 we discuss a convenient falsification test based on this testable implication.

2.2 Identification results

Lemma 1 below presents some preliminary results to facilitate the discussion of causal parameters and identification in our setup.

Lemma 1. *Let Assumptions 1-3 hold. For any* $u \in [0, 1]$,

1. $\lim_{r \to r_0^-} f_{\varepsilon|TR}(e, q_0(r, u), r) = \lim_{r \to r_0^+} f_{\varepsilon|TR}(e, q_1(r, u), r)$
   $= \lim_{r \to r_0} f_{\varepsilon|UR}(e, u, r)$ *for* $e \in \mathcal{E}$.

2. $\lim_{r \to r_0^+} \mathbb{E}[Y|U=u, R=r] - \lim_{r \to r_0^-} \mathbb{E}[Y|U=u, R=r]$
$= \int (G(t_1(u), r_0, e) - G(t_0(u), r_0, e)) F_{\varepsilon|UR}(de, u, r_0)$.

Given $U = u$, T can take on two limiting values as $R \to r_0$, $t_0(u) \equiv q_0(r_0, u)$ and $t_1(u) \equiv q_1(r_0, u)$. By Assumption 2, $\lim_{r \to r_0^-} f_{\varepsilon|TR}(e, q_0(r, u), r) = f_{\varepsilon|TR}(e, t_0(u), r_0)$ and $\lim_{r \to r_0^+} f_{\varepsilon|TR}(e, q_1(r, u), r) = f_{\varepsilon|TR}(e, t_1(u), r_0)$. Lemma 1.1 shows $T \perp \varepsilon | U$, as $R \to r_0$, i.e., conditional on the treatment rank U, any potential changes in T as $R \to r_0$ are independent of ε. Note that conditioning on $U = u$ is implicit in the first equality of Lemma 1.1, since given $R = r$, T and U follow a one-to-one mapping by Assumption 1.

Lemma 1.1 can be seen as a local limiting version of the Imbens and Newey (2009) type of identification condition. The local independence makes U a (local) control variable as defined by Imbens and Newey (2009). The defining feature of any "control variable" is that conditional on this variable (along with possibly other covariates), treatment is exogenous to the outcome of interest.

Here the "IV" $Z \equiv \mathbf{1}(R \geq r_0)$ is binary and is a deterministic function of a possibly endogenous covariate R. Given $U = u$ and $R = r$, T is deterministic, i.e., $T = q_1(r, u)$ for $r \geq r_0$, and $T = q_0(r, u)$ for $r < r_0$. Causal identification with this control variable U is therefore local to the RD cutoff, which is a generic feature of the RD design. In contrast, Imbens and Newey (2009) focus on a continuous IV and aim to identify different causal objects than ours.

Lemma 1.2 provides identification of the reduced-form effect of the "IV" Z on Y, given $U = u$. It states that given $U = u$, the conditional mean change in the outcome at the RD threshold is causally related to the treatment change from $t_0(u)$ to $t_1(u)$. By the potential outcome notation,

$$\int (G(t_1(u), r_0, e) - G(t_0(u), r_0, e)) F_{\varepsilon|UR}(de, u, r_0) = \mathbb{E}\left[Y_{t_1(u)} - Y_{t_0(u)} | U = u, R = r_0\right].$$

It follows that $\lim_{r \to r_0^+} \mathbb{E}[Y|U=u, R=r] - \lim_{r \to r_0^-} \mathbb{E}[Y|U=u, R=r] =$
$\mathbb{E}\left[Y_{t_1(u)} - Y_{t_0(u)} | U = u, R = r_0\right]$. Based on Lemma 1.2, we can define the causal parameters of interest, Q-LATE and WQ-LATE.

Let $\mathcal{U} \equiv \{u \in [0,1]\text{: } |t_1(u) - t_0(u)| > 0\}$. For any $u \in \mathcal{U}$, define Q-LATE as

$$\begin{aligned} \tau(u) &\equiv \int \frac{G(t_1(u), r_0, e) - G(t_0(u), r_0, e)}{t_1(u) - t_0(u)} F_{\varepsilon|UR}(de, u, r_0) && (1) \\ &= \mathbb{E}\left[\frac{Y_{t_1(u)} - Y_{t_0(u)}}{t_1(u) - t_0(u)} | U = u, R = r_0\right], && (2) \end{aligned}$$

where $\frac{G(t_1(u),r_0,e)-G(t_0(u),r_0,e)}{t_1(u)-t_0(u)}$ is the (standardized) individual treatment effect.

$\frac{G(t_1(u),r_0,e)-G(t_0(u),r_0,e)}{t_1(u)-t_0(u)}$ is causal, because T switches from $t_0(u)$ to $t_1(u)$, while R and ε are held fixed. Q-LATE then captures an average causal effect for individuals with treatment rank $U = u$ at the RD threshold. The denominator in equation (2) reflects the fact that T is not a binary variable and that conditional on $U = u$ and $R = r_0$, there are two potential treatment values, $t_0(u)$ and $t_1(u)$. Analogous to the Wald formula, Q-LATE $\tau(u)$ is the ratio of the reduced-form effect of Z on Y to that of Z on T given $U = u$. For example, if the true model for Y given $U = u$ is $Y = b_0(u) + b_1(u) T + b_2(u) R + \varepsilon$, then $\tau(u) = b_1(u)$ for any $u \in \mathcal{U}$.

Q-LATE captures how treatment effects vary with treatment intensities of $t_0(u)$ and $t_1(u)$. For example, in our empirical application, Q-LATE reveals how increased bank capital affects bank outcomes at various levels of bank capital. In studying the returns to medical utilization around the low birth weight cutoff as in Almond et al. (2010), Q-LATE can be used to determine whether there are diminishing returns to medical spending. In exploring the effects of air pollution on life expectancy or mortality as in Chen et al. (2013), Ebenstein et al. (2017), and Fan et al. (2020), Q-LATE can be used to determine whether the effects of air pollution vary with pollution severity.

Further define the weighted average of Q-LATE, WQ-LATE, as

$$\pi(w) \equiv \int_{\mathcal{U}} \tau(u) w(u) du,$$

where $w(u)$ is a properly defined weighting function such that $w(u) \geq 0$ and $\int_{\mathcal{U}} w(u) du = 1$.

When the function $G(T, r, \varepsilon)$ is continuously differentiable in its first argument, both parameters can be expressed as weighted average derivatives of $Y = G(T, r, \varepsilon)$ with respect to T. In

particular, following Lemma 5 of Angrist, Graddy, and Imbens (2000),

$$\begin{aligned}\tau(u) &= \int \left(\int_{t_0(u)}^{t_1(u)} \frac{\partial}{\partial t} G(t, r_0, e)\, dt\right) (\Delta q(u))^{-1} F_{\varepsilon|UR}(de, u, r_0) \\ &= \mathbb{E}\left[\left(\int_{t_0(u)}^{t_1(u)} \frac{\partial}{\partial t} G(t, r_0, \varepsilon)\, dt\right) (\Delta q(u))^{-1} \middle| U = u, R = r_0\right] \\ &= \int_{t_0(u)}^{t_1(u)} \mathbb{E}\left[\frac{\partial}{\partial t} G(t, r_0, \varepsilon) \middle| U = u, R = r_0\right] (\Delta q(u))^{-1} dt,\end{aligned}$$

where $\Delta q(u) \equiv t_1(u) - t_0(u)$. Q-LATE $\tau(u)$ is a weighted average derivative averaging over the change in T at a given quantile u at the RD threshold. It follows that WQ-LATE $\pi(w)$ is also a weighted average derivative, averaging over both changes in T at a given quantile u and over $U \in \mathcal{U}$ at the RD threshold.

Define $q^+(u) \equiv \lim_{r \to r_0^+} q(r, u)$ and $q^-(u) \equiv \lim_{r \to r_0^-} q(r, u)$, where $q(r, u) \equiv q_0(r, u)(1 - Z) + q_1(r, u)Z$ is the conditional u quantile of T given $R = r$. These limits exist, as $q(r, u)$ is right and left continuous in r at $r = r_0$ given smoothness of $q_z(r, u)$ by Assumption 2. Let $m(t, r) \equiv \mathbb{E}[Y|T = t, R = r]$, and define $m^+(u) \equiv \lim_{r \to r_0^+} m(q^+(u), r)$ and $m^-(u) \equiv \lim_{r \to r_0^-} m(q^-(u), r)$. $q^\pm(u)$ and $m^\pm(u)$ can be consistently estimated from the data.

Theorem 1 (Identification)**.** *Under Assumptions 1–4, for any $u \in \mathcal{U}$, Q-LATE $\tau(u)$ is identified and is given by*

$$\tau(u) = \frac{m^+(u) - m^-(u)}{q^+(u) - q^-(u)}. \tag{3}$$

Further, WQ-LATE $\pi(w) \equiv \int_{\mathcal{U}} \tau(u)\, w(u)\, du$ is identified for any known or estimable weighting function $w(u)$ such that $w(u) \geq 0$ and $\int_{\mathcal{U}} w(u)\, du = 1$.

Note that in our setup, $q^+(u) = t_1(u)$ and $q^-(u) = t_0(u)$. In addition, U and T follow a one-to-one mapping, so we condition on $T = q^+(u)$ or $T = q^-(u)$ instead of $U = u$ in the numerator of equation (3).

To aggregate Q-LATE, one simple weighting function is equal weighting, i.e., $w(u) = 1/\int_{\mathcal{U}} 1 du$. One may choose other properly defined weighting functions. $w(u)$ is required to be non-negative; otherwise, when $w(u)$ is allowed to be negative, some weights will be greater than 1 and $\pi(w)$ will

be some weighted difference of the average treatment effects among those who change treatment levels at the RD threshold. The next section shows that the standard RD estimand can be expressed as a WQ-LATE, using a particular weighting function. In the special case when treatment effect is locally constant, the weighting function does not matter. With any valid weighting functions, one can identify the same homogenous treatment effect.

Remark 1 (Quantile effects)**.** *In addition to Q-LATE and WQ-LATE, one may identify potential outcome distributions and further local quantile treatment effects (LQTEs) at each* $u \in \mathcal{U}$*. In particular, under Assumptions 1–4,* $F_{Y_{t_1(u)}|UR}(y, u, r_0) = \lim_{r \to r_0^+} \mathbb{E}\left[\mathbf{1}(Y \leq y) | T = q^+(u), R = r\right]$*, and* $F_{Y_{t_0(u)}|UR}(y, u, r_0) = \lim_{r \to r_0^-} \mathbb{E}\left[\mathbf{1}(Y \leq y) | T = q^-(u), R = r\right]$*. When these potential outcome distributions are invertible, one can invert them to obtain LQTEs,* $F^{-1}_{Y_{t_1(u)}|UR}(\upsilon, u, r_0) - F^{-1}_{Y_{t_1(u)}|UR}(\upsilon, u, r_0)$*, for* $\upsilon \in (0, 1)$ *and* $u \in \mathcal{U}$*.*

Remark 2 (Covariates)**.** *Our basic setup assumes away other covariates other than the running variable. Rank invariance or similarity may be more plausible when conditioning on relevant covariates (see, e.g., discussion in Chernozhokov and Hansen, 2005). Let Assumptions 1 - 4 hold conditional on covariates. Our identification results then hold conditional on covariates. One caveat is that given some covariates* $\mathsf{X} = \mathsf{x}$*, Q-LATE at any treatment rank* $U(\mathsf{x}) = u(\mathsf{x})$ *is now covariate specific. One may average the Q-LATE over* $U(\mathsf{x})$ *to obtain the conditional WQ-LATE given* $\mathsf{X} = \mathsf{x}$*. One may further average the conditional WQ-LATE over the distribution of* X *at* $R{=}r_0$ *to obtain an unconditional WQ-LATE.*

3 Double-robust identification

In this section, we discuss the standard RD estimand and show that it can be expressed as a WQ-LATE, using a particular weighting function. We then discuss a double-robust identification approach and propose a causal estimand that incorporates the standard RD estimand as a special case. See, e.g., Arkhangelsky and Imbens (2021) for a double-robust approach to causal effects in panel data models.

3.1 Standard RD estimand

Consider the standard RD estimand in the form of the standard local Wald ratio, and rewrite it as follows

$$\begin{aligned}
\pi^{RD} &\equiv \frac{\lim_{r\to r_0^+}\mathbb{E}[Y|R=r]-\lim_{r\to r_0^-}\mathbb{E}[Y|R=r]}{\lim_{r\to r_0^+}\mathbb{E}[T|R=r]-\lim_{r\to r_0^-}\mathbb{E}[T|R=r]} \\
&= \frac{\int_0^1\left(\lim_{r\to r_0^+}\mathbb{E}[Y|U=u,R=r]-\lim_{r\to r_0^-}\mathbb{E}[Y|U=u,R=r]\right)du}{\int_0^1\left(\lim_{r\to r_0^+}q(r,u)-\lim_{r\to r_0^-}q(r,u)\right)du} \\
&= \int_{\mathcal{U}}\tau(u)\frac{\Delta q(u)}{\int_{\mathcal{U}}\Delta q(u)\,du}du, \qquad (4)
\end{aligned}$$

where the first equality follows from $T = q(R,U)$ and interchanging limit and integral, which is allowed under our assumptions, and the second equality follows from Lemma 1.2 and the fact that $t_1(u) = \lim_{r\to r_0^+} q(r,u)$ and $t_0(u) = \lim_{r\to r_0^-} q(r,u)$. Therefore, under our assumptions, the standard RD estimand identifies a weighted average of Q-LATEs, using weights $w^{RD}(u) \equiv \Delta q(u) / \int_{\mathcal{U}} \Delta q(u)\,du$.

To ensure $w^{RD}(u) \geq 0$ over $\mathcal{U}$, it is necessary that $\Delta q(u) \geq 0$ or $\Delta q(u) \leq 0$ for all $u \in \mathcal{U}$. Otherwise, when $\Delta q(u)$ can switch signs, π^{RD} would be undefined if the denominator $\int_{\mathcal{U}} \Delta q(u)\,du = 0$, and π^{RD} would be a weighted difference of the average treatment effects for units with positive treatment changes and those with negative treatment changes if $\int_{\mathcal{U}} \Delta q(u)\,du \neq 0$.

Assumption 3b (Monotonicity)**.** $\Pr(T_1 - T_0 \geq 0|R=r_0) = 1$ *or* $\Pr(T_1 - T_0 \leq 0|R=r_0) = 1$.

Assumption 3b requires that treatment T is weakly increasing or weakly decreasing almost surely when crossing the RD threshold. Assumption 3b implies that $\Delta q(U) \geq 0$ or $\Delta q(U) \leq 0$ holds almost surely.

Unlike Assumption 3, which imposes rank restrictions, Assumption 3b imposes a sign restriction on the treatment changes at the RD threshold. Angrist, Graddy, and Imbens (2000) make a similar assumption in identifying a general simultaneous equations system with binary IVs.

When Assumption 3 local treatment rank invariance or similarity does not hold, Q-LATE involved in equation (4) does not have a causal interpretation. However, the RD estimand can still identify a causal parameter under Assumption 3b monotonicity. We formally state this result in the following Lemma 2.

Lemma 2. *Let Assumptions 1, 2, 3b, and 4 hold. Then* π^{RD} *identifies a weighted average effect of* T *on* Y *at* $R = r_0$.

The exact form of the weighted average effect is provided in the proof of Lemma 2 in the Appendix. We show that in this case, the standard RD estimand with a continuous treatment identifies a weighted average of individual treatment effects among those individuals who change their treatment intensity at the RD threshold, i.e., those having $t_1(u_1) - t_0(u_0) > 0$ (or $t_1(u_1) - t_0(u_0) < 0$). The individual treatment effect is given by $\frac{G(t_1(u_1),r_0,e)-G(t_0(u_0),r_0,e)}{t_1(u_1)-t_0(u_0)}$, and the weight is proportional to the individual's treatment change, $t_1(u_1) - t_0(u_0)$. When further $G(T, R, \varepsilon)$ is continuously differentiable in T, the identified effect can be expressed as a weighted average derivative of Y w.r.t. T, as shown in the proof of Lemma 2.

3.2 Double-robust identification

The discussion so far suggests that the standard RD estimand in general requires Assumption 3b monotonicity in order to be causal. Note that the monotonicity and rank assumptions impose different restrictions on the first-stage heterogeneity. Monotonicity imposes a sign restriction on $T_1 - T_0$ at $R = r_0$, while the rank assumption imposes a rank restriction on T_1 and T_0 at $R = r_0$. Neither assumption implies the other. It is therefore useful to have an estimand that is valid under either assumption. Note that the common empirical practice of focusing on some sub-population for which researchers believe the treatment is more affected still requires either monotonicity or rank similarity to hold for such sub-population.

Theorem 2 (Double-Robust Identification)**.** *Let Assumptions 1, 2 and 4 hold. Then under either*

Assumption 3 or 3b,

$$\pi^* \equiv \int_{\mathcal{U}} \frac{m^+(u) - m^-(u)}{q^+(u) - q^-(u)} \frac{|q^+(u) - q^-(u)|}{\int_{\mathcal{U}} |q^+(u) - q^-(u)| du} du \tag{5}$$

identifies a weighted average effect of T on Y at $R = r_0$.

Theorem 2 provides a causal estimand that is valid under either the monotonicity or rank assumption. When monotonicity holds, $\pi^* = \pi^{RD}$. When the rank assumption holds, $\pi^* = \pi(w^*) \equiv \int_{\mathcal{U}} \tau(u) w^*(u) du$ for $w^*(u) \equiv \frac{|\Delta q(u)|}{\int_{\mathcal{U}} |\Delta q(u)| du}$, i.e., π^* identifies a WQ-LATE. Either way, π^* identifies a weighted average of individual treatment effects given by $\frac{G(t_1(u_1), r_0, e) - G(t_0(u_0), r_0, e)}{t_1(u_1) - t_0(u_0)}$ among those individuals who change their treatment intensities at the RD threshold.

The two alternative assumptions put different restrictions on how individuals can change treatment intensities when crossing the RD threshold. Monotonicity requires that U_0 and U_1 are such that $t_1(U_1) - t_0(U_0) \geq 0$ or $t_1(U_1) - t_0(U_0) \leq 0$ almost surely, i.e., individuals change treatment in one direction when crossing the RD threshold. In contrast, the rank assumption requires that given ε, U_0 and U_1 have the same conditional distribution at $R = r_0$, i.e., the probability for an individual to stay at a certain rank of the treatment distribution stays the same when crossing the RD threshold. Our estimand π^* provides a robust way to aggregate the individual treatment effects.

4 Estimation

The proposed estimands for Q-LATE and WQ-LATE involve conditional means and quantiles at a boundary point. Following the standard practice of the RD literature, we estimate Q-LATE and WQ-LATE by local linear mean and quantile regressions.

For simplicity, we use the same kernel function $K(\cdot)$ for all estimation. Let the bandwidths for T and R be h_T and h_R, respectively. The bandwidth sequences h_R and h_T go to zero as the sample size $n \to \infty$. Denote as $\widehat{\theta}$ the estimate of any parameter θ. Given a sample of n $i.i.d.$ observations $\{(Y_i, T_i, R_i)\}_{i=1}^n$ from (Y, T, R), we estimate Q-LATE $\tau(u)$ and WQ-LATE π^* by the following procedure.

Step 1: Let $\mathsf{U}^{(l)} \equiv \{u_1, u_2, ..., u_l\}$ be the set of equally spaced quantiles over the unit interval $(0, 1)$. For $u \in \mathsf{U}^{(l)}$, estimate $q^+(u)$ by $\widehat{q}^+(u) \equiv \widehat{a}_0$ from the local linear quantile regression

$$(\widehat{a}_0, \widehat{a}_1) = \arg\min_{a_0, a_1} \sum_{\{i:\ R_i \geq r_0\}} K\left(\frac{R_i - r_0}{h_R}\right) \rho_u \left(T_i - a_0 - a_1 \left(R_i - r_0\right)\right),$$

where $\rho_u(\alpha) = \alpha(u - \mathbf{1}(\alpha < 0))$ is the standard check function. Estimate $q^-(u)$ similarly using observations below r_0.

Step 2: Let $\widetilde{\mathcal{U}} \equiv \left\{u \in \mathsf{U}^{(l)} : |\Delta\hat{q}(u)| > \epsilon_n\right\}$, where $\Delta\hat{q}(u) \equiv \hat{q}^+(u) - \hat{q}^-(u)$ and the trimming parameter $\epsilon_n \to 0$ is a positive sequence satisfying the conditions in Lemma 6 in Appendix B. For all $u \in \widetilde{\mathcal{U}}$, estimate $m^+(u)$ by $\widehat{m}^+(u) \equiv \widehat{b}_0$ from the local linear regression

$$\begin{aligned}\left(\widehat{b}_0, \widehat{b}_1, \widehat{b}_2\right) \ = \ & \arg\min_{b_0, b_1, b_2} \sum_{\{i:\ R_i \geq r_0\}} K\left(\frac{R_i - r_0}{h_R}\right) K\left(\frac{T_i - \hat{q}^+(u)}{h_T}\right) \\ & \times \left(Y_i - b_0 - b_1\left(R_i - r_0\right) - b_2\left(T_i - \hat{q}^+(u)\right)\right)^2.\end{aligned}$$

Estimate $m^-(u)$ similarly by replacing $\hat{q}^+(u)$ with $\hat{q}^-(u)$ and using observations below r_0.

Step 3: Estimate $\tau(u)$ by the plug-in estimator $\widehat{\tau}(u) = \frac{\hat{m}^+(u) - \hat{m}^-(u)}{\hat{q}^+(u) - \hat{q}^-(u)}$ for $u \in \widetilde{\mathcal{U}}$.

Step 4: Estimate π^* by $\hat{\pi}^* = \sum_{u \in \widetilde{\mathcal{U}}} \widehat{\tau}(u) \frac{|\Delta\hat{q}(u)|}{\sum_{u \in \widetilde{\mathcal{U}}} |\Delta\hat{q}(u)|}$.

Our identification theory requires trimming out treatment quantiles where there are no changes at the RD threshold, i.e., $\Delta q(u) = 0$, whereas in practice we do not know the true $\Delta q(u)$. To avoid any pre-testing problems, we trim out all quantiles such that $|\Delta\hat{q}(u)| \leq \epsilon_n$ for some chosen ϵ_n. Lemma 6 in Appendix B shows that when ϵ_n satisfies the required conditions, this trimming procedure is asymptotically equivalent to trimming out those treatment quantiles where the true changes are zero and hence preserves the asymptotic properties of our estimator. If one wishes to focus on quantiles such that $|\Delta q(u)| > c_0$ for some small $c_0 > 0$, then the trimming parameter can be defined as $c_n = c_0 + \epsilon_n$.

In practice, one can choose $\epsilon_n = \max_{u \in \mathsf{U}^{(l)}} se\left(\Delta\tilde{q}(u)\right) \times 1.96$, where $\Delta\tilde{q}(u)$ is a preliminary Step 1 estimator of the treatment quantile change, using the bandwidth $\tilde{h}_R$ such that $\tilde{h}_R / h_R \to 0$

and $n\tilde{h}_R^2/h_R \to \infty$. We discuss the choice of h_R in Section 5. The associated standard errors satisfy $se(\Delta\tilde{q}(u)) = O_p\big((n\tilde{h}_R)^{-1/2}\big) > se(\Delta\hat{q}(u)) = O_p\big((nh_R)^{-1/2}\big)$. By this procedure, insignificant estimates (at the 5% significance level) of $\Delta\hat{q}(u)$ along with some significant but small estimates will be trimmed out. Since $\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)| - |\Delta q(u)|\big| = O_p((nh_R)^{-1/2})$, the conditions for ϵ_n given in Lemma 6 are satisfied. Consider specifically the bandwidth sequences $h_R = cn^{-a}$ and $\tilde{h}_R = cn^{-b}$ for some constants $0 < a, b < 1$ and $c > 0$. The required conditions for ϵ_n are satisfied when choosing b such that $a < b < (a+1)/2$.

Recall that our identifying assumptions imply a testable condition $\lim_{r\to r_0^+} F_{X|UR}(x,u,r) - \lim_{r\to r_0^-} F_{X|UR}(x,u,r) = 0$ for some observable covariate X. This suggests that one may test that Q-LATEs or WQ-LATEs on the covariate distribution are zero. In practice, one can use $\mathbf{1}(X \le x)$ as an outcome and follow the above estimation procedure to perform falsification tests. Standard multiple testing adjustments may be applied if needed. Any false significant effects on the covariate distribution would cast doubt on the validity of the identifying assumptions.

More formally, one may follow the idea of the RD distributional tests of Shen and Zhang (2016) to implement a Kolmogorov-Smirnov type of test. Such a test compares the estimated conditional distributions $\lim_{r\to r_0^+} F_{X|UR}(x,u,r)$ and $\lim_{r\to r_0^-} F_{X|UR}(x,u,r)$. Developing a full-blown test is beyond the scope of the current paper and is left for future research.

5 Inference

The proposed estimators have several distinct features, which make analyzing their asymptotic properties challenging. First, the local polynomial estimator in Step 2 involves a continuous treatment variable T, in addition to the running variable R. Evaluating T over its interior support and evaluating R at the boundary point r_0 complicates the analysis. Second, we need to account for the sampling variation of $\hat{q}^{\pm}(u)$ from Step 1, which appears in both the numerator and denominator of $\widehat{\tau}(u)$, as well as in the weighting function $\widehat{\omega}^*(u)$ for $\hat{\pi}^*$. Third, our estimation involves a trimming procedure that is based on the estimated quantile change $\Delta\hat{q}(u)$. We overcome these complica-

tions by extending the results of Kong, Linton, and Xia (2010) and Qu and Yoon (2015). Qu and Yoon (2015) provide uniform convergence results for local linear quantile regressions, while Kong, Linton, and Xia (2010) establish uniform convergence results for local polynomial estimators.

To establish our inference procedure, we derive the asymptotically linear representation and asymptotic normality of the estimators $\hat{\tau}(u)$ and $\hat{\pi}^*$. We show that, similar to the standard RD local polynomial estimator, the large sample distributional approximations involve leading biases, which depend on changes in the curvatures of the conditional quantile and mean functions in Step 1 and Step 2 of estimation. There are two common approaches to removing these leading biases, undersmoothing and bias correction. The undersmoothing approach uses a bandwidth sequence that goes to zero fast enough with the sample size, so that the bias is asymptotically negligible relative to the standard error. Nevertheless it is known that this undersmoothing approach prevents a lot of bandwidth choices used in practice. This section focuses on the bias correction approach. Undersmoothing results are presented in Appendix B.2.

We develop robust inference for our bias-corrected estimators, similar to the robust bias-corrected inference of Calonico, Cattaneo, and Titiunik (2014) in the context of the standard RD design. Calonico, Cattaneo, and Farrell (2018, 2019, 2020) further formally establish higher-order improvements of such an approach. Our robust inference takes into account the added variability due to the bias correction in deriving large sample distributions. We also present the optimal bandwidths for both the Q-LATE and WQ-LATE estimators by minimizing the asymptotic mean squared error (AMSE). The robust confidence intervals for the bias-corrected estimators deliver valid inference when these AMSE optimal bandwidths are used.

We impose the following assumptions for asymptotics.

Assumption 5 (Asymptotics)**.** *1. For any $t \in \mathcal{T}_z$, $z = 0, 1$, $r \in \mathcal{R}$, and $u \in \mathcal{U}$, $f_{T_z R}(t, r)$ is bounded and bounded away from zero, and has bounded first order derivatives with respect to (t, r); $\partial^j q_z(r, u)/\partial r^j$ is finite and Lipschitz continuous over (r, u) for $j = 1, 2, 3$; $q_z(r_0, u)$ and $\partial q_z(r_0, u)/\partial u$ are finite and Lipschitz continuous in u.*

2. For any $t \in \mathcal{T}_z$, $z = 0, 1$, and $r \in \mathcal{R}$, $\mathbb{E}\left[G(T_z, R, \varepsilon)|T_z = t, R = r\right]$ has bounded fourth

order derivatives; the conditional variance $\mathbb{V}\big[G(T_z, R, \varepsilon)|T_z = t, R = r\big]$ *is continuous and bounded away from zero; the conditional density* $f_{T_z R|Y}(t, r, y)$ *is bounded for any* $y \in \mathcal{Y}$. $\mathbb{E}\big[\big|Y - \mathbb{E}\big[Y|T_z, R\big]\big|^3\big] < \infty$ *for* $z = 0, 1$.

3. *The kernel function* K *is bounded, positive, compactly supported, symmetric, having finite first-order derivative, and satisfying* $\int_{-\infty}^{\infty} \upsilon^2 K(\upsilon) d\upsilon > 0$.

Assumption 5.1 imposes sufficient smoothness conditions to derive the asymptotically linear representations of $\hat{q}^{\pm}(u)$. In particular, the bounded joint density implies a compact support where the stochastic expansions of $\hat{q}^{\pm}(u)$ hold uniformly over u. Together with the smoothness conditions on $q_z(r, u)$, the remainder terms in the stochastic expansions are controlled to be small. Assumption 5.2 imposes additional conditions to derive the asymptotically linear representation of $\hat{\mathbb{E}}[Y|T, R]$ and asymptotic normality of our estimators. Assumption 5.3 provides the standard regularity conditions for the kernel function.

The asymptotically linear representations and asymptotic normality of the main estimators $\hat{\tau}(u)$ and $\hat{\pi}^*$ are presented in Appendix B, followed by the inference theory based on undersmoothing. In the following sections 5.1 and 5.2, we present the robust bias-corrected inference for Q-LATE $\tau(u)$ and WQ-LATE π^*, respectively.

5.1 Inference on Q-LATE

Denote the leading bias for $\hat{\tau}(u)$ as $h_R^2 \mathsf{B}_{R\tau}(u) + h_T^2 \mathsf{B}_{T\tau}(u)$. The exact forms of $\mathsf{B}_{R\tau}(u)$ and $\mathsf{B}_{T\tau}(u)$ are presented in equations (B.8) and (B.9) in Appendix B, respectively. We propose the bias-corrected estimator for $\tau(u)$

$$\hat{\tau}^{bc}(u) \equiv \hat{\tau}(u) - \left(h_R^2 \widehat{\mathsf{B}}_{R\tau}(u) + h_T^2 \widehat{\mathsf{B}}_{T\tau}(u)\right),$$

where $\widehat{\mathsf{B}}_{R\tau}(u)$ and $\widehat{\mathsf{B}}_{T\tau}(u)$ are consistent estimators for $\mathsf{B}_{R\tau}(u)$ and $\mathsf{B}_{T\tau}(u)$, respectively.

Bias correction reduces biases, but also introduces variability. When the added variability of the estimated bias is not accounted for, the empirical coverage of the resulting confidence inter-

val can be well below their nominal target, which implies that conventional confidence intervals may substantially over-reject the null hypothesis of no treatment effect. We therefore present the asymptotic distributions of the bias-corrected estimators $\hat{\tau}^{bc}(u)$, taking into account the sampling variation induced by bias correction.

Theorem 3 (Asymptotic distribution of $\hat{\tau}^{bc}(u)$)**.** *Let Assumptions 1-5 hold. Let the bandwidths for $\hat{\tau}(u)$ be $h_R = c_R h$, $h_T = c_T h$, the bandwidths used for the bias estimation be $b_R = c_R b$ and $b_T = c_T b$, for some positive constants c_R, c_T, and positive sequences $h = h_n \to 0$ and $b = b_n \to 0$. If $h/b \to \rho \in [0, \infty]$, $n \min\{h^6, b^6\} \max\{h^2, b^2\} \to 0$, $n \min\{h^2, b^6 h^{-4}\} \to \infty$, and $nh^3 \max\{1, h^6/b^6\} \to \infty$, then for any $u \in \mathcal{U}$,*

$$\frac{\hat{\tau}^{bc}(u) - \tau(u)}{\sqrt{\mathsf{V}^{bc}_{\tau,n}(u)}} \xrightarrow{d} \mathcal{N}(0,1), \text{ where } \mathsf{V}^{bc}_{\tau,n}(u) \equiv \left(\frac{\mathsf{V}_\tau(u)}{nh^2} + \frac{\mathsf{V}_{\mathsf{B}_\tau}(u)}{nb^6 h^{-4}} + \frac{\mathsf{C}_\tau(u;\rho)}{nhb} \right) \frac{1}{c_R c_T}.$$

The exact forms of $\mathsf{V}_\tau(u)$, $\mathsf{V}_{\mathsf{B}_\tau}(u)$ and $\mathsf{C}_\tau(u;\rho)$ are given in equations (B.1), (B.2), and (B.3) in Appendix B, respectively.

The variance $\mathsf{V}^{bc}_{\tau,n}(u)$ consists of three terms: $\mathsf{V}_\tau(u)$ is from the variance of the actual estimator $\hat{\tau}(u)$, $\mathsf{V}_{\mathsf{B}_\tau}(u)$ is from the variance of the bias estimator $h_R^2 \widehat{\mathsf{B}}_{R\tau} + h_T^2 \widehat{\mathsf{B}}_{T\tau}$, and $\mathsf{C}_\tau(u;\rho)$ is from the covariance between $\hat{\tau}(u)$ and $h_R^2 \widehat{\mathsf{B}}_{R\tau} + h_T^2 \widehat{\mathsf{B}}_{T\tau}$. Theorem 3 incorporates three limiting cases depending on ρ, the limiting value of h/b. When $h/b \to 0$, $\hat{\tau}(u)$ is first-order and the bias estimator is of smaller order. Thus the variance reduces to $\mathsf{V}^{bc}_{\tau,n}(u) = \mathsf{V}_\tau(u)/(nh^2 c_R c_T)$. When $h/b \to \rho \in (0, \infty)$, both $\hat{\tau}(u)$ and the bias estimator contribute to the asymptotic variance. For example, when $\rho = 1$, $\mathsf{V}^{bc}_{\tau,n}(u) = (\mathsf{V}_\tau(u) + \mathsf{V}_{\mathsf{B}_\tau}(u) + \mathsf{C}_\tau(u;1))/(nh^2 c_R c_T)$. When $h/b \to \infty$, the bias estimator is first-order and $\hat{\tau}(u)$ is of smaller order, so $\mathsf{V}^{bc}_{\tau,n}(u) = \mathsf{V}_{\mathsf{B}_\tau}(u)/(nb^6 h^{-4} c_R c_T)$.

Without loss of generality, we assume that the bandwidths $h_R = c_R h$ and $h_T = c_T h$ are of the same order. We show in Lemma 4 in Appendix B that h_R and h_T have the same first-order impact on $\hat{\tau}(u)$. This is because the local linear estimator of $\mathbb{E}[Y|T, R]$ in Step 2 dominates the first-order asymptotically linear representation, and the quantile regression of T on R in Step 1 is of smaller order. In addition, we derive the optimal bandwidths that minimize the AMSE of $\hat{\tau}(u)$ in Theorem

4 below. The resulting AMSE optimal bandwidths are of the same order $n^{-1/6}$.

Theorem 4 (AMSE optimal bandwidth for $\hat{\tau}(u)$)**.** *Let Assumptions 1-5 hold. If* $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_R h_T^2 \to \infty$, $nh_T h_R^5 \to c \in [0, \infty)$, $nh_R h_T^5 \to c \in [0, \infty)$, *and* $h_R^2 / h_T \to 0$, *then the mean squared error of* $\hat{\tau}(u)$ *is* $\mathbb{E}\left[\left(\hat{\tau}(u) - \tau(u)\right)^2\right] = \left(h_R^2 \mathsf{B}_{R\tau}(u) + h_T^2 \mathsf{B}_{T\tau}(u)\right)^2 + (nh_R h_T)^{-1} \mathsf{V}_\tau(u) + o\left(h_R^4 + h_T^4 + (nh_R h_T)^{-1}\right)$*; further if* $\mathsf{B}_{R\tau}(u) \neq 0$ *and* $\mathsf{B}_{T\tau}(u) \neq 0$, *the bandwidths that minimize the AMSE are* $h_{R\tau}^*(u) = c_R^*(u) n^{-1/6}$ *and* $h_{T\tau}^*(u) = c_T^*(u) n^{-1/6}$, *where* $c_R^*(u) = (\mathsf{V}_\tau(u)/8)^{1/6} (\mathsf{B}_{T\tau}(u)/\mathsf{B}_{R\tau}^5(u))^{1/12}$ *and* $c_T^*(u) = (\mathsf{V}_\tau(u)/8)^{1/6} (\mathsf{B}_{R\tau}(u)/\mathsf{B}_{T\tau}^5(u))^{1/12}$.

The AMSE optimal bandwidths for $\hat{\tau}(u)$ satisfy the bandwidth conditions specified in Theorem 3. Therefore one can apply the above AMSE optimal bandwidths and then conduct the bias-corrected robust inference provided in Theorem 3.

The biases, robust variances, and the AMSE optimal bandwidths can be consistently estimated by plug-in estimators. The biases and variances depend on the second order derivatives of $q^{\pm}(u)$ and $m^{\pm}(u)$, the conditional variance of Y given (T, R), the density f_{TR}, and some constants determined by the kernel function. These involved parameters can be estimated by local quadratic quantile and mean regressions as well as kernel density estimators. Details of the plug-in estimators are provided in Appendix C.

The terms due to the bias correction, $V_{\mathsf{B}_\tau}(u)$ and $C_\tau(u; \rho)$, depend on $V_\tau(u)$ and some kernel-specific constants. As a result, $V_{\tau,n}^{bc}(u)$ only depends on $V_\tau(u)$ and some constants, which implies that estimating the robust variance is not computationally more demanding than estimating the conventional variance $V_\tau(u)$ without the bias correction. For example, for the Uniform kernel and $\rho = 1$, $V_{\tau,n}^{bc}(u) = 13.89 V_\tau(u)/(nh^2)$. Imbens and Kalyanaraman (2012) and Arai and Ichimura (2018) also use similar kernel-specific constants.

5.2 Inference on WQ-LATE

Denote the leading bias for $\hat{\pi}^*$ as $h_R^2 \mathsf{B}_{R\pi} + h_T^2 \mathsf{B}_{T\pi}$. The exact forms of $\mathsf{B}_{R\pi}$ and $\mathsf{B}_{T\pi}$ are given in equations (B.11) and (B.12) in Appendix B, respectively. We propose the bias-corrected estimator

for π^*

$$\hat{\pi}^{bc} \equiv \hat{\pi}^* - \left(h_R^2 \widehat{\mathsf{B}}_{R\pi} + h_T^2 \widehat{\mathsf{B}}_{T\pi}\right),$$

where $\widehat{\mathsf{B}}_{R\pi}$ and $\widehat{\mathsf{B}}_{T\pi}$ are consistent estimators of $\mathsf{B}_{R\pi}$ and $\mathsf{B}_{T\pi}$, respectively.

Theorem 5 (Asymptotic distribution of $\hat{\pi}^{bc}$)**.** *Let Assumptions 1, 2, either 3 or 3b, 4 and 5 hold and* $l^{-1}\sqrt{nh_R} \to 0$. *Let the bandwidths for* $\hat{\pi}^*$ *be* $h_R = c_R h$, $h_T = c_T h$, *the bandwidths used for the bias estimation be* $b_R = c_R b$ *and* $b_T = c_T b$, *for some positive constants* c_R, c_T, *and positive sequences* $h = h_n \to 0$ *and* $b = b_n \to 0$. *If* $h/b \to \rho \in [0, \infty]$, $n \min\{h^5, b^5\} \max\{h^2, b^2\} \to 0$, $n \min\{h, b^5 h^{-4}\} \to \infty$, *and* $nh^4 \max\{1, h^5 b^{-5}\} \to \infty$, *then*

$$\frac{\hat{\pi}^{bc} - \pi^*}{\sqrt{\mathsf{V}_{\pi,n}^{bc}}} \xrightarrow{d} \mathcal{N}(0,1), \text{ where } \mathsf{V}_{\pi,n}^{bc} \equiv \left(\frac{\mathsf{V}_\pi}{nh} + \frac{\mathsf{V}_{\mathsf{B}_\pi}}{nb^5 h^{-4}} + \frac{\mathsf{C}_\pi(\rho)}{nb^2 h^{-1}}\right)\frac{1}{c_R}. \tag{6}$$

The exact forms of V_π, $\mathsf{V}_{\mathsf{B}_\pi}$, *and* $\mathsf{C}_\pi(\rho)$ *are given in equations (B.4), (B.6), and (B.7) in Appendix B, respectively.*

Instead of letting c_T *be a constant, suppose* $h_T = c_T h$ *where* $c_T = c_{Tn}$ *is a positive sequence satisfying* $c_{Tn} \to 0$ *and* $hc_{Tn}^{-3} \to 0$. *Equation (6) still holds.*

$\mathsf{V}_{\pi,n}^{bc}$ consists of three terms: V_π is from the variance of the actual estimator $\hat{\pi}^*$, $\mathsf{V}_{\mathsf{B}_\pi}$ is from the variance of the bias estimator $h_R^2 \widehat{\mathsf{B}}_{R\pi} + h_T^2 \widehat{\mathsf{B}}_{T\pi}$, and $\mathsf{C}_\pi(\rho)$ is from the covariance between $\hat{\pi}^*$ and $h_R^2 \widehat{\mathsf{B}}_{R\pi} + h_T^2 \widehat{\mathsf{B}}_{T\pi}$. Similar to Theorem 3, Theorem 5 incorporates three limiting cases depending on ρ. When $h/b \to \rho = 0$, $\hat{\pi}^*$ is first-order and the bias estimator is of smaller order. Then $\mathsf{V}_{\pi,n}^{bc} \equiv \mathsf{V}_\pi / (nhc_R)$. When $h/b \to \rho \in (0, \infty)$, both $\hat{\pi}^*$ and the bias estimator contribute to the asymptotic variance. When $h/b \to \infty$, the bias estimator is first-order and $\hat{\pi}^*$ is of smaller order. Then $\mathsf{V}_{\pi,n}^{bc} \equiv \mathsf{V}_{\mathsf{B}_\pi} / (nb^5 h^{-4} c_R)$.

Note that Q-LATE $\tau(u)$ is a function of T and R, while WQ-LATE π^* is a weighted average of $\tau(u)$ averaging over T and hence is only a function of R. The asymptotic theory for $\hat{\pi}^*$ in Lemma 5 of Appendix B shows that the leading variance is of order $1/\sqrt{nh_R}$. In theory, one can choose a small bandwidth for T, in particular $h_T = c_{Tn} h$ for $c_{Tn} \to 0$, such that the leading

bias associated with h_T, $h_T^2 \mathsf{B}_{T\pi}$, becomes first-order ignorable compared with the leading bias associated with h_R, $h_R^2 \mathsf{B}_{R\pi}$. The leading bias of $\hat{\pi}^*$ can then be simplified to $h_R^2 \mathsf{B}_{R\pi}$. It follows that the first-order asymptotic property of $\hat{\pi}^*$ will not depend on h_T. These are features of the general marginal integration or partial mean of the nonparametrically estimated conditional mean function (see, e.g., Newey, 1994). Nevertheless, $h_T^2 \mathsf{B}_{T\pi}$ might not be ignorable in finite samples. The finite-sample performance of the bias-corrected estimator could be compromised, if the bias term associated with h_T was ignored. Our bias-corrected estimator $\hat{\pi}^{bc}$ and the associated robust inference therefore take into account $h_T^2 \widehat{\mathsf{B}}_{T\pi}$.

The following Theorem presents the optimal bandwidth that minimizes the AMSE of $\hat{\pi}^*$.

Theorem 6 (AMSE optimal bandwidth for $\hat{\pi}^*$)**.** *Let Assumptions 1, 2, either 3 or 3b, 4 and 5 hold and* $l^{-1}\sqrt{nh_R} \to 0$. *If* $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_R h_T^3 \to \infty$, $nh_R^5 \to c \in [0, \infty)$, *and* $nh_R h_T^4 \to c \in [0, \infty)$, *then the mean squared error of* $\hat{\pi}^*$ *is* $\mathbb{E}\left[\left(\hat{\pi}^* - \pi^*\right)^2\right] = h_R^4 \mathsf{B}_{R\pi}^2 + h_T^4 \mathsf{B}_{T\pi}^2 + (nh_R)^{-1} \mathsf{V}_\pi + o\left(h_R^4 + h_T^4 + (nh_R)^{-1}\right)$*; further if* $nh_R h_T^4 \to 0$, $\mathsf{B}_{R\pi} \neq 0$, *and* $\mathsf{B}_{T\pi} \neq 0$, *then the bandwidth that minimizes the AMSE is* $h_{R\pi}^* = \left(\mathsf{V}_\pi / \left(4\mathsf{B}_{R\pi}^2\right)\right)^{1/5} n^{-1/5}$.

The optimal bandwidth $h_{R\pi}^*$ is derived under the scenario that the leading bias associated with h_T, $h_T^2 \mathsf{B}_{T\pi}$, is first-order asymptotically ignorable. Following Horowitz (2001), we suggest a rule-of-thumb bandwidth for h_T. In particular, $h_{T\pi}^{rot} = h_{R\pi}^* n^{-1/30} \sigma_T / \sigma_R$, where σ_R and σ_T are the standard deviations of R and T, respectively. This rule-of-thumb bandwidth satisfies the conditions $nh_{R\pi}^* h_T^3 \to \infty$ and $nh_{R\pi}^* h_T^4 \to 0$ in Theorem 6. We can use $h_{R\pi}^*$ and $h_{T\pi}^{rot}$ to conduct bias-corrected robust inference provided in Theorem 5.

Remark 3. *Lemmas 4 and 5 in Appendix B present the asymptotically linear representations of* $\hat{\tau}(u)$ *and* $\hat{\pi}^*$, *respectively. We compute the asymptotic unconditional MSE, as in Imbens and Kalyanaraman (2012). In contrast, Calonico, Cattaneo, and Titiunik (2014), Arai and Ichimura (2018), and Calonico, Cattaneo, Farrell, and Titiunik (2019) derive the asymptotic conditional MSE given the sample data. In large samples, these two approaches, approximating the unconditional or conditional MSE, are equivalent. In finite samples, the resulting confidence interval*

based on the conditional variance can be larger or smaller than the confidence interval based on the unconditional variance.

The unconditional MSE simplifies the asymptotic analysis for our multi-step estimators. Note that the Q-LATE estimator $\widehat{\tau}(u)$ *involves two continuous regressors* R *and* T*. In contrast, the standard RD estimator for a binary treatment has only one continuous regressor* R*. Based on the asymptotically linear representation of* $\hat{\tau}(u)$*, the leading unconditional bias is a linear function of the unconditional biases of Step 1 quantile regression and Step 2 mean regression. It follows that the leading unconditional bias of* $\hat{\pi}^*$ *is also a simple linear function of the biases of* $\hat{q}^{\pm}(u)$ *and* $\hat{m}^{\pm}(u)$*.*

Remark 4. *Calonico, Cattaneo, Farrell, and Titiunik (2019) show that inclusion of covariates in the standard RD design can increase efficiency. Intuitively, the efficiency gain may carry over to our WQ-LATE estimator if the covariate adjustment is made additively in a linear-in-parameters way. A full theoretical development can be interesting for future research.*

6 Empirical analysis

This section applies the proposed approach to quantify the impacts of bank capital on banks' short-run responses and long-run failure probabilities. Are banks less likely to fail when they hold more capital? Answering this question can shed light on the role of higher capital in promoting a stable financial system. The minimum capital requirement in the early 20th century United States provides a unique quasi-experiment that allows one to nonparametrically identify the true causal impacts of bank capital. Back then, bank runs and banking panics were prevalent. The minimum capital requirement was set in place to prevent bank from holding too little capital and to thereby promote banking stability.

As shown in Figure 1, the requirement depends on town sizes and changes abruptly at the town population threshold 3,000. The required minimum capital is $25,000 for a bank located in a town with a population less than 3,000, and jumps to $50,000 for a bank located in a town with

a population at or above 3,000. There are two other population thresholds, 6,000 and 50,000, at which the minimum capital requirement changes. Our empirical analysis focuses on the population threshold 3,000, since about 88% of banks in our sample are located in towns with a population below 6,000.

Let the continuous treatment T be bank capital, and the running variable R be town population. Further let Z indicate whether a bank is located in a town with 3,000 or more people. We consider three outcomes of interest (Y): total assets, leverage, and an indicator of whether a bank suspended its operation in the following 24 years. Leverage is defined as the ratio of a bank's total assets to capital, which is a measure of the amount of risk a bank engages in. Logged values are used for bank capital, assets and leverage, as these variables have rather skewed distributions. We estimate the impacts of the minimum capital requirement on the distribution of bank capital (i.e., the first stage impact of Z on T), and further the impacts of higher capital on the three outcomes of interest (i.e., the impacts of T on Y). We also quantify any possible treatment effect heterogeneity at various levels of bank capital.

Our data come from three sources: the annual reports of the Office of the Comptroller of the Currency (OCC), Rand McNally's Bankers Directory, and the United States population census. Our full estimation sample consists of 822 banks in 45 towns, among which 717 are below the relevant policy threshold and 105 are above. In addition to T, Y, and R described above, we gather information on county characteristics that measure their business and agricultural conditions, including the percentage of black population, the percentage of farmland, and manufacturing output per capita per square miles. These covariates (X) are used for validity checks. More information on the data along with sample summary statistics is provided in Appendix D.1.

It is worth mentioning that in our sample, less than 1% of the banks below the regulatory threshold hold the required minimum capital, \$25,000, and less than 2% of the banks above the threshold hold the required minimum capital, \$50,000. Lack of mass points at the required minimum capital levels ensures that our Assumption 1 holds.

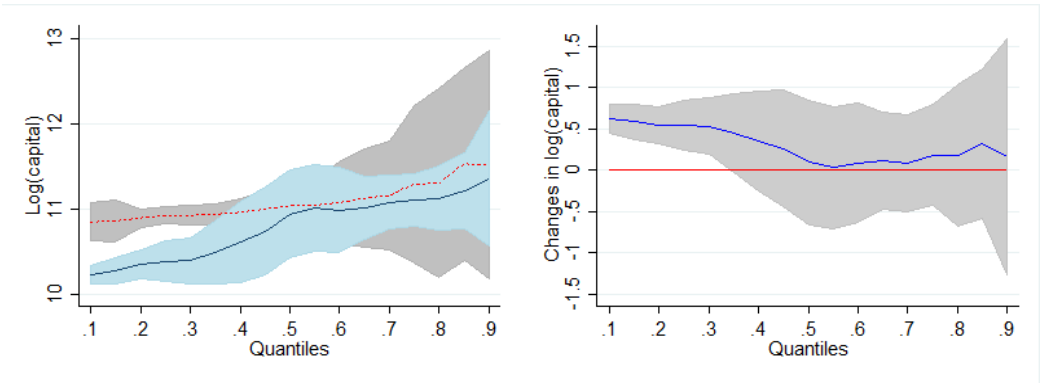

Figure 2: Estimated quantile curves of bank capital above and below the population threshold 3,000 (left) and quantile changes (right).

6.1 Estimation results

Figure 2 visualizes the estimated quantile curves of log capital above or below the policy threshold (left) and the estimated quantile changes (right) along with their 95% point-wise confidence bands. These estimates are generated using $\widehat{h}^*_{R\pi} = 1,462.76$. For simplicity, all estimates in the empirical analysis use uniform kernels, unless otherwise stated. Consistent with the visual evidence in Figure 1, Figure 2 suggests that significant changes only occur at roughly the bottom 30 percentiles of the distribution of log capital. The estimated changes are also larger at lower quantiles. In contrast, the estimated mean change in log capital using $\widehat{h}^*_{R\pi} = 1,462.76$ is 0.107 with a standard error 0.148. The estimated mean change by the default CCT rdrobust package (using $\widehat{h}_R = 803.58$ and a triangular kernel) is 0.141 with a standard error 0.171. The lack of a significant mean change in bank capital suggests that the standard fuzzy RD design does not apply.

Figure 3 illustrates the bias-corrected estimates of Q-LATEs at different quantiles along with their 95% confidence intervals. The main bandwidths used for estimation are $\widehat{h}^*_{R_\pi} = 1,426.76$ and $\widehat{h}^{rot}_{T\pi} = 0.441$. The bandwidth for bias estimation is set to be $4.5n^{-1/8} = 2,308.67$, corresponding to $\rho = 0.618$ (See Appendix C.3 for details). A preliminary bandwidth $3/4\widehat{h}^*_{R_\pi} = 1,097.07$ is used to determine the trimming thresholds. Alternative results based on undersmoothing or bootstrapped standard errors (with or without being clustered at the town level) are presented in Appendix D.2. Clustering seems to have little impacts based on the bootstrapped standard errors.

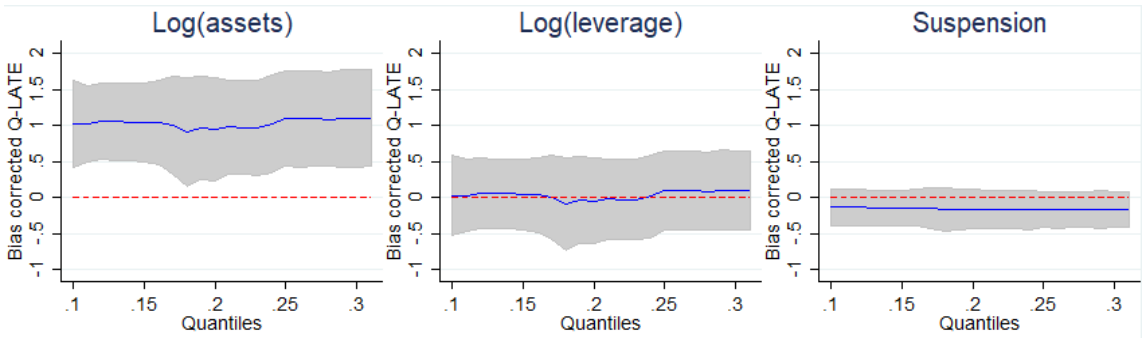

Figure 3: Bias-corrected estimates of Q-LATEs at different quantiles

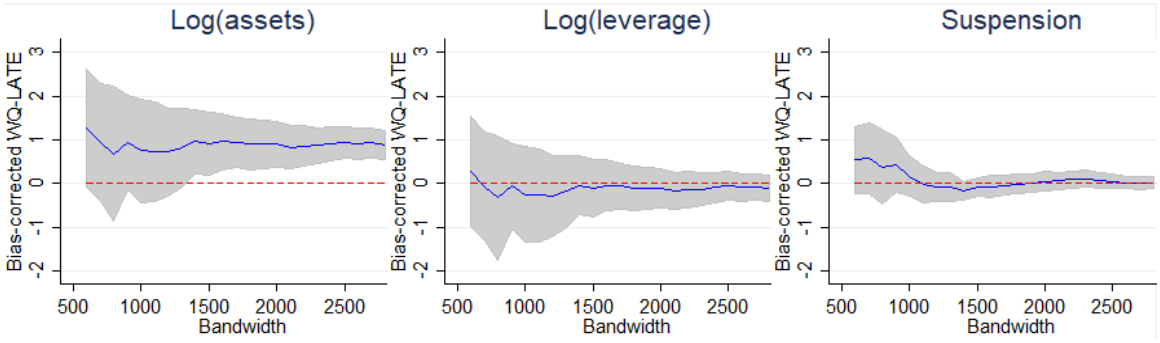

Figure 4: Bias-corrected estimates of WQ-LATEs by different bandwidths

Our analytical standard errors therefore do not take into account possible clustering at the town level.

As shown in Figure 3, the estimated Q-LATEs for log assets are around 1 at various low quantiles of log capital. All estimates are significant at the 1% level. The corresponding WQ-LATE is estimated to be 1.034, which is also significant at the 1% level, so on average, a 1% increase in capital leads to roughly a 1% increase in assets among those banks at lower quantiles of the capital distribution. The estimated impacts on log leverage and those on the long-run risk of suspending operation are small and insignificant.

Figure 4 further plots the bias-corrected estimates of WQ-LATEs (along with the 95% confidence intervals) against different bandwidth choices. The point estimates of WQ-LATEs are robust to a wide range of bandwidth choices, even though as expected, the confidence intervals get wider as the bandwidth gets smaller. Calonico, Cattaneo, and Farrell (2020) develop a new bandwidth selector for robust bias-corrected confidence intervals with minimal coverage error, in the context

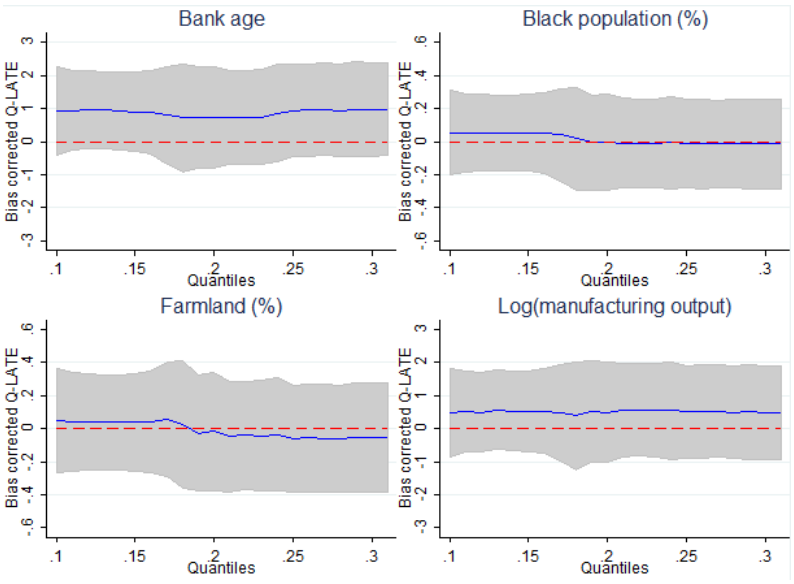

Figure 5: Bias-corrected estimates of Q-LATEs on covariates (first moments)

of the standard RD design. A formal development of such coverage-error optimal bandwidths for the Q-LATE and WQ-LATE estimators is out of the scope of this paper, but we can implement the rule-of-thumb bandwidth suggested in Calonico, Cattaneo, and Farrell (2020), i.e., the rescaled AMSE optimal bandwidth $n^{-1/20}\hat{h}^*_{R\pi} = 1045.75$. As shown in Figure 4, at this bandwidth, the point estimates of WQ-LATEs are largely consistent with those estimates at our AMSE optimal bandwidth, even though the confidence intervals are much wider.

Overall, our empirical analysis suggests that while the minimum capital requirement induces small banks (i.e., banks at the bottom 30% of the capital distribution) to hold more capital, these banks adjust their assets proportionately. That is, banks simply scale up without a ratio regulation. As a result, their leverages and long-run risk of failure remain almost unchanged. These results help us better understand the frequent bank runs and banking panics prior to the Great Depression.

6.2 Validity checks

Validity of our estimates requires our identifying assumptions to hold. This section performs the proposed joint specification tests. For simplicity, instead of testing the entire distribution of covariates, we test the low order (raw) moments of covariates. That is, we replace the outcome variable

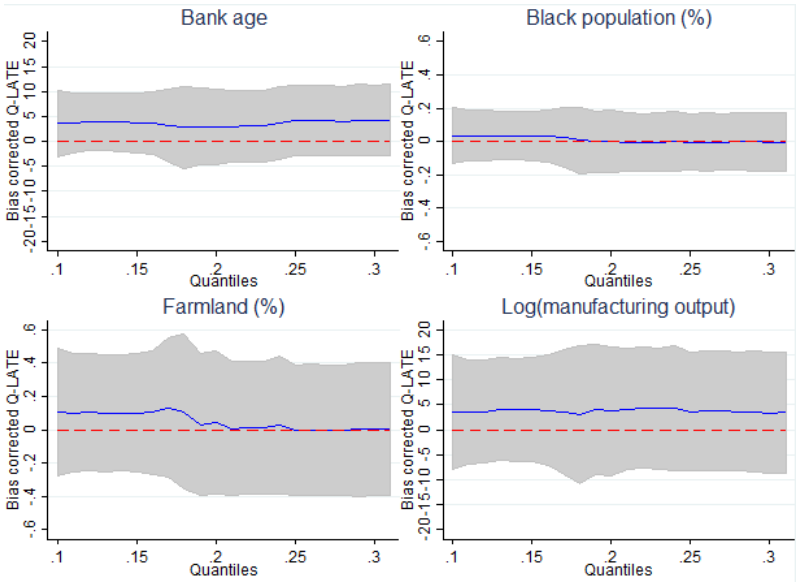

Figure 6: Bias-corrected estimates of Q-LATEs on covariates (second moments)

by each of the first and second moments of the four covariates (i.e., bank age, percentage of black population, percentage of farmland, and log of manufacturing output per capita) and re-estimate Q-LATEs. We use the same bandwidths and specification as those used for our main estimation. Results of these falsification tests are visualized in Figures 5 and 6. Table D3.1 in the Appendix further reports the bias-corrected estimates of WQ-LATEs on the first two moments of the covariates. None of these estimates are statistically significant.

In addition to our joint tests, we also perform the standard RD validity checks, including the density test and covariates smoothness test. Details of these tests and formal testing results are provided in Appendix D.3. Figure 7 presents the histogram of the town population (left) and the log frequency of the town population within each bin of 200 population (right). Superimposed on the right graph is the estimated log density along with the 95% confidence interval. Figure 8 plots the mean of the covariate in a bin of town population against the mid-point of the bin. The bars mark the 95% confidence intervals. Overall we do not find evidence that banks took advantage of the lower capital requirement and hence were more likely to operate in towns with populations just under 3,000. Results of our validity checks strongly support the plausibility of our assumptions.

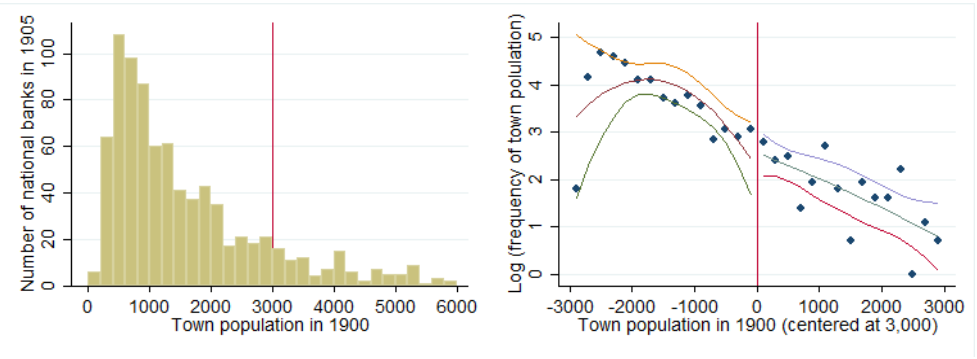

Figure 7: Histogram and the empirical density of town population

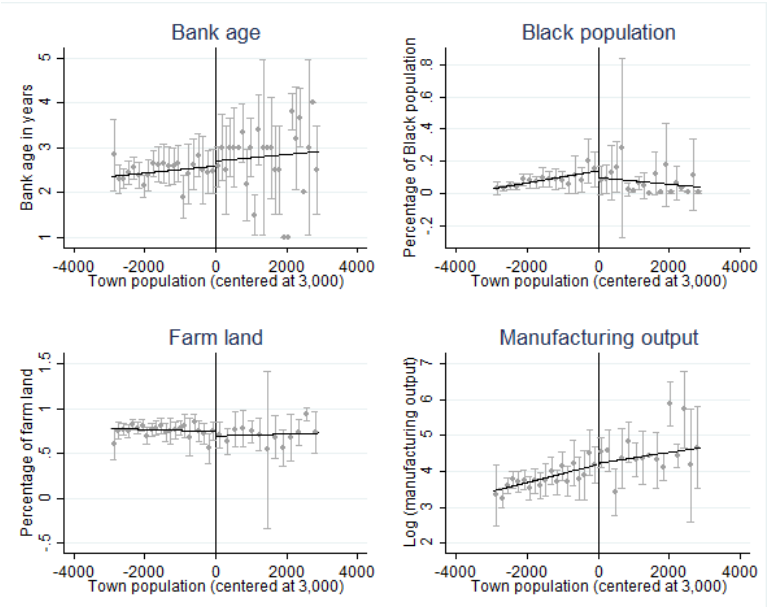

Figure 8: Conditional means of covariates conditional on town population

7 Conclusion

An empirically important class of fuzzy RD designs involve continuous treatments. This paper provides nonparametric identification and robust bias-corrected inference for such RD designs. We utilize for identification any distributional changes in the continuous treatment at the RD threshold, including the usual mean change as a special case. Our model can potentially apply to a large class of policies that target parts or features of the treatment distribution, such as changing the mean, changing the variance or shifting one or both tails of the distribution. Treatment changes in general are responses to relevant policies. By focusing on where the true changes are in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects. Our empirical application demonstrates the usefulness of the proposed approach.

References

[1] Almond D., Doyle J. J., Kowalski A. E., and Williams H. (2010): "Estimating marginal returns to medical care: evidence from at-risk newborns," The Quarterly Journal of Economics, 125(2), 591-634.

[2] Angrist, J. D., G. Imbens, and K. Graddy, (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," The Review of Economic Studies, 67, 499-527.

[3] Arai, Y. and H. Ichimura (2018): "Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator," Quantitative Economics, 9(1), 441-482.

[4] Arkhangelsky D. and G. W. Imbens (2021): "Double-Robust Identification for Causal Panel Data Models," NBER working paper No. w28364.

[5] Caetano, C., G. Caetano and J. C. Escanciano, (2020): "Regression Discontinuity Design with Multivalued Treatments," Working paper.

[6] Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Bias Corrected Inference in Regression Discontinuity Design," Econometrica, 82(6), 2295-2326.

[7] Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," Journal of the American Statistical Association 113(522), 767-779.

[8] Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019): "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," arXiv:1808.01398.

[9] Calonico, S., M. D. Cattaneo, and M. H. Farrell (2020): "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression-Discontinuity Designs," Econometrics Journal, 23, 192-210.

[10] Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on causal effects in a generalized regression kink design," Econometrica, 83(6), 2453-2483.

[11] Cattaneo, M. D., N. Idrobo and R. Titiunik (2019): A Practical Introduction to Regression Discontinuity Designs: Foundations. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

[12] Cattaneo, M. D., N. Idrobo and R. Titiunik (2020a): A Practical Introduction to Regression Discontinuity Designs: Extensions. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

[13] Cattaneo, M. D., N. Idrobo and R. Titiunik (2020b): The Regression Discontinuity Design. Handbook of Research Methods in Political Science and International Relations.

[14] Chen, Y., A. Ebenstein, M. Greenstone, and H. Li (2013): "Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy," PNAS, 110 (32), 12936-12941.

[15] Chernozhukov, V. and Hansen, C. (2005): "An IV Model of Quantile Treatment Effects," Econometrica, 73, 245-261.

[16] D'haultfoeuille, X. and P. Février (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," Econometrica, 83(3), 1199-1210.

[17] Ebenstein, A., M. Fan, M. Greenstone, G. He, and M. Zhou (2017): "New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy," PNAS, 114 (39) 10384-10389.

[18] Frandsen B., M. Frölich, and B. Melly (2012): "Quantile Treatment Effects in the Regression Discontinuity Design," Journal of Econometrics, 168, 382-395.

[19] Giuntella, O. and F. Mazzonna (2019): "Sunset time and the economic effects of social jetlag: evidence from US time zone borders," Journal of Health Economics, 65, 210-226.

[20] Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69(1), 201-209.

[21] Horowitz, J. (2001): "Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function, " Econometrica, 69(2), 499-513.

[22] Imbens, G. and K. Kalaynaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," The Review of Economic Studies, 79(3), 933-959.

[23] Imbens, G. W. and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics, 142, 615-635.

[24] Imbens, G. and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Econometrica, 77, 1481-1512.

[25] Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and its Application to the Additive Model," Econometric Theory, 26(5), 1529-1564.

[26] Litschig, S., and K. Morrison (2010): "Government Spending and Re-election: Quasi-Experimental Evidence from Brazilian Municipalities," UPF Discussion Paper.

[27] Newey, W. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," Econometric Theory, 10(2), 233-253.

[28] Qu, Z. and J. Yoon (2015): "Nonparametric Estimation and Inference on Conditional Quantile Processes," Journal of Econometrics, 185(1), 1-19.

[29] Shen, S. and X. Zhang (2016): "Distributional Tests for Regression Discontinuity: Theory and Empirical Examples," The Review of Economics and Statistics, 98(4): 685-700.

[30] Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," Econometrica, 83(3),1185-1197.

Supplemental Appendix for "Regression Discontinuity Designs with a Continuous Treatment"

Yingying Dong, Ying-Ying Lee, Michael Gou

This Appendix is organized as follows. Section A provides proofs for the lemmas, theorem, and corollary presented in Section 2 Identification. Sections B.1 and B.2 provide some preliminary lemmas along with their proofs to facilitate deriving the asymptotic properties for the proposed estimators. Sections B.3 and B.4 then present proofs for the theorems presented in Section 5 Inference. Section C describes how to estimate the biases, variances of the Q-LATE and WQ-LATE estimators as well as the AMSE optimal bandwidths discussed in Sections 5.2 and 5.3. Section D provides data description and various additional results of the empirical analysis.

A Proofs for Section 2 Identification

Proof of Lemma 1.1 By Bayes' Theorem, Assumption 3 $U_0|(\varepsilon, R=r_0) \sim U_1|(\varepsilon, R=r_0)$ means $\varepsilon|(U_0=u, R=r_0) \sim \varepsilon|(U_1=u, R=r_0)$, or $f_{\varepsilon|U_1R}(e,u,r_0) = f_{\varepsilon|U_0R}(e,u,r_0)$. Further,

$$\begin{aligned} f_{\varepsilon|U_1R}(e,u,r_0) &= f_{\varepsilon|U_0R}(e,u,r_0) \overset{(1)}{\Longleftrightarrow} \\ \lim_{r\to r_0^+} f_{\varepsilon|U_1R}(e,u,r) &= \lim_{r\to r_0^-} f_{\varepsilon|U_0R}(e,u,r) = \lim_{r\to r_0} f_{\varepsilon|UR}(e,u,r) \overset{(2)}{\Longleftrightarrow} \\ \lim_{r\to r_0^+} f_{\varepsilon|TR}(e,q_1(r,u),r) &= \lim_{r\to r_0^-} f_{\varepsilon|TR}(e,q_0(r,u),r) = \lim_{r\to r_0} f_{\varepsilon|UR}(e,u,r), \end{aligned}$$

where equivalence (1) follows from continuity of $f_{\varepsilon|U_zR}(e,u,r)$ in Assumption 2 and the definition $U \equiv U_1\mathbf{1}(R \geq r_0) + U_0\mathbf{1}(R < r_0)$, and (2) follows from the fact that given $R=r$ for $r>0$ ($r<0$), u and $q_1(r,u)$ ($q_0(r,u)$) is a one-to-one mapping. Note that by continuity of $q_z(r,u)$ and $f_{\varepsilon|U_zR}(e,u,r)$, $z=0,1$, $\lim_{r\to r_0^+} f_{\varepsilon|TR}(e,q_1(r,u),r) = f_{\varepsilon|TR}(e,t_1(u),r_0)$ and $\lim_{r\to r_0^-} f_{\varepsilon|TR}(e,q_0(r,u),r) = f_{\varepsilon|TR}(e,t_0(u),r_0)$, so the above shows $f_{\varepsilon|TR}(e,t_1(u),r_0) = f_{\varepsilon|TR}(e,t_0(u),r_0)$. That is, given $U=u$, any potential changes in T when $R \to r_0$ are independent of ε.

Proof of Lemma 1.2

$$\begin{aligned} &\lim_{r\to r_0^+} \mathbb{E}[Y|U=u, R=r] - \lim_{r\to r_0^-} \mathbb{E}[Y|U=u, R=r] \\ &= \lim_{r\to r_0^+} \mathbb{E}\left[Y|T=q_1(r,u), U_1=u, R=r\right] - \lim_{r\to r_0^-} \mathbb{E}\left[Y|T=q_0(r,u), U_0=u, R=r\right] \\ &= \lim_{r\to r_0^+} \mathbb{E}\left[G(q_1(r,u),r,\varepsilon)|U_1=u, R=r\right] - \lim_{r\to r_0^-} \mathbb{E}\left[G(q_0(r,u),r,\varepsilon)|U_0=u, R=r\right] \\ &= \mathbb{E}[G(t_1(u),r_0,\varepsilon)|U_1=u, R=r_0] - \mathbb{E}[G(t_0(u),r_0,\varepsilon)|U_0=u, R=r_0] \\ &= \int (G(t_1(u),r_0,e) - G(t_0(u),r_0,e))\, F_{\varepsilon|UR}(de,u,r_0), \end{aligned}$$

where the first equality follows from Assumption 1; the second equality follows from the definition $Y = G(T, R, \varepsilon)$; the third equality follows from the continuity conditions in Assumption 2 and compact support, which together ensure interchangeability of limit and expectation (integral). It follows that $\mathbb{E}\left[G\left(q_z(r,u), r, \varepsilon\right)|U_z = u, R = r\right] = \int_{\mathcal{E}} G\left(q_z\left(r,u\right), r, \varepsilon\right) f_{\varepsilon|U_z R}\left(e, u, r\right) de$, $z = 0, 1$, is continuous in r. The last equality follows from the fact that Assumption 3 implies $f_{\varepsilon|U_1 R}\left(e, u, r_0\right) = f_{\varepsilon|U_0 R}\left(e, u, r_0\right) = f_{\varepsilon|U R}\left(e, u, r_0\right)$.

Proof of Theorem 1 By definition, $T = q(r,u) = q_0(r,u)\left(1 - Z\right) + q_1(r,u)Z$. Further by smoothness of $q_z(r,u)$, $z = 0, 1$ in Assumption 2, the right and left limits of $q(r,u)$ at $r = r_0$ exist, i.e., $\lim_{r \to r_0^+} q(r,u) = q_1(r_0,u) \equiv t_1\left(u\right)$ and $\lim_{r \to r_0^+} q(r,u) = q_1(r_0,u) \equiv t_0\left(u\right)$. Equation (3) holds following Lemma 1. $\pi\left(w\right) \equiv \int_{\mathcal{U}} \tau\left(u\right) w\left(u\right) du$ is identified since 1) $\tau\left(u\right)$ is identified, 2) the weighting function $w\left(u\right)$ is assumed to be known or estimable, and 3) the set $\mathcal{U} \equiv \{u \in [0,1]: |t_1\left(u\right) - t_0\left(u\right)| > 0\}$ is identified given that $q_z(r,u)$, $z = 0, 1$ is identified.

Proof of Lemma 2 Assumption 3b monotonicity states $\Pr\left(t_1(U_1) \geq t_0(U_0)|R = r_0\right) = 1$ or $\Pr\left(t_1(U_1) \leq t_0(U_0)|R = r_0\right) = 1$. Without loss of generality, we assume the former is true. Given the smoothness conditions in Assumption 2, we have

$$
\begin{aligned}
\pi^{RD} &\equiv \frac{\lim_{r \to r_0^+} \mathbb{E}\left[Y|R = r\right] - \lim_{r \to r_0^-} \mathbb{E}\left[Y|R = r\right]}{\lim_{r \to r_0^+} \mathbb{E}\left[T|R = r\right] - \lim_{r \to r_0^-} \mathbb{E}\left[T|R = r\right]} \\
&= \frac{\lim_{r \to r_0^+} \mathbb{E}\left[G\left(q_1(r, U_1), r, \varepsilon\right)|R = r\right] - \lim_{r \to r_0^-} \mathbb{E}\left[G\left(q_0(r, U_0), r, \varepsilon\right)|R = r\right]}{\lim_{r \to r_0^+} \mathbb{E}\left[q_1(r, U_1|R = r\right] - \lim_{r \to r_0^-} \mathbb{E}\left[q_0(r, U_0)|R = r\right]} \\
&= \frac{\mathbb{E}\left[G\left(t_1(U_1), r_0, \varepsilon\right)|R = r_0\right] - \mathbb{E}\left[G\left(t_0\left(U_0\right)\right), r_0, \varepsilon\right)|R = r_0]}{\mathbb{E}\left[t_1(U_1)|R = r_0\right] - \mathbb{E}\left[t_0(U_0)|R = r_0\right]} \\
&= \frac{\mathbb{E}\left[G\left(t_1(U_1), r_0, \varepsilon\right) - G\left(t_0\left(U_0\right)\right), r_0, \varepsilon\right)|R = r_0]}{\mathbb{E}\left[t_1(U_1) - t_0(U_0)|R = r_0\right]} \\
&= \frac{\iint\int \Delta_Y\left(u_0, u_1, e\right) F_{\varepsilon|U_0 U_1, R = r_0}\left(de, u_0, u_1\right) F_{U_0 U_1|R = r_0}\left(du_0, du_1\right)}{\iint \Delta_T\left(u_0, u_1\right) F_{U_0 U_1|R = r_0}\left(du_0, du_1\right)} \\
&= \iint_{\mathcal{I}}\int \frac{\Delta_Y\left(u_0, u_1, e\right)}{\Delta_T\left(u_0, u_1\right)} \widetilde{w}^{RD}\left(u_0, u_1\right) F_{\varepsilon|U_0 U_1, R = r_0}\left(de, u_0, u_1\right) F_{U_0 U_1|R = r_0}\left(du_0, du_1\right),
\end{aligned}
$$

where $\Delta_Y\left(u_0, u_1, e\right) \equiv G\left(t_1(u_1), r_0, e\right) - G\left(t_0(u_0), r_0, e\right)$, $\Delta_T\left(u_0, u_1\right) \equiv t_1(u_1) - t_0(u_0)$, $\widetilde{w}^{RD}\left(u_0, u_1\right) \equiv \frac{\Delta_T(u_0, u_1)}{\iint_{\mathcal{I}} \Delta_T(u_0, u_1) F_{U_0 U_1|R = r_0}(du_0, du_1)}$, $\mathcal{I} \equiv \{u_0, u_1 \in [0,1]: t_1(u_1) - t_0\left(u_0\right) > 0\}$.

Under Assumption 3b monotonicity, $\widetilde{w}^{RD}\left(u_0, u_1\right) > 0$ and $\iint_{\mathcal{I}} \widetilde{w}^{RD}(u_0, u_1) F_{U_0 U_1|R = r_0}(du_0, du_1)$ $= 1$. Therefore, under Assumptions 2, 3b, and 4, π^{RD} identifies a weighted average of individual causal effects, $\frac{\Delta_Y(u_0, u_1, e)}{\Delta_T(u_0, u_1)} \equiv \frac{G(t_1(u_1), r_0, e) - G(t_0(u_0), r_0, e)}{t_1(u_1) - t_0(u_0)}$, among those having $t_1 u_1) - t_0\left(u_0\right) > 0$.

Further, when the function $G\left(T, R, \varepsilon\right)$ is continuously differentiable in its first argument, we

have

$$
\begin{aligned}
\pi^{RD} &\equiv \frac{\lim_{r\to r_0^+}\mathbb{E}[Y|R=r]-\lim_{r\to r_0^-}\mathbb{E}[Y|R=r]}{\lim_{r\to r_0^+}\mathbb{E}[T|R=r]-\lim_{r\to r_0^-}\mathbb{E}[T|R=r]} \\
&= \frac{\mathbb{E}[G(t_1(U_1),r_0,\varepsilon)-G(t_0(U_0),r_0,\varepsilon)|R=r_0]}{\mathbb{E}[t_1(U_1)-t_0(U_0)|R=r_0]} \\
&= \frac{\mathbb{E}\left[\int_{t_0(U_0)}^{t_1(U_1)}\frac{\partial G(t,r_0,\varepsilon)}{\partial t}dt\middle|R=r_0\right]}{\mathbb{E}\left[\int_{t_0(U_0)}^{t_1(U_1)}1dt\middle|R=r_0\right]} \\
&= \frac{\mathbb{E}\left[\int\frac{\partial G(t,r_0,\varepsilon)}{\partial t}\mathbf{1}(t_0(U_0)\le t\le t_1(U_1))\,dt\middle|R=r_0\right]}{\mathbb{E}\left[\int\mathbf{1}(t_0(U_0)\le t\le t_1(U_1))\,dt\middle|R=r_0\right]} \\
&= \int\mathbb{E}\left[\frac{\partial G(t,r_0,\varepsilon)}{\partial t}\middle|R=r_0,t_0(U_0)\le t\le t_1(U_1)\right] \\
&\quad\times\frac{\Pr(t_0(U_0)\le t\le t_1(U_1)|R=r_0)}{\int\Pr(t_0(U_0)\le t\le t_1(U_1)|R=r_0)\,dt}dt \\
&= \int\mathbb{E}\left[\frac{\partial G(t,r_0,\varepsilon)}{\partial t}\middle|R=r_0,t_0(U_0)\le t\le t_1(U_1)\right]\overline{w}^{RD}dt,
\end{aligned}
$$

where $\overline{w}^{RD}\equiv\frac{\Pr(t_0(U_0)\le t\le t_1(U_1))|R=r_0)}{\int\Pr(t_0(U_0)\le t\le t_1(U_1))dt}$, the third equality follows from Lemma 5 in the Appendix of Angrist, Graddy and Imbens (2000), and the fifth equality follows from the law of iterated expectations and interchanging the order of integration under the standard regularity conditions. See Theorem 1 of Angrist and Imbens (1995) for a similar expression when they discuss the LATE model with a binary IV and variable treatment intensity.

Note that Lemma 2 describes the two types of averaging that characterize the standard RD estimand in the case of a continuous treatment. First, there is averaging over some of the individuals at a given treatment level t. This is reflected in the expectation $\mathbb{E}\left[\frac{\partial G(t,r_0,\varepsilon)}{\partial t}\middle|R=r_0,t_0(U_0)\le t\le t_1(U_1)\right]$. For any treatment level t, only those individuals whose potential treatments at $R=r_0$, namely $t_1(U_1)$ and $t_0(U_0)$, bracket this treatment t enter into the expectation. Second, there is averaging over different treatment levels even for the same individuals. Averaging over different treatment levels is reflected in the outer integration and the weighting function $\overline{w}^{RD}$. The weight given to any particular treatment level is proportional to the fraction of individuals whose treatment changes bracket this treatment level.

Proof of Theorem 2 When Assumption 3 the local treatment rank restriction holds along with Assumptions 1, 2 and 4, $\pi^*\equiv\int_{\mathcal{U}}\frac{m^+(u)-m^-(u)}{q^+(u)-q^-(u)}\frac{|q^+(u)-q^-(u)|}{\int_{\mathcal{U}}|q^+(u)-q^-(u)|du}du$ identifies $\pi(w^*)$, which is a special case of the WQ-LATE in Theorem 1 using a weighting function $w^*(u)\equiv\frac{|\Delta q(u)|}{\int_0^1|\Delta q(u)|du}$. Note that $w^*(u)>0$ by construction, so $\pi(w^*)$ is a weighted average effect by Theorem 1 and the discussion in the main text.

Alternatively, when Assumption 3b monotonicity holds along with Assumptions 1, 2 and 4,

$\pi^* = \pi^{RD}$. π^{RD} identifies a weighted average effect by Lemma 2.

B Proofs for Section 5 Inference

This section proceeds as follows. We first introduce notation. Section B.1 presents preliminary lemmas to facilitate establishing asymptotics. Section B.2 presents asymptotic theorems under undersmoothing. These lemmas and theorems can also be of independent interest. Section B.3 collects the proofs of the lemmas in Section B.1. Section B.4 provides the proofs of Theorem 7, Theorem 3, and Theorem 4 in Section 5, which pertain to $\widehat{\tau}(u)$. Section B.5 presents the proofs of Theorem 8, Theorem 5, and Theorem 6 in Section 5, which pertain to $\hat{\pi}^*$.

Notation. Let $f^{\pm}_{T|R}(u) \equiv \lim_{r\to r_0^{\pm}} f_{T|R}(q^{\pm}(u), r)$, $q''^{\pm}_r(u) \equiv \lim_{r\to r_0^{\pm}} \partial^2 q(r,u)/\partial r^2$, $m'^{\pm}_t(u) \equiv \lim_{r\to r_0^{\pm}} \partial \mathbb{E}\big[Y|T=t,\, R=r\big]/\partial t|_{t=q^{\pm}(u)}$, $m''^{\pm}_t(u) \equiv \lim_{r\to r_0^{\pm}} \partial^2 \mathbb{E}\big[Y|T=t,\, R=r\big]/\partial t^2|_{t=q^{\pm}(u)}$, $m''^{\pm}_r(u) \equiv \lim_{r\to r_0^{\pm}} \partial^2 \mathbb{E}\big[Y|T=q^{\pm}(u),\, R=r\big]/\partial r^2$, and $\sigma^{2\pm}(u) \equiv \lim_{r\to r_0^{\pm}} \mathbb{V}\big[Y\big|T=q^{\pm}(u),\, R=r\big]$. Define $\Lambda^{\pm}(u) \equiv \big(m'^{\pm}_t(u) - \pi^*\big)w^*(u)/\Delta q(u)$.

The following constants are defined by the kernel function. $\kappa_j \equiv \int_0^{\infty} v^j K(v)dv$, $\lambda_j \equiv \int_0^{\infty} v^j K^2(v)dv$, $C_{\mathsf{V}} \equiv 4\big(\kappa_2^2\lambda_0 - 2\kappa_1\kappa_2\lambda_1 + \kappa_1^2\lambda_2\big)\big(\kappa_2 - 2\kappa_1^2\big)^{-2}$, $C_{\mathsf{B}} \equiv \big(\kappa_2^2 - \kappa_1\kappa_3\big)\big(\kappa_2 - 2\kappa_1^2\big)^{-1}$, and $C_{\mathsf{C}}(\rho) \equiv \int_0^{\infty} K(v/\rho)K(v)dv\big(\rho\kappa_2 \int_0^{\infty} K(v/\rho)K(v)dv - \kappa_1 \int_0^{\infty} vK(v/\rho)K(v)dv\big)$.[1] Define the 6×6 symmetric matrices

$$S_2 \equiv \begin{pmatrix} 1/2 & \kappa_1 & 0 & \kappa_2 & 0 & \kappa_2 \\ & \kappa_2 & 0 & \kappa_3 & 0 & 2\kappa_2\kappa_1 \\ & & \kappa_2 & 0 & 2\kappa_2\kappa_1 & 0 \\ & & & \kappa_4 & 0 & 2\kappa_2^2 \\ & & & & 2\kappa_2^2 & 0 \\ & & & & & \kappa_4 \end{pmatrix} \text{ and } \Lambda_2 \equiv \begin{pmatrix} \lambda_0 & \lambda_1 & 0 & \lambda_2 & 0 & 0 \\ & \lambda_2 & 0 & \lambda_3 & 0 & 0 \\ & & 0 & 0 & 0 & 0 \\ & & & \lambda_4 & 0 & 0 \\ & & & & 0 & 0 \\ & & & & & 0 \end{pmatrix}.$$

Let e_j be the 6×1 jth unit column vector, i.e., it has 1 as the jth entry and 0's as all other entries.

For the variances of $\hat{\tau}(u)$ and $\hat{\tau}^{bc}(u)$,

$$\mathsf{V}_{\tau}(u) \equiv \frac{2\lambda_0 C_{\mathsf{V}}}{(\Delta q(u))^2\, f_R(r_0)}\left(\frac{\sigma^{2+}(u)}{f^{+}_{T|R}(u)} + \frac{\sigma^{2-}(u)}{f^{-}_{T|R}(u)}\right) \tag{B.1}$$

$$\mathsf{V}_{\mathsf{B}_{\tau}}(u) \equiv \mathsf{V}_{\tau}(u) C_{\mathsf{V}}^{-1} 4\lambda_0 \left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^{\top} S_2^{-1} e_1 e_1^{\top} S_2^{-1} \left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right) \tag{B.2}$$

$$\mathsf{C}_{\tau}(u;\rho) \equiv -\mathsf{V}_{\tau}(u) \frac{8\left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^{\top} S_2^{-1} e_1}{\lambda_0 C_{\mathsf{V}}(\kappa_2 - 2\kappa_1^2)} C_{\mathsf{C}}(\rho) \tag{B.3}$$

[1] For the Uniform kernel, $\lambda_0 = 1/4$, $C_{\mathsf{V}} = 4$, $C_{\mathsf{B}} = -1/12$, $C_{\mathsf{C}}(\rho) = \rho^3/384$ if $\rho \le 1$, and $C_{\mathsf{C}}(\rho) = 0.03125(\rho/3 - 0.25)$ if $\rho > 1$. For the Epanechnikov kernel, $\lambda_0 = 0.3$, $C_{\mathsf{V}} = 0.243$, $C_{\mathsf{B}} = 0.07414$, $C_{\mathsf{C}}(\rho) = 0.$if $\rho = 0$, and $C_{\mathsf{C}}(\rho) = \lambda_0(\kappa_2\lambda_0 - \kappa_1\lambda_1)$ if $\rho = 1$.

For the variance of $\hat{\pi}^*$ and $\hat{\pi}^{bc}$,

$$\mathsf{V}_\pi \equiv \mathsf{V}_\pi^m + \mathsf{V}_\pi^q, \text{ where} \tag{B.4}$$

$$\mathsf{V}_\pi^m \equiv \frac{C_\mathsf{V} \int_\mathcal{U} \left(\sigma^{2+}(u) + \sigma^{2-}(u)\right) du}{f_R(r_0) \left(\int_\mathcal{U} |\Delta q(u)| du\right)^2} \tag{B.5}$$

$$\mathsf{V}_\pi^q \equiv \frac{C_\mathsf{V}}{f_R(r_0)} \int_\mathcal{U} \int_\mathcal{U} (\min\{u, v\} - vu) \left(\frac{\Lambda^+(u)\Lambda^+(v)}{f_{T|R}^+(u)\, f_{T|R}^+(v)} + \frac{\Lambda^-(u)\Lambda^-(v)}{f_{T|R}^-(u)\, f_{T|R}^-(v)} \right) dv du$$

$$\mathsf{V}_{\mathsf{B}_\pi} \equiv \mathsf{V}_\pi^m C_\mathsf{V}^{-1} 4 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top S_2^{-1} \Lambda_2 S_2^{-1} \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right) \tag{B.6}$$

$$\mathsf{C}_\pi(\rho) \equiv - \mathsf{V}_\pi^m \frac{8 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top S_2^{-1}}{C_\mathsf{V} \left(\kappa_2 - 2\kappa_1^2\right)} \int_0^\infty K(v) K(v/\rho) \mathbf{v} (\kappa_2 - \kappa_1 v/\rho) dv \tag{B.7}$$

where $\mathbf{v} \equiv (1, v, 0, v^2, 0, 0)^\top$. V_π^m is due to estimation of $\Delta\hat{m}(u)$ in Step 2 and V_π^q is due to estimation of $\Delta\hat{q}(u)$ in Step 1. For $\rho = 1$, the integration in $\mathsf{C}_\pi(\rho)$ becomes $\left(\kappa_2\lambda_0 - \kappa_1\lambda_1, \kappa_2\lambda_1 - \kappa_1\lambda_2, 0, \kappa_2\lambda_2 - \kappa_1\lambda_3, 0, 0\right)^\top$.

For the bias of $\hat{\tau}(u)$,

$$\mathsf{B}_{R\tau}(u) \equiv \left(\mathsf{B}_{R2}(u) + \mathsf{B}_1^+(u) \left(m_t'^+(u) - \tau(u)\right) - \mathsf{B}_1^-(u) \left(m_t'^-(u) - \tau(u)\right) \right) \Big/ \Delta q(u) \tag{B.8}$$

$$\mathsf{B}_{T\tau}(u) \equiv \mathsf{B}_{T2}(u) \big/ \Delta q(u) \tag{B.9}$$

$$\mathsf{B}_\tau(u) \equiv c_R^2 \mathsf{B}_{R\tau}(u) + c_T^2 \mathsf{B}_{T\tau}(u) \tag{B.10}$$

where $\mathsf{B}_{R2}(u) \equiv C_\mathsf{B}\left(m_r''^+(u) - m_r''^-(u)\right)$, $\mathsf{B}_{T2}(u) \equiv \kappa_2\left(m_t''^+(u) - m_t''^-(u)\right)$, and $\mathsf{B}_1^\pm(u) \equiv C_\mathsf{B} q_r''^\pm(u)$.

For the bias of $\hat{\pi}^*$,

$$\mathsf{B}_{R\pi} \equiv \int_\mathcal{U} \mathsf{B}_{R\tau}(u) w^*(u)\, du + \int_\mathcal{U} \left(\mathsf{B}_1^+(u) - \mathsf{B}_1^-(u)\right) \left(\tau(u) - \pi^*\right) \frac{w^*(u)}{\Delta q(u)} du \tag{B.11}$$

$$\mathsf{B}_{T\pi} \equiv \int_\mathcal{U} \mathsf{B}_{T\tau}(u) w^*(u)\, du \tag{B.12}$$

$$\mathsf{B}_\pi \equiv \int_\mathcal{U} \mathsf{B}_\tau(u) w^*(u)\, du + c_R^2 \int_\mathcal{U} \left(\mathsf{B}_1^+(u) - \mathsf{B}_1^-(u)\right) \left(\tau(u) - \pi^*\right) \frac{w^*(u)}{\Delta q(u)} du \tag{B.13}$$

Let $\mathbb{B}\left[\hat{\beta}\right] \equiv \mathbb{E}\left[\hat{\beta}\right] - \beta$ denote the bias for a generic estimator $\hat{\beta}$ of the parameter β and $\mathbb{C}$ [X, Y] denote the covariance of any two random variables X and Y. Let $\|\cdot\|_\infty$ be the sup-norm, i.e., $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

B.1 Preliminary asymptotic results

In the following, Lemma 3 presents the asymptotically linear representations for $\Delta\hat{q}(u)$ and $\Delta\hat{m}(u)$. Lemma 4(I) and Lemma 5(I) present the asymptotically linear representations for $\hat{\tau}(u)$ and $\hat{\pi}^*$, re-

spectively. Lemma 4(D) and Lemma 5(D) present the asymptotic distributions of $\hat{\tau}(u)$ and $\hat{\pi}^*$, respectively.

Lemma 3. *Let Assumptions 1-5 hold. Then uniformly in $u \in \mathcal{U}$,*

(Q) $\Delta\hat{q}(u) - \Delta q(u) - h_R^2\left(\mathsf{B}_1^+(u) - \mathsf{B}_1^-(u)\right) = n^{-1}\sum_{i=1}^n Z_i\Phi_{1i}^+(u) - (1-Z_i)\Phi_{1i}^-(u) + O_p(h_R^3) + o_p\left((nh_R)^{-1/2}\right)$, *where*

$$\Phi_{1i}^+(u) \equiv (u - \mathbf{1}(T_i \le q_1(R_i,u)))\frac{2(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_{TR}^+(u)(\kappa_2 - 2\kappa_1^2)}\frac{1}{h_R}K\left(\frac{R_i - r_0}{h_R}\right)$$

and $\Phi_{1i}^-(u)$ is defined analogously by replacing $q_1(R_i,u)$ with $q_0(R_i,u)$.

(M) $\Delta\hat{m}(u) - \Delta m(u) - \left(h_R^2\mathsf{B}_{R2}(u) + h_T^2\mathsf{B}_{T2}(u) + h_R^2\mathsf{B}_1^+(u)m_t'^+(u) - h_R^2\mathsf{B}_1^-(u)m_t'^-(u)\right) = n^{-1}\sum_{i=1}^n Z_i\left(\phi_{2i}^+(u) + \Phi_{1i}^+(u)m_t'^+(u)\right) - (1-Z_i)\left(\phi_{2i}^-(u) + \Phi_{1i}^-(u)m_t'^-(u)\right) + Rem$, *where*

$$\phi_{2i}^{\pm}(u) \equiv \left(Y_i - \left(m^{\pm}(u) + m_r'^{\pm}(u)(R_i - r_0) + m_t'^{\pm}(u)\left(T_i - q^{\pm}(u)\right)\right)\right) \times \frac{2(\kappa_2 - \kappa_1(R_i - r_0)/(h_R))}{f_{TR}^{\pm}(u)(\kappa_2 - 2\kappa_1^2)}\frac{1}{h_T}K\left(\frac{T_i - q^{\pm}(u)}{h_T}\right)\frac{1}{h_R}K\left(\frac{R_i - r_0}{h_R}\right)$$

and the remainder term $Rem = O_p\Big(\left(\log n/(nh_Th_R)\right)^{3/4} + \Big(\left(\log n/\left(nh_Rh_T^3\right)\right)^{1/2} + h_R + h_T\Big)\Big(\left(\log n/(nh_R)\right)^{1/2} + h_R^2\Big)\Big)$.

Lemma 4. *Let Assumptions 1-5 hold.*

(I) *Then uniformly in $u \in \mathcal{U}$, $\hat{\tau}(u) - \tau(u) - h_R^2\mathsf{B}_{R\tau}(u) - h_T^2\mathsf{B}_{T\tau}(u) = n^{-1}\sum_{i=1}^n IF_{\tau i}(u) + Rem$, where the influence function $IF_{\tau i}(u) \equiv \left(Z_i(\phi_{2i}^+(u) + \Phi_{1i}^+(u)(m_t'^+(u) - \tau(u))) - (1 - Z_i)(\phi_{2i}^-(u) + \Phi_{1i}^-(u)(m_t'^-(u) - \tau(u)))\right)(\Delta q(u))^{-1}$, and $\Phi_{1i}^{\pm}(u)$, $\phi_{2i}^{\pm}(u)$, and Rem are given in Lemma 3.*

(D) *If $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_Rh_T^2 \to \infty$, $nh_Th_R^5 \to c \in [0,\infty)$, $nh_Rh_T^5 \to c \in [0,\infty)$, and $h_R^2/h_T \to 0$, then for $u \in \mathcal{U}$, $\sqrt{nh_Rh_T}\left(\hat{\tau}(u) - \tau(u) - h_R^2\mathsf{B}_{R\tau}(u) - h_T^2\mathsf{B}_{T\tau}(u)\right) \to_d \mathcal{N}\left(0, \mathsf{V}_\tau(u)\right)$.*

The bandwidths conditions on h_T and h_R in Lemma 4 are symmetric, suggesting that h_R and h_T have the same first-order impact on $\hat{\tau}(u)$. This is because the local linear estimator of $\mathbb{E}[Y|T,R]$ in Step 2 dominates the first-order asymptotically linear representation, i.e., $\hat{q}^{\pm}$ from Step 1 is of smaller order.[2]

Lemma 5. *Let Assumptions 1, 2, either 3 or 3b, 4 and 5 hold and $l^{-1}\sqrt{nh_R} \to 0$.*

(I) *Then $\hat{\pi}^* - \pi^* - h_R^2\mathsf{B}_{R\pi} - h_T^2\mathsf{B}_{T\pi} = n^{-1}\sum_{i=1}^n IF_{\pi i} + Rem$, where the influence function*

[2]The condition $h_R^2/h_T \to 0$ is to control the remainder term of the product of the estimation errors of $\partial\hat{\mathbb{E}}\left[Y|T=t, R=r^+\right]/\partial t$ and $\hat{q}^+(u)$. This is from linearizing $\hat{m}^{\pm}(u) = \hat{\mathbb{E}}\left[Y|T=\hat{q}^{\pm}(u), R=r^+\right]$ in the estimation errors.

$$IF_{\pi i} \equiv Z_i \Phi_{21i}^{+} - (1-Z_i)\Phi_{21i}^{-} + \int_{\mathcal{U}} \left(Z_i \Phi_{1i}^{+}(u)\Lambda^{+}(u) - (1-Z_i)\Phi_{1i}^{-}(u)\Lambda^{-}(u) \right) du,$$

$$\begin{aligned}\Phi_{21i}^{\pm} \equiv & \left(Y_i - m(T_i, r_0^{\pm}) - m_r'(T_i, r_0^{\pm})\,(R_i - r_0) \right) \frac{\omega^*\left(F_{T|R}(T_i, r_0^{\pm})\right)}{\Delta q\left(F_{T|R}(T_i, r_0^{\pm})\right)} \\ & \times \frac{1}{h_R} K\left(\frac{R_i - r_0}{h_R}\right) \int \mathbf{1}\left(F_{T|R}(T_i + s h_T, r_0^{\pm}) \in \mathcal{U}\right) K(s) ds \\ & \times \frac{2\left(\kappa_2 - \kappa_1 (R_i - r_0)/(h_R)\right)}{f_R(r_0)\left(\kappa_2 - 2\kappa_1^2\right)},\end{aligned}$$

$m_r'(T_i, r_0^{\pm}) \equiv \lim_{r \to r_0^{\pm}} \partial \mathbb{E}[Y|T = T_i, R = r]/\partial r$, *and* $\Phi_{1i}^{\pm}(u)$ *and Rem are given in Lemma 3.*

(D) *If* $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_R h_T^3 \to \infty$, $nh_R^5 \to c \in [0, \infty)$, *and* $nh_R h_T^4 \to c \in [0, \infty)$, *then* $\sqrt{nh_R}\left(\hat{\pi}^* - \pi^* - h_R^2 \mathsf{B}_{R\pi} - h_T^2 \mathsf{B}_{T\pi}\right) \to_d \mathcal{N}(0, \mathsf{V}_\pi)$.

Define $\chi(u) = \mathbf{1}(|\Delta q(u)| > 0)$. Rewrite $\pi^* = \int_0^1 \tau(u)\omega^*(u)\chi(u)du$. In estimation, we replace $\chi(u)$ by $\hat{\chi}(u) = \mathbf{1}\left(|\Delta \hat{q}(u)| > \epsilon_n\right)$. Lemma 6 below shows that using $\hat{\chi}(u)$ is asymptotically equivalent to using $\chi(u)$.

Lemma 6. *Let the trimming parameter* ϵ_n *satisfy* $\epsilon_n^{-1} \sup_{u \in \mathcal{U}} \left| |\Delta \hat{q}(u)| - |\Delta q(u)| \right| = o_p(1)$ *and* $\epsilon_n^2 \left(\sup_{u \in \mathcal{U}} \left| |\Delta \hat{q}(u)| - |\Delta q(u)| \right| \right)^{-1} = o_p(1)$. *Then* $\int_0^1 \Delta \hat{q}(u)\left(\hat{\chi}(u) - \chi(u)\right) du = o_p\left(\sup_{u \in \mathcal{U}} \left| |\Delta \hat{q}(u)| - |\Delta q(u)| \right| \right)$.

Given the above Lemma 6, in the following proofs for $\hat{\pi}$ and $\hat{\pi}^{bc}$ we focus on estimators using the infeasible trimming function $\chi(u)$.

B.2 Asymptotic distributions under undersmoothing

Theorem 7 below presents the asymptotic distribution of $\hat{\tau}(u)$ under bandwidth sequences that go to zero fast enough with the sample size n (i.e., satisfying $nh_R^5 h_T \to 0$ and $nh_T^5 h_R \to 0$ instead of converging to $c \in (0, \infty)$), so that the bias is asymptotically negligible.

Theorem 7 (Asymptotic distribution of $\hat{\tau}(u)$)**.** *Let Assumptions 1-5 hold. If* $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_R h_T^2 \to \infty$, $nh_T h_R^5 \to 0$, $nh_R h_T^5 \to 0$, *and* $h_R^2 / h_T \to 0$, *then for* $u \in \mathcal{U}$

$$\frac{\hat{\tau}(u) - \tau(u)}{\sqrt{\mathsf{V}_{\tau,n}(u)}} \longrightarrow_d \mathcal{N}(0,1), \text{ where } \mathsf{V}_{\tau,n}(u) \equiv \frac{\mathsf{V}_\tau(u)}{nh_R h_T}.$$

The exact form of $\mathsf{V}_\tau(u)$ *is given by equation (B.1).*

The bandwidth conditions in Theorem 7 imply a bandwidth choice $h_R = h_{Rn} = c_R n^{-a}$ and $h_T = h_{Tn} = c_T n^{-a}$ for some constant $a \in (1/6, 1/3)$ and $c_R, c_T \in (0, \infty)$. Theorem 7 implies $\sqrt{nh_R h_T}\left(\hat{\tau}(u) - \tau(u)\right) \to_d \mathcal{N}(0, \mathsf{V}_\tau(u))$, where $\mathsf{V}_\tau(u)$ is the asymptotic variance of $\sqrt{nh_R h_T}\hat{\tau}(u)$. The $100(1-\alpha)\%$ confidence interval for $\tau(u)$ is then given by $\left[\hat{\tau}(u) \pm \Phi_{1-\alpha/2}^{-1}\sqrt{\mathsf{V}_\tau(u)/(nh_R h_T)}\right]$, where $\Phi_{1-\alpha/2}^{-1}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. One can estimate $\mathsf{V}_\tau(u)$ by the usual plug-in estimator, i.e., replacing the unknown parameters involved with their consistent estimates.

Theorem 8 below similarly presents the asymptotic distribution of $\hat{\pi}^*$ using bandwidth sequences that go to zero fast enough with the sample size (i.e., satisfying $nh_R^5 \to 0$ and $nh_R h_T^4 \to 0$ instead of converging to $c \in (0, \infty)$), so that the bias is asymptotically negligible.

Theorem 8 (Asymptotic distribution of $\hat{\pi}^*$)**.** *Let Assumptions 1, 2, either 3 or 3b, 4 and 5 hold and* $l^{-1}\sqrt{nh_R} \to 0$. *If* $h_R = h_{Rn} \to 0$, $h_T = h_{Tn} \to 0$, $nh_R h_T^3 \to \infty$, *and* $nh_R^5 \to 0$, $nh_R h_T^4 \to 0$, *then*

$$\frac{\hat{\pi}^* - \pi^*}{\sqrt{\mathsf{V}_{\pi,n}}} \to_d \mathcal{N}(0,1), \text{ where } \mathsf{V}_{\pi,n} \equiv \frac{\mathsf{V}_\pi}{nh_R}.$$

The exact form of V_π *is given by equation (B.4).*

The bandwidth conditions in Theorem 8 imply a bandwidth choice $h_R \sim h_T \sim n^{-a}$ for $a \in (1/5, 1/4)$. Based on Theorem 8, $\sqrt{nh_R}\left(\hat{\pi}^* - \pi^*\right) \to_d \mathcal{N}(0, \mathsf{V}_\pi)$, where V_π is the asymptotic variance of $\sqrt{nh_R}\hat{\pi}^*$.

The asymptotic distributions of $\hat{\tau}(u)$ and $\hat{\pi}^*$ presented here are valid only when the bandwidths shrink to zero fast enough with the sample size, which prevents overly large bandwidth choices, as are typical in empirical practice.

B.3 Proofs for Section B.1

The following proofs focus on $\hat{q}^+(u)$ and $\hat{m}^+(u)$ using observations above the RD threshold. Results for $\hat{q}^-(u)$ and $\hat{m}^-(u)$ can be analogously derived.

Proof of Lemma 3.

(Q) Proof for $\Delta\hat{q}(u)$**.** By Theorem 1.2 of Qu and Yoon (2015), we can show that the leading bias of $\hat{q}^+(u)$ with a small enough h_R is given by

$$\mathsf{B}_1^+(u) \equiv q_r''^+(u)\frac{1}{2}(1,0)\, N_{h_R}^{+\,-1} \int_{\mathcal{D}_{h_R}^+} v^2 (1,v)^\top K(v)dv, \text{ where}$$

$$N_{h_R}^+ \equiv \int_{\mathcal{D}_{h_R}^+} \begin{pmatrix} 1 & v \\ v & v^2 \end{pmatrix} K(v)dv = N_1 \equiv \begin{pmatrix} 1/2 & \kappa_1 \\ \kappa_1 & \kappa_2 \end{pmatrix}$$

and $\mathcal{D}_{h_R}^+ \equiv \left[0, (\bar{r} - r_0)/h_R\right) \cap Supp(K)$ if $\mathcal{R} = (\underline{r}, \bar{r})$. Note that $(1,0)N_1^{-1} = (2\kappa_2, -2\kappa_1, 0)/(\kappa_2 - 2\kappa_1^2)$, so $\mathsf{B}_1^+(u) \equiv C_{\mathsf{B}} q_r''^+(u)$.

By the Taylor expansion in Step 3 of the proof of Theorem 1 in Qu and Yoon (2015) and by their notation, $e_i(u) = -h_R^2 \frac{1}{2}\left(\frac{R_i - r_0}{h_R}\right)^2 \frac{\partial^2 q(u,r)}{\partial r^2} - h_R^3 \frac{1}{3!}\left(\frac{R_i - r_0}{h_R}\right)^3 \frac{\partial^3 q(u,r)}{\partial r^3} + o(h_R^3)$. Following the same arguments as those in their proof and assuming that $\partial^3 q(u,r)/\partial r^3$ is bounded, the second-order bias of $\hat{q}^+(u)$ is $O(h_R^3)$.

(M) Proof for $\Delta\hat{m}(u)$**.** Kong, Linton, and Xia (2010) provide a uniform Bahadur representation for the local polynomial regression that is uniform over the interior support of the regressors. In the following, we extend their results to the case when one of the regressors R is evaluated at the boundary point r_0.

Decompose $\hat{m}^+(u) - m^+(u) = \hat{m}^+(u) - \tilde{m}^+(u) + \tilde{m}^+(u) - m^+(u)$, where $\tilde{m}^+(u) = \hat{\mathbb{E}}\big[Y|T = q^+(u), R = r_0\big]$ is the infeasible estimator using the true $q^+(u)$. By Corollary 1 of Kong, Linton, and Xia (2010), the following asymptotically linear representation holds: $\tilde{m}^+(u) - m^+(u) - \mathbb{B}\left[\tilde{m}^+(u)\right] = n^{-1}\sum_{i=1}^n Z_i\phi_{2i}^+(u) + O_p\left((\log n/(nh_Rh_T))^{3/4}\right)$ uniformly over $u \in \mathcal{U}$, where the bias

$$\begin{aligned}
&\mathbb{B}\left[\tilde{m}^+(u)\right] \\
=&(1,0,0)S_1^{-1}Q_1\left(\frac{h_R^2}{2}m_r''^+(u), h_Rh_T\lim_{r\to r_0^+}\frac{\partial^2 m(t,r)}{\partial r\partial t}\Big|_{t=q^+(u)}, \frac{h_T^2}{2}m_t''^+(u)\right)^\top \\
&+ O(h_R^3 + h_T^3),
\end{aligned}$$

$$S_1 = \begin{pmatrix} 1/2 & \kappa_1 & 0 \\ \kappa_1 & \kappa_2 & 0 \\ 0 & 0 & \kappa_2 \end{pmatrix}, \text{and } Q_1 = \begin{pmatrix} \kappa_2 & 0 & \kappa_2 \\ \kappa_3 & 0 & 2\kappa_2\kappa_1 \\ 0 & 2\kappa_2\kappa_1 & 0 \end{pmatrix}.$$

Note $(1,0,0)S_1^{-1} = \big(2\kappa_2, -2\kappa_1, 0\big)/\big(\kappa_2 - 2\kappa_1^2\big)$ and $(1,0,0)S_1^{-1}Q_1 = 2(C_{\mathsf{B}}, 0, \kappa_2)$. Then $\mathbb{B}\big[\tilde{m}^+(u)\big] - \mathbb{B}\big[\tilde{m}^-(u)\big] = h_R^2\mathsf{B}_{R2}(u) + h_T^2\mathsf{B}_{T2}(u) + o(h_R^2 + h_T^2)$.

Applying Theorem 1 of Kong, Linton, and Xia (2010) and Lemma 3(Q), we have

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\left|\hat{m}^+(u) - \tilde{m}^+(u) - \big(\hat{q}^+(u) - q^+(u)\big)\frac{\partial}{\partial t}\mathbb{E}\left[Y|T = t, R = r_0\right]\Big|_{t=q^+(u)}\right| \\
&= O_p\Big(\Big(\sup_{u\in\mathcal{U}}\left|\hat{q}^+(u) - q^+(u)\right|\Big)^2 + \sup_{t\in\mathcal{T}_0}\Big(\Big|\frac{\partial}{\partial t}\hat{\mathbb{E}}\left[Y|T = t, R = r_0\right] \\
&\quad - \frac{\partial}{\partial t}\mathbb{E}\left[Y|T = t, R = r_0\right]\Big|\Big)\sup_{u\in\mathcal{U}}\left|\hat{q}^+(u) - q^+(u)\right|\Big) \\
&= O_p\left(\log n/(nh_R) + h_R^4 + \left(\left(\log n/\left(nh_Rh_T^3\right)\right)^{1/2} + h_R + h_T\right)\left((\log n/(nh_R))^{1/2} + h_R^2\right)\right) \\
&= O_p\left(\left(\left(\log n/\left(nh_Rh_T^3\right)\right)^{1/2} + h_R + h_T\right)\left((\log n/(nh_R))^{1/2} + h_R^2\right)\right),
\end{aligned}$$

where the compact set $\mathcal{T}_0 \subset \mathcal{T}$. We then obtain $\hat{m}^+(u) - m^+(u) - \mathbb{B}\big[\tilde{m}^+(u)\big] - h_R^2\mathsf{B}_1^+(u)m_t'^+(u) = n^{-1}\sum_{i=1}^n \Phi_{1i}^+(u)m_t'^+(u)Z_i + \phi_{2i}^+(u)Z_i + Rem$.

Proof of Lemma 4.

(I) From the proof of Lemma 3, $\|\Delta\hat{q} - \Delta q\|_\infty = O_p\left((\log n/(nh_R))^{1/2} + h_R^2\right)$, $\|\Delta\hat{m} - \Delta m\|_\infty = O_p\left((\log n/\left(nh_Rh_T\right))^{1/2} + h_R^2 + h_T^2\right)$, and uniformly over $u \in \mathcal{U}$,

$$\begin{aligned}
\hat{\tau}(u) - \tau(u) =& \frac{\Delta\hat{m}(u) - \Delta m(u)}{\Delta q(u)} - \frac{\tau(u)}{\Delta q(u)}\left(\Delta\hat{q}(u) - \Delta q(u)\right) \\
&+ O_p\left(\left\|\Delta\hat{m} - \Delta m\right\|_\infty\left\|\Delta\hat{q} - \Delta q\right\|_\infty\right).
\end{aligned}$$

By Lemma 3, we obtain the influence function $IF_{\tau i}(u)$ and the bias.

(D) Consider the asymptotic variance $\mathsf{V}_\tau(u)$. $\mathbb{E}\big[Z\big(u-\mathbf{1}\left(T \le q_1(R,u)\right)\big)\big|R\big]=0$, so $\mathbb{E}\big[Z_i\Phi_{1i}^+\big]=0$. Further, $\lim_{r\to r_0^+}\mathbb{E}\big[Y-\big(m^+(u)+m_r'^+(u)(R-r_0)+m_t'^+(u)(T-q^+(u))\big)\big|T=q^+(u),R=r\big]=0$, so we can show $\mathbb{E}\big[Z_i\phi_{2i}^+\big]=O(h)$. Then the sampling variation from $\hat{m}(u)$ in Step 2 contributes

$$
\begin{aligned}
&h_Rh_T\mathbb{V}\left[Z_i\phi_{2i}^+(u)\right]\\
=&\,h_Rh_T\mathbb{E}\Bigg[Z\mathbb{E}\Bigg[\left(Y-\left(m^+(u)+m_r'^+(u)\left(R-r_0\right)+m_t'^+(u)\left(T-q^+(u)\right)\right)\right)^2\Bigg|T,R\Bigg]\\
&\times\left(\frac{2\left(\kappa_2-\kappa_1(R-r_0)/h_R\right)}{f_{TR}^+(u)\left(\kappa_2-2\kappa_1^2\right)}\right)^2\frac{1}{h_T^2}K^2\left(\frac{T-q^+(u)}{h_T}\right)\frac{1}{h_R^2}K^2\left(\frac{R-r_0}{h_R}\right)\Bigg]+o(1)\\
=&2\lambda_0C_{\mathsf{V}}\frac{\sigma^{2+}(u)}{f_{TR}^+(u)}+o(1),
\end{aligned}
$$

where $C_{\mathsf{V}}=4\int_0^\infty\left(\kappa_2-\kappa_1\upsilon\right)^2K^2(\upsilon)d\upsilon/\left(\kappa_2-2\kappa_1^2\right)^2=4\left(\kappa_2^2\lambda_0-2\kappa_1\kappa_2\lambda_1+\kappa_1^2\lambda_2\right)\left(\kappa_2-2\kappa_1^2\right)^{-2}$. The sampling variation from $\Delta\hat{q}$ in Step 1 contributes

$$
\begin{aligned}
&h_Rh_T\mathbb{V}\left[Z_i\Phi_{1i}^+(u)\right]\\
=&h_Rh_T\mathbb{E}\Bigg[Z\mathbb{E}\left[\left(u-\mathbf{1}\left(T\le q_1(R,u)\right)\right)^2\Big|R\right]\left(\frac{2\left(\kappa_2-\kappa_1(R-r_0)/h_R\right)}{f_{TR}^+(u)\left(\kappa_2-2\kappa_1^2\right)}\right)^2\\
&\times\frac{1}{h_R^2}K^2\Big(\frac{R-r_0}{h_R}\Big)\Bigg]\\
=&h_TC_{\mathsf{V}}u(1-u)\frac{f_R^+(r_0)}{f_{TR}^{+2}(u)}+o(h_T)=O(h_T).
\end{aligned}
$$

Thus the sampling variation from the first step estimator $\Delta\hat{q}$ is of smaller order compared with the sampling variation from the second step estimator $\hat{m}(u)$. Therefore we obtain the asymptotic variance $\mathsf{V}_\tau(u)$.

For asymptotic normality, we apply Lyapounov CLT with the third absolute moment. The Lyapounov condition holds, i.e., $\left(\sum_{i=1}^n\mathbb{V}\big[IF_{\tau i}(u)\big]\right)^{-3/2}\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau i}(u)|^3\big]=O\big((nh_R^{-1}h_T^{-1})^{-3/2}\big)\times\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau i}(u)|^3\big]=O\big(\left(nh_Rh_T\right)^{-1/2}\big)=o(1)$. The bandwidth conditions guarantees $\sqrt{nh_Rh_T}Rem=o_p(1)$.

Proof of Lemma 5.

(I) The proof is for the estimator using the infeasible trimming, i.e., we use $\tilde{w}^*(u)\equiv\frac{|\Delta\hat{q}(u)|}{\int_{\mathcal{U}}|\Delta\hat{q}(u)|du}$ for $\mathcal{U}=\{u\in(0,1):|\Delta q(u)|>0\}$. Denote this infeasible estimator as $\tilde{\pi}\equiv\int_{\mathcal{U}}\hat{\tau}(u)\tilde{w}^*(u)du$. We show that as l goes to infinity and $n\to\infty$, $\hat{\pi}^*-\tilde{\pi}=o_p((nh_R)^{-1/2})$ at the end of the proof.

Let $w^*(u)\equiv\frac{|\Delta q(u)|}{\int_{\mathcal{U}}|\Delta q(u)|du}\equiv\frac{A(u)}{B}$ and $\tilde{w}^*(u)\equiv\frac{|\Delta\hat{q}(u)|}{\int_{\mathcal{U}}|\Delta\hat{q}(u)|du}\equiv\frac{\hat{A}(u_j)}{\hat{B}}$. A linear expansion

$\tilde{w}^*(u) - w^*(u) = \frac{\hat{A}(u)-A(u)}{B} - \frac{w^*(u)}{B}\big(\hat{B}-B\big) + O_p\big(\big\|\hat{A}-A\big\|_\infty\big|\hat{B}-B\big|\big) = O_p\big(\|\hat{q}-q\|_\infty\big) = O_p\big((\log n/(nh_R))^{1/2} + h_R^2\big)$. Then

$$\begin{aligned}\tilde{\pi} - \pi^* &= \int_{\mathcal{U}} \hat{\tau}(u)\hat{w}^*(u)du - \int_{\mathcal{U}} \tau(u)w^*(u)\,du \\ &= \int_{\mathcal{U}} \big(\hat{\tau}(u)-\tau(u)\big)\,w^*(u)\,du + \int_{\mathcal{U}} \tau(u)\,\big(\tilde{w}^*(u)-w^*(u)\big)\,du \\ &\quad + \int_{\mathcal{U}} \big(\hat{\tau}(u)-\tau(u)\big)\,\big(\tilde{w}^*(u)-w^*(u)\big)\,du,\end{aligned} \tag{B.14}$$

where the last term is $O_p\big(\big((\log n/(nh_Rh_T))^{1/2}+h_R^2+h_T^2\big)\big((\log n/(nh_R))^{1/2}+h_R^2\big)\big)$ by Lemma 4.

First consider the estimation error in the estimated weighting function $\tilde{w}^*(u)$ in equation (B.14). Let $\phi_{1i}(u) \equiv \phi_{1i}^+(u) - \phi_{1i}^-(u)$, where $\phi_{1i}^+(u) \equiv Z_i\Phi_{1i}^+(u) + h_R^2\mathsf{B}_1^+(u)$ and $\phi_{1i}^-(u) \equiv (1-Z_i)\Phi_{1i}^-(u) + h_R^2\mathsf{B}_1^-(u)$, so $\Delta\hat{q}(u) - \Delta q(u) = n^{-1}\sum_{i=1}^n \phi_{1i}(u) + O_p(h_R^3) + o_p((nh_R)^{-1/2})$. The absolute value function is Hadamard directionally differentiable.

By the delta method in Example 2.1 of Fang and Santos (2019), $\hat{A}(u) - A(u) = |\Delta\hat{q}(u)| - |\Delta q(u)| = n^{-1}\sum_{i=1}^n \phi_{1i}(u)\,(\mathbf{1}(\Delta q(u) > 0) - \mathbf{1}(\Delta q(u) < 0)) + O_p(h_R^3) + o_p\left((nh_R)^{-1/2}\right) = O_p\left((nh_R)^{-1/2} + h_R^2\right)$, since $\mathbf{1}(\Delta q(u) = 0) = 0$ for $u \in \mathcal{U}$. It follows that $\hat{B} - B = \int_{\mathcal{U}}\big(\hat{A}(u) - A(u)\big)du + o(1) = n^{-1}\sum_{i=1}^n \int_{\mathcal{U}} \phi_{1i}(u)\big(\mathbf{1}(\Delta q(u) > 0) - \mathbf{1}(\Delta q(u) < 0)\big)du + o_p\big((nh_R)^{-1/2}\big) = O_p\big((nh_R)^{-1/2} + h_R^2\big)$. Then

$$\begin{aligned}&\int_{\mathcal{U}} \tau(u)\,\big(\tilde{w}^*(u)-w^*(u)\big)\,du \\ &= \int_{\mathcal{U}} \frac{\tau(u)}{B}\left(\hat{A}(u)-A(u)\right)du - \frac{\pi^*}{B}\left(\hat{B}-B\right) \\ &\quad + O_p\left(\int_{\mathcal{U}} |\tau(u)|du\,\left\|\hat{A}-A\right\|_\infty\left|\hat{B}-B\right|\right) \\ &= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{U}} \left(\frac{\tau(u)}{B} - \frac{\pi^*}{B}\right)\phi_{1i}(u)\,(\mathbf{1}(\Delta q(u) > 0) - \mathbf{1}(\Delta q(u) < 0))\,du \\ &\quad + \ O_p\left(\log n/(nh_R) + h_R^4\right) + o_p\left((nh_R)^{-1/2}\right) \\ &= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{U}} \big(\tau(u)-\pi^*\big)\,\phi_{1i}(u)\frac{w^*(u)}{\Delta q(u)}du + O_p\left(\log n/(nh_R) + h_R^4\right) + o_p\left((nh_R)^{-1/2}\right)\end{aligned} \tag{B.15}$$

since $w^*(u)/\Delta q(u) = \big(\mathbf{1}(\Delta q(u) > 0) - \mathbf{1}(\Delta q(u) < 0)\big)/B$.

Next consider the first term in (B.14). Let $\mathsf{m}^+(v) \equiv \lim_{r\to r_0^+}\mathbb{E}\big[Y\big|T=v,\,R=r\big]$, $\mathsf{m}_r^+(v) \equiv \lim_{r\to r_0^+}\partial\mathbb{E}\big[Y\big|T=v,\,R=r\big]/\partial r$, and $\mathsf{m}_t^+(v) \equiv \lim_{r\to r_0^+}\partial\mathbb{E}\big[Y\big|T=t,\,R=r\big]/\partial t|_{T=v}$. By change of variable $v = q^+(u)$, $dv = du\partial q^+(u)/\partial u = du\,f_R(r_0)/f_{TR}^+(u)$. Then $\phi_{2i}^+(u)$ defined in

Lemma 3 becomes

$$\phi_{2i}^{+}(F_{T_1|R}(v,r_0)) = \left(Y_i - \left(\mathsf{m}^{+}(v) + \mathsf{m}_r'^{+}(v)\,(R_i - r_0) + \mathsf{m}_t'^{+}(v)\,(T_i - v)\right)\right) \times \frac{2\,(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_{TR}^{+}(F_{T_1|R}(v,r_0))\left(\kappa_2 - 2\kappa_1^2\right)} \frac{1}{h_T} K\left(\frac{T_i - v}{h_T}\right) K_{h_R}(R_i - r_0).$$

Let $\mathcal{U}^{+} \equiv \left[\underline{u}, \bar{u}\right] \subseteq \mathcal{U}$ such that $\Delta q(u) > 0$ for all $u \in \mathcal{U}^{+}$. Then

$$\begin{aligned}
&\int_{\mathcal{U}^{+}} \phi_{2i}^{+}(u) \frac{w^{*}(u)}{\Delta q(u)} du \\
&= \int_{q^{+}(\underline{u})}^{q^{+}(\bar{u})} \left(Y_i - \left(\mathsf{m}^{+}(v) + \mathsf{m}_r'^{+}(v)\,(R_i - r_0) + \mathsf{m}_t'^{+}(v)\,(T_i - v)\right)\right) \\
&\quad \times \frac{2\,(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_{TR}^{+}(F_{T_1|R}(v,r_0))\left(\kappa_2 - 2\kappa_1^2\right)} K_{h_T}(T_i - v)\, K_{h_R}(R_i - r_0)\, \frac{f_{TR}^{+}(F_{T_1|R}(v,r_0))}{f_R(r_0) B} dv \\
&= \int_{\frac{q^{+}(\underline{u}) - T_i}{h_T}}^{\frac{q^{+}(\bar{u}) - T_i}{h_T}} \left(Y_i - \left(\mathsf{m}^{+}(T_i + h_T s) + \mathsf{m}_r'^{+}(T_i + h_T s)\,(R_i - r_0)\right.\right. \\
&\quad \left.\left. - \mathsf{m}_t'^{+}(T_i + h_T s)\,(h_T s)\right)\right) K(s) ds \frac{2\,(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_R(r_0) B\left(\kappa_2 - 2\kappa_1^2\right)} K_{h_R}(R_i - r_0) \\
&= \Phi_{21i}^{+} + O_p(h_T).
\end{aligned}$$

The last equality follows by letting $\mathrm{U}_{zi} \equiv F_{T_z|R}(T_{zi}, r_0) \sim Unif(0,1)$ for $z \in \{0,1\}$. Thus $T_{1i} = q^{+}(\mathrm{U}_{1i})$ and $\mathsf{m}^{+}(T_{1i}) = m^{+}(\mathrm{U}_{1i})$. The same argument applies to $\mathcal{U}^{-}$, where $\Delta q(u) < 0$ for $u \in \mathcal{U}^{-}$. Then together with the influence function derived in Lemma 4, the first term in equation (B.14) is given by

$$\begin{aligned}
&\int_{\mathcal{U}} \left(\hat{\tau}(u) - \tau(u)\right) w^{*}(u) du \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(Z_i \Phi_{21i}^{+} - (1 - Z_i)\Phi_{2i}^{-}\right) \mathbf{1}\left(\mathrm{U}_i \in \mathcal{U}\right) + \int_{\mathcal{U}} \left(\phi_{1i}^{+}\left(m_t'^{+}(u) - \tau(u)\right)\right. \\
&\quad \left. - \phi_{1i}^{-}\left(m_t'^{-}(u) - \tau(u)\right) + h_R^2 \mathsf{B}_{R2}(u) + h_T^2 \mathsf{B}_{T2}(u)\right) \frac{w^{*}(u)}{\Delta q(u)} du + Rem.
\end{aligned}$$

Together with (B.15), we obtain the asymptotically linear representation for $\hat{\pi}^{*}$.

(D) The asymptotic variance V_{π} is derived using the influence function in Lemma 5(I),

$$\begin{aligned}
\mathsf{V}_{\pi} = \lim_{n \to \infty} h_R \mathbb{V}\Big[&\left(Z_i \Phi_{21i}^{+} - (1 - Z_i)\Phi_{21i}^{-}\right) \\
&+ \int_{\mathcal{U}} \left(Z_i \Phi_{1i}^{+}(u)\Lambda^{+}(u) - (1 - Z_i)\Phi_{1i}^{-}(u)\Lambda^{-}(u)\right) du\Big].
\end{aligned}$$

$\lim_{r\to r_0^+}\mathbb{E}\big[(Y-(m^+(\mathrm{U})+m_r^+(\mathrm{U})(R-r_0)))w^*(\mathrm{U})/\Delta q(\mathrm{U})\big|\mathrm{U}=F_{T_1|R}(T_1,r_0),R=r\big]=0$, so we can show $\mathbb{E}\big[Z_i\Phi_{21i}^+\big]=O(h_R)$ and $\mathbb{E}\big[Z_i\Phi_{21i}^+\int_{\mathcal{U}}Z_i\Phi_{1i}^+(u)\Lambda^+(u)du\big]=O(h_R)$. Then for V_π^q,

$$
\begin{aligned}
&\lim_{n\to\infty}h_R\mathbb{V}\left[\int_{\mathcal{U}}Z_i\Phi_{1i}^+(u)\Lambda^+(u)du\right]\\
&=\lim_{n\to\infty}h_R\int_{r_0}^{\infty}\int_{\mathcal{U}}\int_{\mathcal{U}}\mathbb{E}\left[\left(u-\mathbf{1}\left(T\le q^+(u)\right)\right)\left(v-\mathbf{1}\left(T\le q^+(v)\right)\right)|R\right]\\
&\quad\times\frac{\Lambda^+(u)}{f_{TR}^+(u)}du\frac{\Lambda^+(v)}{f_{TR}^+(v)}dv\left(\frac{2\left(\kappa_2-\kappa_1(R-r_0)/h_R\right)}{\left(\kappa_2-2\kappa_1^2\right)}\frac{1}{h_R}K\left(\frac{R-r_0}{h_R}\right)\right)^2f_R(R)dR.
\end{aligned}
$$

In the following derivation for V_π^m, we sometimes suppress the notation $+/-$ for simplicity. Let $\mathcal{T}=\big[\underline{t},\bar{t}\big],\mathcal{U}^+=\big[\underline{u},\bar{u}\big],\bar{q}\equiv q^+(\bar{u}),\underline{q}\equiv q^+(\underline{u})$, and $Q\equiv\bar{q}-\underline{q}$. The second equality below is by change of variable $s=(T-q(u))/h_T$. The fourth equality below is by change of variable $a=(q(u)-q(v))/h_T$, so $du=f_{T|R}^+(q(v)+ah_T)h_Tda$.

$$
\begin{aligned}
&h_R\mathbb{E}\left[\left(\int_{\mathcal{U}^+}\phi_{2i}^+(u)\frac{w^*(u)}{\Delta q(u)}du\right)^2\right]\\
&=\frac{C_{\mathsf{V}}}{f_R(r_0)B^2}\int_{\mathcal{U}^+}\int_{\mathcal{U}^+}\int_{\mathcal{T}}\mathbb{E}\big[\big(Y-\mathsf{m}^+(q(u))-\mathsf{m}_t'^+(q(u))\,(T-q(u))\big)\big(Y-\mathsf{m}^+(q(v))\\
&\quad-\mathsf{m}_t'^+(q(v))\,(T-q(v))\big)\big|T,R=r_0^+\big]K_{h_T}(T-q(u))K_{h_T}(T-q(v))\,f_{T|R}(T,r_0^+)dT\\
&\quad\times\frac{1}{f_{T|R}^+(u)\,f_{T|R}^+(v)}dudv\\
&=\frac{C_{\mathsf{V}}}{f_R(r_0)B^2}\int_{\mathcal{U}^+}\int_{\mathcal{U}^+}\int_{\frac{\underline{t}-q(u)}{h_T}}^{\frac{\bar{t}-q(u)}{h_T}}\mathbb{E}\big[\big(Y-\mathsf{m}^+(q(u))-\mathsf{m}_t'^+(q(u))sh_T\big)\big(Y-\mathsf{m}^+(q(v))\\
&\quad-\mathsf{m}_t'^+(q(v))\,(q(u)-q(v)+sh_T)\big)\big|T=q(u)+sh_T,R=r_0^+\big]K(s)\\
&\quad\times\frac{1}{h_T}K\left(\frac{q(u)-q(v)}{h_T}+s\right)f_{T|R}(q(u)+sh_T,r_0^+)ds\frac{dudv}{f_{T|R}^+(u)\,f_{T|R}^+(v)}\\
&=\int_{\mathcal{U}^+}\int_{\mathcal{U}^+}\int_{\frac{\underline{t}-q(u)}{h_T}}^{\frac{\bar{t}-q(u)}{h_T}}\mathbb{E}\big[\left(Y-\mathsf{m}^+(q(u))\right)\left(Y-\mathsf{m}^+(q(v))\right.\\
&\quad-\mathsf{m}_t'^+(q(v))\,(q(u)-q(v))\big)\big|T=q(u),R=r_0^+\big]K(s)\frac{1}{h_T}K\left(\frac{q(u)-q(v)}{h_T}+s\right)\\
&\quad\times f_{T|R}(q(u),r_0^+)ds\frac{dudv}{f_{T|R}^+(u)\,f_{T|R}^+(v)}\frac{C_{\mathsf{V}}}{f_R(r_0)B^2}+O(h_T)\\
&=\frac{C_{\mathsf{V}}}{f_R(r_0)B^2}\int_{\mathcal{U}^+}\int_{\frac{\underline{q}-q(v)}{h_T}}^{\frac{\bar{q}-q(v)}{h_T}}\int_{\frac{\underline{t}-q(v)}{h_T}-a}^{\frac{\bar{t}-q(v)}{h_T}-a}\mathbb{E}\big[\big(Y-\mathsf{m}^+(q(v)+ah_T)\big)\big(Y-\mathsf{m}^+(q(v))\\
&\quad-\mathsf{m}_t'^+(q(v))ah_T\big)\big|T=q(v)+ah_T,R=r_0^+\big]K(s)K\left(a+s\right)dsdadv+O(h_T)
\end{aligned}
$$

$$= \frac{C_{\mathsf{V}}}{f_R(r_0)B^2}\int_{\mathcal{U}^+}\int_{\frac{\underline{q}-q(v)}{h_T}}^{\frac{\bar{q}-q(v)}{h_T}}\int_{\frac{\underline{t}-q(v)}{h_T}-a}^{\frac{\bar{t}-q(v)}{h_T}-a} K(s)K(a+s)\,dsda\sigma^{2+}(v)dv + O(h_T)$$

$$= \frac{C_{\mathsf{V}}}{f_R(r_0)B^2}\int_{\mathcal{U}^+}\int_{\frac{\underline{q}-q(v)}{h_T}}^{\frac{\bar{q}-q(v)}{h_T}}\int_{\frac{\underline{t}-q(v)}{h_T}}^{\frac{\bar{t}-q(v)}{h_T}} K(s-a)K(s)\,dsda\sigma^{2+}(v)dv + O(h_T), \tag{B.16}$$

where the crude bound $O(h_T)$ comes from the integration over the sub-support $\mathcal{U}^+$. Let $G(u) \equiv \int_{-\infty}^{u} K(s)ds$. By integration by parts and change of variable,

$$\begin{aligned}
&\int_{\frac{\underline{q}-q(v)}{h_T}}^{\frac{\bar{q}-q(v)}{h_T}}\int_{\frac{\underline{t}-q(v)}{h_T}}^{\frac{\bar{t}-q(v)}{h_T}} K(s-a)K(s)\,dsda \\
&= \int_{\frac{\underline{t}-q(v)}{h_T}}^{\frac{\bar{t}-q(v)}{h_T}}\int_{\frac{\underline{q}-q(v)}{h_T}-s}^{\frac{\bar{q}-q(v)}{h_T}-s} K(a)daK(s)\,ds \\
&= \int_{\frac{\underline{t}-q(v)}{h_T}}^{\frac{\bar{t}-q(v)}{h_T}}\left(G\left(\frac{\bar{q}-q(v)}{h_T}-s\right)-G\left(\frac{\underline{q}-q(v)}{h_T}-s\right)\right)K(s)\,ds \\
&= \int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} G\left(\frac{\bar{q}-q(v)}{h_T}-s\right)K(s)ds - \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\underline{q}-\underline{t}}{h_T}} G\left(\frac{\underline{q}-q(v)}{h_T}-s\right)K(s)\,ds \\
&\quad + G\left(\frac{\bar{t}-q(v)}{h_T}\right)\left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right)-G\left(\frac{\underline{q}-\bar{t}}{h_T}\right)\right) \\
&\quad - G\left(\frac{\underline{t}-q(v)}{h_T}\right)\left(G\left(\frac{\bar{q}-\underline{t}}{h_T}\right)-G\left(\frac{\underline{q}-\underline{t}}{h_T}\right)\right).
\end{aligned}$$

Note in the first two terms in the above equation, the range of integration does not depend on v. So we can change the order of integrations in (B.16).

Let $\sigma^{2+}(v) \equiv \mathbb{E}\big[\left(Y-\mathsf{m}^+(T)\right)^2\big|T=q(v), R=r_0^+\big]$ and $\mathsf{V}(v) \equiv \int_0^v \sigma^{2+}(u)du$. We compute

$$\begin{aligned}
&\int_{\mathcal{U}^+}\sigma^{2+}(v)G\left(\frac{\bar{q}-q(v)}{h_T}-s\right)dv \\
&= \mathsf{V}(v)G\left(\frac{\bar{q}-q(v)}{h_T}-s\right)\Big|_{\underline{u}}^{\bar{u}} + \int_{\underline{u}}^{\bar{u}}\mathsf{V}(v)K\left(\frac{\bar{q}-q(v)}{h_T}-s\right)\frac{q'(v)}{h_T}dv \\
&= \mathsf{V}(\bar{u})G(-s) - \mathsf{V}(\underline{u})G(Q/h_T-s) + \int_{-Q/h_T}^{0}\mathsf{V}(F_{T|R}(\bar{q}+ah_T,r_0^+))K(s+a)da \\
&= \mathsf{V}(\bar{u})G(-s) - \mathsf{V}(\underline{u})G(Q/h_T-s) + \mathsf{V}(\bar{u})\left(G(s)-G(-Q/h_T+s)\right) + O(h_T) \\
&= \big(\mathsf{V}(\bar{u})-\mathsf{V}(\underline{u})\big)G(Q/h_T-s) + O(h_T),
\end{aligned}$$

where the second equality is by change of variable $a = (q(v)-\bar{q})/h_T$. Similarly

$\int_{\mathcal{U}^+} \sigma^{2+}(v) G\left(\frac{\underline{q}-q(v)}{h_T} - s\right) dv = \left(\mathsf{V}(\bar{u}) - \mathsf{V}(\underline{u})\right) G(-Q/h_T - s) + O(h_T)$. And

$$
\begin{aligned}
&\int_{\mathcal{U}^+} \sigma^{2+}(v) G\left(\frac{\bar{t}-q(v)}{h_T}\right) dv \\
&= \mathsf{V}(v) G\left(\frac{\bar{t}-q(v)}{h_T}\right)\Big|_{\underline{u}}^{\bar{u}} + \int_{\underline{u}}^{\bar{u}} \mathsf{V}(v) K\left(\frac{\bar{t}-q(v)}{h_T}\right) \frac{q'(v)}{h_T} dv \\
&= \mathsf{V}(\bar{u}) G\left(\frac{\bar{t}-\bar{q}}{h_T}\right) - \mathsf{V}(\underline{u}) G\left(\frac{\bar{t}-\underline{q}}{h_T}\right) + \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\bar{t}}{h_T}} \mathsf{V}(F_{T|R}(\bar{t}+ah_T, r_0^+)) K(a) da \\
&= \mathsf{V}(\bar{u}) G\left(\frac{\bar{t}-\bar{q}}{h_T}\right) - \mathsf{V}(\underline{u}) G\left(\frac{\bar{t}-\underline{q}}{h_T}\right) + \mathsf{V}(1)\left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right) - G\left(\frac{\underline{q}-\bar{t}}{h_T}\right)\right) + O(h_T),
\end{aligned}
$$

where the second equality is by change of variable $a = (q(v) - \bar{t})/h_T$. Similarly $\int_{\mathcal{U}^+} \sigma^{2+}(v) G\left(\frac{\underline{t}-q(v)}{h_T}\right) dv = \mathsf{V}(\bar{u}) G\left(\frac{\underline{t}-\bar{q}}{h_T}\right) - \mathsf{V}(\underline{u}) G\left(\frac{\underline{t}-\underline{q}}{h_T}\right) + O(h_T)$, because $\mathsf{V}(F_{T|R}(\underline{t}, r_0^+)) = \mathsf{V}(0) = 0$.

The main component in (B.16) becomes

$$
\begin{aligned}
&\int_{\mathcal{U}^+} \int_{\frac{\underline{q}-q(v)}{h_T}}^{\frac{\bar{q}-q(v)}{h_T}} \int_{\frac{\underline{t}-q(v)}{h_T}}^{\frac{\bar{t}-q(v)}{h_T}} K(s-a) K(s)\, ds da \sigma^{2+}(v) dv \\
&= \int_{\mathcal{U}^+} \Bigg\{ \int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} G\left(\frac{\bar{q}-q(v)}{h_T} - s\right) K(s) ds - \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\underline{q}-\underline{t}}{h_T}} G\left(\frac{\underline{q}-q(v)}{h_T} - s\right) K(s)\, ds \\
&\quad + G\left(\frac{\bar{t}-q(v)}{h_T}\right)\left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right) - G\left(\frac{\underline{q}-\bar{t}}{h_T}\right)\right) \\
&\quad - G\left(\frac{\underline{t}-q(v)}{h_T}\right)\left(G\left(\frac{\bar{q}-\underline{t}}{h_T}\right) - G\left(\frac{\underline{q}-\underline{t}}{h_T}\right)\right) \Bigg\} \sigma^{2+}(v) dv \\
&= \int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} \int_{\mathcal{U}^+} G\left(\frac{\bar{q}-q(v)}{h_T} - s\right) \sigma^{2+}(v) dv K(s) ds \\
&\quad - \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\underline{q}-\underline{t}}{h_T}} \int_{\mathcal{U}^+} G\left(\frac{\underline{q}-q(v)}{h_T} - s\right) \sigma^{2+}(v) dv K(s)\, ds \\
&\quad + \int_{\mathcal{U}^+} G\left(\frac{\bar{t}-q(v)}{h_T}\right) \sigma^{2+}(v) dv \left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right) - G\left(\frac{\underline{q}-\bar{t}}{h_T}\right)\right) \\
&\quad - \int_{\mathcal{U}^+} G\left(\frac{\underline{t}-q(v)}{h_T}\right) \sigma^{2+}(v) dv \left(G\left(\frac{\bar{q}-\underline{t}}{h_T}\right) - G\left(\frac{\underline{q}-\underline{t}}{h_T}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} \big(\mathsf{V}(\bar{u}) - \mathsf{V}(\underline{u})\big) G\Big(\frac{Q}{h_T} - s\Big) K(s) ds - \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\underline{q}-\underline{t}}{h_T}} \big(\mathsf{V}(\bar{u}) - \mathsf{V}(\underline{u})\big) G\Big(-\frac{Q}{h_T} - s\Big) K(s)\, ds \\
&\quad + \left\{ \mathsf{V}(\bar{u}) G\left(\frac{\bar{t}-\bar{q}}{h_T}\right) - \mathsf{V}(\underline{u}) G\left(\frac{\bar{t}-\underline{q}}{h_T}\right) \right\} \left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right) - G\left(\frac{\underline{q}-\bar{t}}{h_T}\right) \right) \\
&\quad + \mathsf{V}(1) \left(G\left(\frac{\bar{q}-\bar{t}}{h_T}\right) - G\left(\frac{\underline{q}-\bar{t}}{h_T}\right) \right)^2 \\
&\quad - \left\{ \mathsf{V}(\bar{u}) G\left(\frac{\underline{t}-\bar{q}}{h_T}\right) - \mathsf{V}(\underline{u}) G\left(\frac{\underline{t}-\underline{q}}{h_T}\right) \right\} \left(G\left(\frac{\bar{q}-\underline{t}}{h_T}\right) - G\left(\frac{\underline{q}-\underline{t}}{h_T}\right) \right) + O(h_T) \qquad \text{(B.17)}
\end{aligned}
$$

whose limit is $\mathsf{V}(\bar{u}) - \mathsf{V}(\underline{u}) = \int_{\mathcal{U}^+} \sigma^{2+}(u) du$ as $h_T \to 0$. Then we obtain V_π^m in (B.5).

Below we discuss that in finite samples the bandwidth h_T might not be small relative to $\bar{q} - \underline{q}$, $\bar{t} - \bar{q}$, and $\underline{q} - \underline{t}$. We suggest an adjustment term to estimate V_π^m in Section C. Let the support of the kernel K be $\big[-\bar{k}, \bar{k}\big]$. Let $\underline{h} \equiv Q/(2\bar{k})$. For $h_T > \underline{h}$, $s \in \big[-\bar{k}, \bar{k} - Q/h_T\big]$, and $\tilde{k} \in \big[Q/h_T + s, \bar{k}\big]$, $|G(Q/h_T + s) - G(\bar{k})| \le |K(\tilde{k})||\bar{k} - Q/h_T - s| = O(Qh_T/\underline{h}^2) = O(h_T/Q)$. So when $h_T > Q/(2\bar{k})$, equation (B.17) becomes $\int_{\underline{u}}^{\bar{u}} \sigma^2(u) du + O(h_T/Q)$. In finite samples, $O(h_T/Q)$ might not be ignorable. Thus in the first two terms in (B.17),

$$
\begin{aligned}
&\int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} G\Big(\frac{Q}{h_T} - s\Big) K(s) ds - \int_{\frac{\underline{q}-\bar{t}}{h_T}}^{\frac{\underline{q}-\underline{t}}{h_T}} G\Big(-\frac{Q}{h_T} - s\Big) K(s)\, ds \\
&= G\left(\frac{\bar{t}-\bar{q}}{h_T}\right) - G\left(\frac{\underline{t}-\bar{q}}{h_T}\right) - \int_{\frac{\bar{q}-\bar{t}}{h_T}}^{\frac{\bar{q}-\underline{t}}{h_T}} G\Big(s - \frac{Q}{h_T}\Big) K(s) ds - \int_{\frac{\underline{t}-\underline{q}}{h_T}}^{\frac{\bar{t}-\underline{q}}{h_T}} G\Big(s - \frac{Q}{h_T}\Big) K(s)\, ds. \qquad \text{(B.18)}
\end{aligned}
$$

The last three terms in (B.17) are of smaller order $o(h_T/\min\{\bar{t} - \bar{q}, \underline{q} - \underline{t}\})$.

To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. By the bandwidth conditions, the Lyapounov condition $\big(\sum_{i=1}^n \mathbb{V}\big[IF_{\pi i}\big]\big)^{-3/2} \sum_{i=1}^n \mathbb{E}\big[|IF_{\pi i}|^3\big] = O\big((nh^{-1})^{-3/2}\big) \sum_{i=1}^n \mathbb{E}\big[|IF_{\pi i}|^3\big] = O\big((nh_R)^{-1/2}\big) = o(1)$ holds.

Finally, we argue that as the number of grid points $l = l_n \to \infty$ and $l^{-1}\sqrt{nh_R} \to 0$, we can work with $\tilde{\pi}$ in the above proof by showing that $\hat{\pi}^* - \tilde{\pi} = o_p((nh_R)^{-1/2})$. Decompose $\hat{\pi}^* - \tilde{\pi}$ to

$$
\begin{aligned}
&l^{-1} \sum_{u_j \in \widehat{\mathcal{U}}} \hat{\tau}(u_j) \hat{\omega}(u_j) - l^{-1} \sum_{u_j \in \widehat{\mathcal{U}}} \hat{\tau}(u_j) \tilde{\omega}^*(u_j) \qquad \text{(B.19)} \\
&+ l^{-1} \sum_{u_j \in \widehat{\mathcal{U}}} \hat{\tau}(u_j) \tilde{\omega}^*(u_j) - \int_{\widehat{\mathcal{U}}} \hat{\tau}(u) \tilde{\omega}^*(u) du \qquad \text{(B.20)} \\
&+ \int_{\widehat{\mathcal{U}}} \hat{\tau}(u) \tilde{\omega}^*(u) du - \int_{\mathcal{U}} \hat{\tau}(u) \tilde{\omega}^*(u) du. \qquad \text{(B.21)}
\end{aligned}
$$

For the term (B.21), we argue that using the estimated trimming $\widehat{\mathcal{U}}$ is asymptotically equivalent to using the unknown $\mathcal{U}$. By Lemma 6, $\int_{\widehat{\mathcal{U}}} |\Delta\hat{q}(u)| du - \int_{\mathcal{U}} |\Delta\hat{q}(u)| du = \int_0^1 |\Delta\hat{q}(u)| (\hat{\chi}(u) -$

$\chi(u))du = o_p((nh_R)^{-1/2})$. The smoothness condition in Assumption 5.2 implies Lipschitz continuity, so $\tau(u)w^*(u)\int_{\mathcal{U}}|\Delta q(u)|du = O(|\Delta q(u)|)$. Thus $\left|\int_{\widehat{\mathcal{U}}}\hat{\tau}(u)\tilde{w}^*(u)du - \int_{\mathcal{U}}\hat{\tau}(u)\tilde{w}^*(u)du\right| = O_p\big(\int_0^1|\Delta\hat{q}(u)|(\hat{\chi}(u)-\chi(u))du\big) = o_p((nh_R)^{-1/2})$ by Lemma 6.

Next consider the term (B.20). Let $\hat{f}(u) \equiv \hat{\tau}(u)\tilde{w}^*(u)$ that is a smooth function of u. By a Taylor series expansion, the approximation error of the Riemann sum is $\left|l^{-1}\sum_{u_j\in\widehat{\mathcal{U}}}\hat{f}(u_j) - \int_{\widehat{\mathcal{U}}}\hat{f}(u)du\right| \leq l^{-1}\sum_{u_j\in\widehat{\mathcal{U}}}\left|\hat{f}(u_j)-\hat{f}(u_{j-1})\right| \leq l^{-1}\sum_{u_j\in\widehat{\mathcal{U}}}\left|\hat{f}'(\bar{u}_j)\right|(u_j-u_{j-1}) \leq l^{-1}\max_{u\in\widehat{\mathcal{U}}}\left|\hat{f}'(u)\right| = O_p(l^{-1})$, where $\bar{u}_j \in (u_{j-1},u_j)$. Therefore (B.20) is $O_p(l^{-1}) = o_p((nh_R)^{-1/2})$. The same arguments for (B.20) and (B.21) imply that (B.19) is $o_p((nh_R)^{-1/2})$.

Proof of Lemma 6. Rewrite

$$\begin{aligned}
&\hat{\chi}(u)-\chi(u)\\
&= \mathbf{1}\big(|\Delta\hat{q}(u)|>\epsilon_n, |\Delta q(u)|\leq 0\big) - \mathbf{1}\big(|\Delta\hat{q}(u)|\leq\epsilon_n, |\Delta q(u)|>0\big)\\
&= \mathbf{1}\big(|\Delta\hat{q}(u)|>\epsilon_n, |\Delta q(u)|\leq 0\big) - \mathbf{1}\big(|\Delta\hat{q}(u)|\leq\epsilon_n < 2\epsilon_n < |\Delta q(u)|\big) \qquad \text{(B.22)}\\
&\quad - \mathbf{1}\big(|\Delta\hat{q}(u)|\leq\epsilon_n, 0<|\Delta q(u)|\leq 2\epsilon_n\big). \qquad \text{(B.23)}
\end{aligned}$$

By the condition $\epsilon_n^{-1}\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big| = o_p(1)$, the first term in (B.22) $\mathbf{1}\big(|\Delta\hat{q}(u)|>\epsilon_n, |\Delta q(u)|\leq 0\big) \leq \mathbf{1}\big(\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|>\epsilon_n\big) = 0$ with probability approaching one (w.p.a.1) for any $u\in\mathcal{U}$. Thus $\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|\big)^{-1}\int_0^1|\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)|>\epsilon_n, |\Delta q(u)|\leq 0\big)du = 0$ w.p.a.1. It then implies that $\int_0^1|\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)|>\epsilon_n, |\Delta q(u)|\leq 0\big)du$ $= o_p\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|\big)$. The same argument applies to the second term in (B.22) and implies that $\int_0^1|\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)|\leq\epsilon_n<2\epsilon_n<|\Delta q(u)|\big)du = o_p\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|\big)$.

For the term in (B.23), note that $\int_0^1\mathbf{1}\big(0<|\Delta q(u)|\leq 2\epsilon_n\big)du = F(2\epsilon_n)$ denotes the CDF of $|\Delta q(U)|$ with $U\sim Unif(0,1)$. By the smoothness Assumption 5.1, we can apply a Taylor series expansion $F(2\epsilon_n) = F'(0)2\epsilon_n + o(\epsilon_n) = O(\epsilon_n)$. Therefore

$$\begin{aligned}
&\int_0^1|\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)|\leq\epsilon_n, 0<|\Delta q(u)|\leq 2\epsilon_n\big)du\\
&\leq \epsilon_n\int_0^1\mathbf{1}\big(0<|\Delta q(u)|\leq 2\epsilon_n\big)du = O(\epsilon_n^2) = o\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|\big)
\end{aligned}$$

by the condition $\epsilon_n^2\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)|-|\Delta q(u)|\big|\big)^{-1} = o_p(1)$. The result is then implied.

B.4 Proofs of Theorem 7, Theorem 3, and Theorem 4 for $\tau(u)$

Proof of Theorem 7. Lemma 4 implies Theorem 7 by letting the bias be of smaller order, i.e., $\sqrt{nh_Rh_T}\big(h_R^2\mathsf{B}_{R\tau}(u) - h_T^2\mathsf{B}_{T\tau}(u)\big) = o(1)$.

Proof of Theorem 3. The following derives the terms $\mathsf{V}_{\mathsf{B}_\tau}(u)$ and $\mathsf{C}_\tau(u;\rho)$ in the variance of $\hat{\tau}^{bc}(u)$, which are due to bias-correction.

Let $\mathsf{B}_2^{\pm}(u) \equiv c_R^2 C_{\mathsf{B}} m_r''^{\pm}(u) + c_T^2 \kappa_2 m_t''^{\pm}(u)$ and $\mathsf{B}_2(u) \equiv \mathsf{B}_2^{+}(u) - \mathsf{B}_2^{-}(u)$. For notational simplicity, we suppress the notation for u in the functions of u. Let $\widehat{\mathsf{B}}_\tau - \mathsf{B}_\tau \equiv \widehat{\mathsf{B}}_\tau^{+} - \mathsf{B}_\tau^{+} - \left(\widehat{\mathsf{B}}_\tau^{-} - \mathsf{B}_\tau^{-}\right)$, where B_τ is defined in (B.10). We linearize the estimator and focus on the part above the threshold: $\widehat{\mathsf{B}}_\tau^{+} - \mathsf{B}_\tau^{+} = \left\{\widehat{\mathsf{B}}_2^{+} - \mathsf{B}_2^{+} - c_R^2 \mathsf{B}_1^{+}\left(\hat{\tau} - \tau\right) + c_R^2\left(\widehat{\mathsf{B}}_1^{+} - \mathsf{B}_1^{+}\right)\left(m_t'^{+} - \tau\right)\right\} \big/ \Delta q + Rem_\tau$. Corollary 1 of Kong, Linton, and Xia (2010) for the local quadratic estimator implies the asymptotically linear representation for $\widehat{\mathsf{B}}_2^{+} - \mathsf{B}_2^{+}$ in (B.24) below and the convergence rates of the derivatives in $\widehat{\mathsf{B}}_2^{+}$: $\|\hat{m}_r''^{+} - m_r''^{+}\|_\infty = O_p\left(\left(\log n/(nb^6)\right)^{1/2} + b\right)$, $\|\hat{m}_t''^{+} - m_t''^{+}\|_\infty = O_p\left(\left(\log n/(nb^6)\right)^{1/2} + b\right)$, and $\|\hat{m}_t'^{+} - m_t'^{+}\|_\infty = O_p\left(\left(\log n/(nb^4)\right)^{1/2} + b^2\right)$. Lemma 3 in Qu and Yoon (2019) suggests $\|\hat{q}_r''^{+} - q_r''^{+}\|_\infty = O_p\left(\left(\log n/(nb^5)\right)^{1/2} + b\right)$. Thus it can be shown that the term associated with $\hat{q}_r''^{+}$ in $\widehat{\mathsf{B}}_1^{+}$ and the remainder terms Rem_τ are of smaller order.

$$
\begin{aligned}
\widehat{\mathsf{B}}_\tau^{+} - \mathsf{B}_\tau^{+} = {} & \frac{1}{\Delta q}\left\{b^{-2}\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^\top \beta_{n2}^{*+} + \mathbb{B}\left[\widehat{\mathsf{B}}_2^{+}\right]\right\} + O_p\left(\frac{1}{b^2}\left(\frac{\log n}{nb^2}\right)^{3/4}\right) \qquad \text{(B.24)} \\
& - \frac{\mathsf{B}_1^{+}}{\Delta q}\left(\hat{\tau} - \tau\right) + \frac{C_{\mathsf{B}} c_R^2}{\Delta q}\left(\hat{q}_r''^{+} - q_r''^{+}\right)\left(m_t'^{+} - \tau\right) + Rem_\tau \\
= {} & O_p\left(\left(\log n/(nb^6)\right)^{1/2} + b + \left(\log n/(nh_R h_T)\right)^{1/2} + h_R^2 + h_T^2\right),
\end{aligned}
$$

where $\mathbb{B}\left[\widehat{\mathsf{B}}_2^{+}\right] = O(b)$ and

$$
\beta_{n2}^{*+}(u) \equiv \frac{W_2 S_2^{-1} B_n^{-1}}{n f_{TR}^{+}(u)} \sum_{i=1}^{n} K_b(\underline{X}_i - \underline{x})\left(Y_i - \mu\left(\underline{X}_i - \underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X}_i - \underline{x}\right) Z_i,
$$

where $K_b(\underline{X}_i - \underline{x}) \equiv (b_R b_T)^{-1} K\left(\frac{T_i - q^{+}(u)}{b_T}\right) K\left(\frac{R_i - r_0}{b_R}\right)$, $W_2 \equiv diag\{1,1,1,2,1,2\}$, $B_n = diag\{1, b, b, b^2, b^2, b^2\}$, $\underline{X}_i \equiv (T_i/c_T, R_i/c_R)^\top$, $\underline{x} \equiv (q^{+}(u)/c_T, r_0/c_R)^\top$,
$\mu\left(\underline{X}\right) \equiv \left(1, R/c_R, T/c_T, R^2/c_R^2, RT/(c_R c_T), T^2/c_T^2\right)^\top$, and
$\beta_2(\underline{x}) \equiv \left(m^{+}, m_r'^{+} c_R, m_t'^{+} c_T, m_r''^{+} c_R^2, \lim_{r \to r_0^{+}} \frac{\partial^2 m(t,r)}{\partial r \partial t}\big|_{t = q^{+}(u)} c_R c_T, m_t''^{+} c_T^2\right)^\top$. β_{n2}^{*-} is defined as β_{n2}^{*+} by replacing Z_i with $1 - Z_i$ and $+$ with $-$.

Together with Lemma 4, the asymptotically linear representation for $\hat{\tau}^{bc}$ is

$$
\begin{aligned}
\hat{\tau}^{bc} - \tau = {} & \hat{\tau} - \tau - h^2\left(\widehat{\mathsf{B}}_\tau - \mathsf{B}_\tau\right) - h^2 \mathsf{B}_\tau \\
= {} & \frac{1}{n}\sum_{i=1}^{n} IF_{\tau^{bc} i} - h^2\Bigg(\frac{\mathbb{B}\left[\widehat{\mathsf{B}}_2^{+} - \widehat{\mathsf{B}}_2^{-}\right]}{\Delta q} - \frac{c_R^2(\mathsf{B}_1^{+} - \mathsf{B}_1^{-})}{\Delta q}\left(\hat{\tau} - \tau\right) \\
& + \left(\hat{q}_r''^{+} - q_r''^{+}\right)\frac{C_{\mathsf{B}} c_R^2}{\Delta q}\left(m_t'^{+} - \tau\right) - \left(\hat{q}_r''^{-} - q_r''^{-}\right)\frac{C_{\mathsf{B}} c_R^2}{\Delta q}\left(m_t'^{-} - \tau\right)\Bigg) \\
& + O_p\left(\frac{h^2}{b^2}\left(\frac{\log n}{nb^2}\right)^{3/4}\right) + Rem
\end{aligned}
$$

$$= \frac{1}{n}\sum_{i=1}^{n} IF_{\tau^{bc}i} + O_p\left(h^2 b + h^3 + \frac{h^2\sqrt{\log n}}{\sqrt{nb^5}} + \frac{h}{\sqrt{n}} + \frac{\log n}{\sqrt{n^2 h^5}} + (1+\rho^2)\left(\frac{\log n}{nb^2}\right)^{3/4}\right),$$

where the influence function

$$IF_{\tau^{bc}i} \equiv \frac{1}{\Delta q}\left\{ Z_i\left(\phi_{2i}^{+} + \Phi_{1i}^{+}\left(m_t'^{+} - \tau\right)\right) - \frac{h^2}{b^2}\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top}\beta_{n2}^{*+} \right.$$
$$\left. - (1 - Z_i)\left(\phi_{2i}^{-} + \Phi_{1i}^{-}\left(m_t'^{-} - \tau\right)\right) + \frac{h^2}{b^2}\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top}\beta_{n2}^{*-}\right\}. \tag{B.25}$$

Next we derive the asymptotic variance $\mathbb{V}\left[\beta_{n2}^{*+}\right]$ to be

$$\frac{W_2 S_2^{-1} B_n^{-1}}{n f_{TR}^{+2}} \mathbb{V}\left[K_b(\underline{X} - \underline{x})\left(Y_i - \mu\left(\underline{X} - \underline{x}\right)^{\top} W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X} - \underline{x}\right) Z_i\right] B_n^{-1} S_2^{-1} W_2,$$

where the second moment term

$$\begin{aligned}
&\mathbb{V}\left[K_b(\underline{X} - \underline{x})\left(Y - \mu\left(\underline{X} - \underline{x}\right)^{\top} W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X}_i - \underline{x}\right) Z_i\right] \\
=&\int_{\mathcal{T}}\int_{r_0}^{\infty} K_b^2(\underline{X} - \underline{x})\mathbb{E}\left[\left(Y - \mu\left(\underline{X} - \underline{x}\right)^{\top} W_2^{-1}\beta_2(\underline{x})\right)^2 \middle| R, T\right]\mu\left(\underline{X} - \underline{x}\right)\mu\left(\underline{X} - \underline{x}\right)^{\top} \\
&\times f_{TR}(T, R)dTdR \\
=&\int_{-\infty}^{\infty}\int_{0}^{\infty} K^2(v)K^2(s)\mathbb{E}\left[\left(Y - \mu\left(\underline{X} - \underline{x}\right)^{\top} W_2^{-1}\beta_2(\underline{x})\right)^2 \middle| T = q^{+} + b_T s,\, R = r_0 + b_R v\right] \\
&\times \mu((b_T s, b_R v)^{\top})\mu((b_T s, b_R v)^{\top})^{\top} f_{TR}(q^{+} + b_T s, r_0 + b_R v)dvds\frac{1}{b^2 c_T c_R} \\
=&\frac{2\lambda_0^2}{b^2 c_T c_R}\mathbb{V}\left[Y \middle| T = q^{+}, R = r_0\right] f_{TR}(q^{+}, r_0)e_1 e_1^{\top} + O\left(b^{-1}\right).
\end{aligned}$$

Therefore

$$\mathbb{V}\left[\beta_{n2}^{*+}\right] = \frac{2\lambda_0^2\sigma^{2+}}{nb^2 c_T c_R f_{TR}^{+}} W_2 S_2^{-1} e_1 e_1^{\top} S_2^{-1} W_2 + O\left((nb)^{-1}\right).$$

Thus the variance of $\widehat{\mathsf{B}}_{\tau}$ contributes to the asymptotic variance of $\hat{\tau}^{bc}$ by a term of order $\rho^4(nb^2)^{-1} = (nh^2\rho^{-6})^{-1}$. Since $\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top} W_2 = 2\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top}$, we obtain $\mathsf{V}_{\mathsf{B}_{\tau}}(u)$ defined in (B.2) by showing that the sample above the threshold contributes

$$\frac{\sigma^{2+}}{f_{TR}^{+}}\frac{1}{c_T c_R(\Delta q)^2} 8\lambda_0^2\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top} S_2^{-1} e_1 e_1^{\top} S_2^{-1}\left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right).$$

For the covariance term,

$$
\begin{aligned}
&\mathbb{C}\left[Z_i\phi_{2i}^{+}, \beta_{n2}^{*+}\right] \\
&= \frac{1}{n}\frac{2W_2S_2^{-1}B_n^{-1}}{f_{TR}^{+2}\left(\kappa_2 - 2\kappa_1^2\right)}\mathbb{E}\Bigg[K_h\left(T-q^{+}, R-r_0\right)K_b\left(T-q^{+}, R-r_0\right) \\
&\quad\times\left(Y-\left(m^{+}+m_r^{\prime+}\left(R-r_0\right)+m_t^{\prime+}\left(T-q^{+}\right)\right)\right)\left(Y-\mu\left(\underline{X}-\underline{x}\right)^{\top}W_2^{-1}\beta_2(\underline{x})\right) \\
&\quad\times\left(\kappa_2-\kappa_1(R-r_0)/h\right)\mu\left(\underline{X}-\underline{x}\right)Z\Bigg],
\end{aligned}
$$

where the expectation term is

$$
\begin{aligned}
&\int_{\mathcal{T}}\int_{r_0}^{\infty}K_h\left(T-q^{+}, R-r_0\right)K_b\left(T-q^{+}, R-r_0\right)\left(\kappa_2-\kappa_1(R-r_0)/h\right)\mu\left(\underline{X}-\underline{x}\right) \\
&\times\mathbb{E}\Bigg[\left(Y-\left(m^{+}+m_r^{\prime+}\left(R-r_0\right)+m_t^{\prime+}\left(T-q^{+}\right)\right)\right) \\
&\times\left(Y-\mu\left(\underline{X}-\underline{x}\right)^{\top}W_2^{-1}\beta_2(\underline{x})\right)\Bigg|T, R\Bigg]f_{TR}(T, R)dRdT \\
&= \frac{2\sigma^{2+}}{h^2c_Tc_R}\Bigg(\kappa_2\left(\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\right)^2 \\
&\quad-\frac{\kappa_1}{\rho}\int_0^{\infty}\upsilon K(\upsilon/\rho)K(\upsilon)d\upsilon\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\Bigg)e_1f_{TR}^{+}+O\left(bh^{-2}\right).
\end{aligned}
$$

Since $B_n^{-1}e_1 = e_1$, the covariance

$$
\begin{aligned}
&\mathbb{C}\left[Z_i\phi_{2i}^{+}, -\frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}\beta_{n2}^{*+}\right]\frac{1}{(\Delta q)^2} \\
&= -\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}\frac{1}{(\Delta q)^2}\rho^2\mathbb{C}\left[Z_i\phi_{2i}^{+}, \beta_{n2}^{*+}\right] \\
&= -\frac{1}{nb^2c_Tc_R}\frac{\sigma^{2+}}{f_{TR}^{+}}\frac{8\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}S_2^{-1}e_1}{\left(\kappa_2-2\kappa_1^2\right)(\Delta q)^2}\Bigg\{\kappa_2\left(\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\right)^2 \\
&\quad-\frac{\kappa_1}{\rho}\int_0^{\infty}\upsilon K(\upsilon/\rho)K(\upsilon)d\upsilon\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\Bigg\}+O\left((nb)^{-1}\right).
\end{aligned}
$$

A similar derivation yields

$$
\begin{aligned}
&\mathbb{C}\left[Z_i\Phi_{1i}^{+}\left(m_t^{\prime+}-\tau\right), -\frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}\beta_{n2}^{*+}\right]\frac{1}{(\Delta q)^2} \\
&= -\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}\frac{\left(m_t^{\prime+}-\tau\right)}{(\Delta q)^2}\rho^2\mathbb{C}\left[Z_i\Phi_{1i}^{+}, \beta_{n2}^{*+}\right]=o\left((nb^2)^{-1}\right).
\end{aligned}
$$

Thus the covariance between the $\widehat{\mathsf{B}}_\tau$ and $\widehat{\tau}$ contributes to the asymptotic variance of $\hat{\tau}^{bc}$ by a term

of order $(nb^2\rho)^{-1} = (nh^2\rho^{-1})^{-1}$. We obtain $\mathsf{C}_\tau(u;\rho)$ defined in (B.3) by showing that the sample above the threshold contributes

$$-\frac{\sigma^{2+}}{f_{TR}^{+}}\frac{16\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}e_1}{c_T c_R(\kappa_2-2\kappa_1^2)(\Delta q)^2}\int_0^\infty K(\upsilon/\rho)K(\upsilon)d\upsilon$$
$$\times\left(\rho\kappa_2\int_0^\infty K(\upsilon/\rho)K(\upsilon)d\upsilon-\kappa_1\int_0^\infty \upsilon K(\upsilon/\rho)K(\upsilon)d\upsilon\right).$$

Therefore $\mathsf{V}_\tau^{bc}=O\big((nh^2)^{-1}+h^4(nb^6)^{-1}\big)$ and $\mathbb{B}\big[\hat\tau^{bc}\big]=-h^2\mathbb{B}\big[\widehat{\mathsf{B}}_\tau\big]+O(h^3)=O\big(h^3+h^2b\big)$ is of smaller order by the conditions $n\min\{h^6,b^6\}\max\{h^2,b^2\}\to 0$. We have the asymptotically linear representation in (B.25), $\hat\tau^{bc}-\tau=n^{-1}\sum_{i=1}^n IF_{\tau^{bc}i}+o_p\big((nh^2)^{-1/2}+h^2(nb^6)^{-1/2}\big)$.

For asymptotic normality, we apply Lyapounov CLT with third absolute moment. When $h/b\to\rho\in(0,\infty)$, (B.25) implies $\sqrt{nh^2}\big(\hat\tau^{bc}-\tau-\mathbb{B}\big[\hat\tau^{bc}\big]\big)=\sqrt{nh^2}n^{-1}\sum_{i=1}^n IF_{\tau^{bc}i}+o_p(1)$. The Lyapounov condition holds, $\big(\sum_{i=1}^n\mathbb{V}\big[IF_{\tau^{bc}i}\big]\big)^{-3/2}\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big((nh^{-2})^{-3/2}\big)\times\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big(n^{-1/2}h^3\big(h^{-4}+\rho^6b^{-4}\big)\big)=O\big((nh^2)^{-1/2}\big)=o(1)$. Then $\sqrt{nh^2}\big(\hat\tau^{bc}(u;h,b)-\tau(u)\big)\to_d\mathcal{N}\big(0,\mathsf{V}_\tau^{bc}(u)\big)$.

When $h/b\to\infty$, $\sqrt{nb^6h^{-4}}\big(\hat\tau^{bc}-\tau-\mathbb{B}\big[\hat\tau^{bc}\big]\big)=\sqrt{nb^6h^{-4}}n^{-1}\sum_{i=1}^n IF_{\tau^{bc}i}+o_p(1)$. The Lyapounov condition holds, $\big(\sum_{i=1}^n\mathbb{V}\big[IF_{\tau^{bc}i}\big]\big)^{-3/2}\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big((nb^{-6}h^4)^{-3/2}\big)\sum_{i=1}^n\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big(n^{-1/2}b^9h^{-6}\rho^6b^{-4}\big)=O\big((nb^2)^{-1/2}\big)=o(1)$. Then $\sqrt{nb^6h^{-4}}\big(\hat\tau^{bc}(u;h,b)-\tau(u)\big)\to_d\mathcal{N}\big(0,\mathsf{V}_{\mathsf{B}_\tau}(u)\big)$.

Proof of Theorem 4. Theorem 4 follows by minimizing the AMSE implied by Lemma 4. The asymptotic distribution becomes $n^{1/3}\big(\hat\tau(u)-\tau(u)\big)\to_d\mathcal{N}\big(c_R^*(u)c_T^*(u)\mathsf{B}_\tau(u),(c_R^*(u)c_T^*(u))^{-1}\mathsf{V}_\tau(u)\big)$, where $c_R^*(u)=(\mathsf{V}_\tau(u)/8)^{1/6}(\mathsf{B}_{T\tau}(u)/\mathsf{B}_{R\tau}^5(u))^{1/12}$ and $c_T^*(u)=(\mathsf{V}_\tau(u)/8)^{1/6}(\mathsf{B}_{R\tau}(u)/\mathsf{B}_{T\tau}^5(u))^{1/12}$.

B.5 Proofs of Theorem 8, Theorem 5, and Theorem 6 for π^*

Proof of Theorem 8. Lemma 5 implies Theorem 8 by letting the bias be of smaller order, i.e., $\sqrt{nh_R}\big(h_R^2\mathsf{B}_{R\pi}+h_T^2\mathsf{B}_{T\pi}\big)=o(1)$.

Proof of Theorem 5. The following derives the terms $\mathsf{V}_{\mathsf{B}_\pi}$ and $\mathsf{C}_\pi(\rho)$ in the asymptotic variance of $\sqrt{nh_R}\hat\pi^{bc}$, which are due to bias correction.

Similar to the proof of Lemma 5, the proof below is for the estimator using the infeasible trimming function $\chi(u)$, denoted by $\tilde{\mathsf{B}}_\pi\equiv\int_{\mathcal{U}}\widehat{\mathsf{B}}_\tau(u)\tilde\omega^*(u)du+\int_{\mathcal{U}}\Big(\widehat{\mathsf{B}}_1^+(u)-\widehat{\mathsf{B}}_1^-(u)\Big)\big(\hat\tau(u)-\hat\pi\big)\tilde\omega^*(u)/\Delta\hat q(u)du$. Following the same arguments as in Lemma 6, we have $\tilde\pi^{bc}-\hat\pi^{bc}=o_p((nh)^{-1/2})$.

First derive the asymptotically linear representation

$$\hat\pi^{bc}-\pi^*=\frac{1}{n}\sum_{i=1}^n IF_{\pi^{bc}i}+o_p\left((nh)^{-1/2}+\rho^2(nb)^{-1/2}\right),$$

where the influence function

$$
\begin{aligned}
IF_{\pi^{bc}i} \equiv Z_i \Bigg\{ & \int_{\mathcal{U}} \Phi_{1i}^{+}(u)\Lambda^{+}(u)du - \rho^2 \left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top} \Phi_{22i}^{+}(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right) \\
& + \Phi_{21i}^{+}(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right) \Bigg\} - (1 - Z_i) \Bigg\{ \Phi_{21i}^{-}(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right) \\
& + \int_{\mathcal{U}} \Phi_{1i}^{-}(u)\Lambda^{-}(u)du - \rho^2 \left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^{\top} \Phi_{22i}^{-}(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right) \Bigg\}
\end{aligned} \tag{B.26}
$$

with $\Phi_{21i}^{\pm}(h)$ defined in Lemma 5 and

$$
\begin{aligned}
\Phi_{22i}^{\pm}(b) \equiv & \left(Y_i - \left(m^{\pm}(\mathrm{U}_i) + m_r'^{\pm}(\mathrm{U}_i)\left(R_i - r_0\right) + \frac{1}{2} m_r''^{\pm}(\mathrm{U}_i)\left(R_i - r_0\right)^2 \right) \right) \frac{w^*(\mathrm{U}_i)}{\Delta q(\mathrm{U}_i)} \\
& \times \frac{W_2 S_2^{-1}}{f_R(r_0)} \left(1, \frac{R_i - r_0}{b_R}, 0, \left(\frac{R_i - r_0}{b_R} \right)^2, 0, 0 \right)^{\top} \frac{1}{b_R} K\left(\frac{R_i - r_0}{b_R} \right).
\end{aligned}
$$

To derive $\Phi_{22i}^{\pm}(b)$, linearize $\hat{\mathsf{B}}_\pi - \mathsf{B}_\pi$, where B_π is defined in (B.13), to be

$$
\int_{\mathcal{U}} \left(\widehat{\mathsf{B}}_\tau(u) - \mathsf{B}_\tau(u) \right) w^*(u)du + \int_{\mathcal{U}} \left(\mathsf{B}_1^{+}(u) - \mathsf{B}_1^{-}(u) \right) \left(\hat{\tau}(u) - \tau(u) \right) \frac{w^*(u)}{\Delta q(u)} du + Rem_\pi .
$$

The leading term in Rem_π is $O_p(\|\widehat{\mathsf{B}}_\tau - \mathsf{B}_\tau\|_\infty \|\Delta\hat{q} - \Delta q\|_\infty) = O_p\big(\big(\big(\log n/(nb^6)\big)^{1/2} + b + \big(\log n/(nh_R h_T)\big)^{1/2} + h_R^2 + h_T^2\big)\big(\big(\log n/(nh_R)\big)^{1/2} + h_R^2\big)\big)$ by the proof of Theorem 3. And the terms associated with the cross products of $\widehat{\mathsf{B}}_1^{+} - \mathsf{B}_1^{+}$, $\Delta\hat{q} - \Delta q$, $\hat{\tau} - \tau$, and $\hat{\pi}^* - \pi^*$ in Rem_π are of smaller order.

Together with Lemma 4 and Lemma 5,

$$
\begin{aligned}
& \hat{\pi}^{bc} - \pi^* \\
= & \hat{\pi}^* - \pi^* - h^2 \mathsf{B}_\pi - h^2(\hat{\mathsf{B}}_\pi - \mathsf{B}_\pi) \\
= & \frac{1}{n}\sum_{i=1}^{n} IF_{\pi i} - h^2 \int_{\mathcal{U}} \left(\hat{\mathsf{B}}_\tau(u) - \mathsf{B}_\tau(u) \right) w^*(u)du \\
& - c_R^2 h^2 \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{U}} IF_{\tau i}(u) \left(\mathsf{B}_1^{+}(u) - \mathsf{B}_1^{-}(u) \right) \frac{w^*(u)}{\Delta q(u)} du \\
& - c_R^2 h^4 \int_{\mathcal{U}} \mathsf{B}_\tau(u) \left(\mathsf{B}_1^{+}(u) - \mathsf{B}_1^{-}(u) \right) \frac{w^*(u)}{\Delta q(u)} du + Rem + O_p\left(h^5 + h^2(Rem + Rem_\pi) \right).
\end{aligned}
$$

By the same argument in the proof of Lemma 5, the third term associated with $IF_{\tau i}(u)$ is $O_p\big(h_R^2\big((nh_R)^{-1/2} + h_R^2\big)\big)$, which is of smaller order. We focus on the second term $\int_{\mathcal{U}} \big(\widehat{\mathsf{B}}_\tau(u) -$

$\mathsf{B}_\tau(u)\big)w^*(u)du$ using the expansion in (B.24). One can show that

$$\int_{\mathcal{U}} \frac{w^*(u)}{\Delta q(u)}\left\{b^{-2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \beta_{n2}^{*+}(u)+\mathbb{B}\big[\widehat{\mathsf{B}}_2^+\big]-\mathsf{B}_1^+(u)\left(\hat{\tau}(u)-\tau(u)\right)\right\}du \tag{B.27}$$
$$= O_p\left((nb^5)^{-1/2}+b+(nh_R)^{-1/2}+h_R^2+h_T^2\right).$$

To see why, the second term associated with $\mathbb{B}\big[\widehat{\mathsf{B}}_2^+\big]$ is $O(b)$ and the third term associated with $\hat{\tau}-\tau$ is $O_p\left((nh_Rh_T)^{-1/2}+h_R^2+h_T^2\right)$ by the proof of Lemma 5 with the additional weight $\mathsf{B}_1^+(\mathrm{U}_i)/\Delta q(\mathrm{U}_i)$. For the first term in (B.27), we use the same arguments as those in deriving (B.14) in the proof of Lemma 5. By change of variable $v=q^+(u)$ and $s=(v-T_i)/b_T$, we have

$$\begin{aligned}
&\int_{\mathcal{U}} \frac{w^*(u)}{\Delta q(u)}\beta_{n2}^{*+}(u)du\\
&=\frac{W_2S_2^{-1}B_n^{-1}}{n}\sum_{i=1}^n\int_{\mathcal{U}}\frac{w^*(u)}{f_{TR}^+(u)\Delta q(u)}K_b(\underline{X}_i-\underline{x})\left(Y_i-\mu\left(\underline{X}_i-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\\
&\quad\times\mu\left(\underline{X}_i-\underline{x}\right)du\,Z_i\\
&=\frac{W_2S_2^{-1}B_n^{-1}}{n}\sum_{i=1}^n\int_{-\infty}^{\infty}\frac{w^*\left(F_{T_1|R}(T_i+b_Ts,r_0)\right)}{\Delta q\left(F_{T|R}(T_i+b_Ts,r_0)\right)}\frac{K(s)}{f_R(r_0)}\mathbf{1}\left(F_{T_1|R}(T_i+b_Ts,r_0)\in\mathcal{U}\right)\\
&\quad\times\left(Y_i-\mu\left((b_Ts,R_i-r_0)^\top\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left((b_Ts,R_i-r_0)^\top\right)ds\,Z_iK_b(R_i-r_0)\\
&=\frac{1}{n}\sum_{i=1}^n Z_i\Phi_{22i}^+(b)\mathbf{1}\left(\mathrm{U}_i\in\mathcal{U}\right)(1+O_p(b^2)).
\end{aligned}$$

For the asymptotic variance contributed by $\widehat{\mathsf{B}}_\pi$, $\mathsf{V}_{\mathsf{B}_\pi}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\Phi_{22i}^{+2}(b)\mathbf{1}\left(\mathrm{U}_i\in\mathcal{U}\right)Z_i\right]\\
&=W_2S_2^{-1}\mathbb{E}\Bigg[\left(Y-\left(m^+(\mathrm{U})+m_r'^+(\mathrm{U})\left(R-r_0\right)+\frac{1}{2}m_r''^+(\mathrm{U})\left(R-r_0\right)^2\right)\right)^2\\
&\quad\times\left(1,\frac{R-r_0}{b_R},0,\left(\frac{R-r_0}{b_R}\right)^2,0,0\right)^\top\left(1,\frac{R-r_0}{b_R},0,\left(\frac{R-r_0}{b_R}\right)^2,0,0\right)\\
&\quad\times\left(\frac{w^*(\mathrm{U})}{\Delta q(\mathrm{U})}\right)^2K_b^2\left(R-r_0\right)\mathbf{1}\left(\mathrm{U}\in\mathcal{U}\right)Z\Bigg]\frac{S_2^{-1}W_2}{f_R^2(r_0)}\\
&=W_2S_2^{-1}\int_0^\infty\int_{\mathcal{T}}\mathbf{1}\left(F_{T_1|R}(T|r_0)\in\mathcal{U}\right)\mathbf{v}^\top\mathbf{v}K^2\left(v\right)\mathbb{E}\Bigg[\Bigg(Y-\Bigg(m^+(\mathrm{U})+m_r'^+(\mathrm{U})(b_Rv)\\
&\quad+\frac{1}{2}m_r''^+(\mathrm{U})(b_Rv)^2\Bigg)\Bigg)^2\Bigg|\mathrm{U}=F_{T_1|R}(T|r_0),R=r_0+b_Rv\Bigg]f_{TR}(T,r_0+b_Rv)dTdv\\
&\quad\times\frac{S_2^{-1}W_2}{b_RB^2f_R^2(r_0)}
\end{aligned}$$

$$= \frac{\mathbb{E}\left[\mathbb{V}\left[Y\middle|\mathrm{U},\, R\right]\mathbf{1}\left(\mathrm{U}\in\mathcal{U}\right)\middle|R=r_0^+\right]}{b_R B^2 f_R(r_0)} W_2 S_2^{-1}\Lambda_2 S_2^{-1} W_2 + o(b^{-1}) = O(b^{-1}).$$

Thus the first term in (B.27) is $O_p((nb^5)^{-1/2})$. Then $\rho^2\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \Phi_{22i}^+(b)$ contributes to the asymptotic variance of $\hat{\pi}^{bc}$ by a term of order $\rho^4(nb)^{-1} = (nh\rho^{-5})^{-1}$. We obtain $\mathsf{V}_{\mathsf{B}_\pi}$ defined in (B.6) by showing that the sample above the cutoff contributes

$$\frac{4\int_{\mathcal{U}}\sigma^{2+}(u)du}{c_R B^2 f_R(r_0)}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}\Lambda_2 S_2^{-1}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right).$$

The asymptotic covariance is $\lim_{n\to\infty} -2h\rho^2\big(C_{\mathsf{B}}e_4+\kappa_2 e_6\big)^\top \mathbb{C}\big[Z_i\Phi_{21i}^+(h)\mathbf{1}\big(\mathrm{U}_i\in\mathcal{U}\big),\, Z_i\Phi_{22i}^+(b)\mathbf{1}\big(\mathrm{U}_i\in\mathcal{U}\big)\big] = \lim_{n\to\infty} -2h\rho^2\big(C_{\mathsf{B}}e_4+\kappa_2 e_6\big)^\top \mathbb{E}\big[Z_i\Phi_{21i}^+(h)\Phi_{22i}^+(b)\mathbf{1}\big(\mathrm{U}_i\in\mathcal{U}\big)\big]$, where

$$\begin{aligned}
&\mathbb{E}\left[Z_i\Phi_{21i}^+(h)\Phi_{22i}^+(b)\mathbf{1}\left(\mathrm{U}_i\in\mathcal{U}\right)\right]\\
&= \frac{2W_2S_2^{-1}}{B^2 f_R^2(r_0)\left(\kappa_2-2\kappa_1^2\right)}\mathbb{E}\Bigg[ZK_h\left(R-r_0\right)K_b\left(R-r_0\right)\left(Y-m^{\pm}(\mathrm{U})-m_r'^{\pm}(\mathrm{U})\left(R-r_0\right)\right)\\
&\quad\times\left(Y-m^{\pm}(\mathrm{U})-m_r'^{\pm}(\mathrm{U})\left(R-r_0\right)-\frac{m_r''^{\pm}(\mathrm{U})}{2}\left(R-r_0\right)^2\right)\\
&\quad\times\left(1,\frac{R-r_0}{b_R},0,\left(\frac{R-r_0}{b_R}\right)^2,0,0\right)^\top\left(\kappa_2-\kappa_1\frac{R-r_0}{h_R}\right)\mathbf{1}\left(\mathrm{U}\in\mathcal{U}\right)\Bigg].
\end{aligned}$$

By change of variable $\upsilon = (R-r_0)/b_R$, the above expectation term is

$$\begin{aligned}
&\frac{1}{\rho b}\int_0^\infty\int_{\mathcal{T}} K(\upsilon)K(\upsilon/\rho)\mathbb{V}\left[Y|\mathrm{U}=F_{T_1|R}(T,r_0),\, R=r_0+\upsilon b_R\right]\mathbf{v}\left(\kappa_2-\kappa_1\upsilon/\rho\right)\\
&\times\mathbf{1}\left(F_{T_1|R}(T,r_0)\in\mathcal{U}\right)\, f_{TR}(T,r_0+\upsilon b_R)dTd\upsilon = O\left((\rho b)^{-1}\right).
\end{aligned}$$

Thus the covariance between $\rho^2\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top\Phi_{22i}^+(b)$ and $\Phi_{21i}^+(h)$ contributes to the asymptotic variance of $\hat{\pi}^{bc}$ by a term of order $\rho^2(n\rho b)^{-1} = (nh\rho^{-2})^{-1}$. We obtain $\mathsf{C}_\pi(\rho)$ defined in (B.7) by showing that the sample above the cutoff contributes

$$-\frac{8\int_{\mathcal{U}}\sigma^{2+}(u)du}{B^2 f_R(r_0)\left(\kappa_2-2\kappa_1^2\right)}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}\int_0^\infty K(\upsilon)K(\upsilon/\rho)\mathbf{v}(\kappa_2-\kappa_1\upsilon/\rho)d\upsilon.$$

Therefore $\mathbb{V}\big[\hat{\pi}^{bc}\big] = O\big((nh)^{-1}+(nb^5h^{-4})^{-1}\big)$ and $\mathbb{B}\big[\hat{\pi}^{bc}\big] = O\big(h^2(h+b)\big)$ that is smaller-order by the bandwidth conditions $n\min\{h^5,b^5\}\max\{h^2,b^2\}\to 0$. To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. When $h/b\to\rho\in(0,\infty)$, (B.26) implies $\sqrt{nh}\big(\hat{\pi}^{bc}-\pi^*-\mathbb{B}\big[\hat{\pi}^{bc}\big]\big) = \sqrt{nh}n^{-1}\sum_{i=1}^n IF_{\pi^{bc}i}+o_p(1)$. The Lyapounov condition $\big(\sum_{i=1}^n\mathbb{V}\big[IF_{\pi^{bc}i}\big]\big)^{-3/2}\sum_{i=1}^n\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big] = O\big((nh^{-1})^{-3/2}\big)\sum_{i=1}^n\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big] = O\big(n^{-1/2}h^{3/2}h^{-2}\big) = O\big((nh)^{-1/2}\big) = o(1)$ holds. Then $\sqrt{nh}\big(\hat{\pi}^{bc}(h,b)-\pi^*\big)\to_d\mathcal{N}\big(0,\mathsf{V}_\pi^{bc}\big)$.

When $h/b\to\infty$, $\sqrt{nb^5h^{-4}}\big(\hat{\pi}^{bc}-\pi^*-\mathbb{B}\big[\hat{\pi}^{bc}\big]\big) = \sqrt{nb^5h^{-4}}n^{-1}\sum_{i=1}^n IF_{\pi^{bc}i}+o_p(1)$. The Lyapounov condition holds, $\big(\sum_{i=1}^n\mathbb{V}\big[IF_{\pi^{bc}i}\big]\big)^{-3/2}\sum_{i=1}^n\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big] = O\big((nb^{-5}h^4)^{-3/2}\big)\times$

$\sum_{i=1}^{n}\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big] = O\big(n^{-1/2}b^{15/2}h^{-6}\rho^6 b^{-2}\big) = O\big((nb)^{-1/2}\big) = o(1)$. Then $\sqrt{nb^5h^{-4}}\big(\hat{\pi}^{bc}(h,b) - \pi^*\big) \to_d \mathcal{N}\big(0, \mathsf{V}_{\mathsf{B}_\pi}\big)$.

Proof of Theorem 6. Theorem 6 follows by minimizing the AMSE implied by Lemma 5. The asymptotic distribution becomes $n^{2/5}\big(\hat{\pi}^* - \pi^*\big) \to_d \mathcal{N}\big(c_R^{*2}\mathsf{B}_{R\pi}, c_R^{*-1}\mathsf{V}_\pi\big)$, where $c_R^* \equiv \big(\mathsf{V}_\pi/\big(4\mathsf{B}_{R\pi}^2\big)\big)^{1/5}$.

C Estimation of the biases, variances, and AMSE optimal bandwidths

This section describes how to estimate the biases $\mathsf{B}_\tau(u)$ and B_π for $\widehat{\tau}(u)$ and $\hat{\pi}^*$, respectively, and the asymptotic variances $\mathsf{V}_\tau(u)$ and V_π for $\widehat{\tau}(u)$ and $\hat{\pi}^*$, respectively. We also describe how to estimate their associated AMSE optimal bandwidths $h_{R\tau}^*(u)$, $h_{T\tau}^*(u)$, $h_{R\pi}^*$, and $h_{T\pi}^{rot}$. With suitable choices of some preliminary bandwidths given below, these estimators for the biases, variances, and optimal bandwidths are consistent.

Consistency of each unknown element in these plug-in estimators requires standard bandwidth and regularity conditions. These conditions are well known in the literature (see, e.g., Kong, Linton, and Xia, 2010, Calonico, Cattaneo, and Titiunik, 2014, and Qu and Yoon, 2015) and are satisfied by our estimators. In the following we focus on estimating the unknown parameters defined above the RD cutoff. Corresponding parameters defined below the cutoff are estimated analogously.

C.1 Biases estimation

Consider the biases of $\widehat{\tau}(u)$, $\mathsf{B}_{R\tau}(u)$ and $\mathsf{B}_{T\tau}(u)$ in equations (B.8) and (B.9), respectively. C_B is a constant depending on the kernel function. For the Uniform kernel, $C_\mathsf{B} = -1/12$. $\Delta q(u)$ in the denominator of $\tau(u)$ is estimated in Step 1 of the estimation procedure described in the main text.

The remaining unknowns are $m_t'^{+}(u)$, $q_r''^{+}(u)$, $m_r''^{+}(u)$, and $m_t''^{+}(u)$. They can be estimated by local quadratic quantile and mean regressions. In particular, $q_r''^{+}(u)$ is estimated by $2\widehat{\alpha}_2$ from the local quadratic quantile regression with a chosen bandwidth b,

$$\begin{aligned}&(\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2)\\ = \;& \arg\min_{\alpha_0,\alpha_1,\alpha_2} \sum_{\{i:\, R_i \geq r_0\}} K\left(\frac{R_i - r_0}{b_R}\right)\rho_u\left(T_i - \alpha_0 - \alpha_1(R_i - r_0) - \alpha_2(R_i - r_0)^2\right).\end{aligned}$$

Further, $m_t'^{+}(u)$, $m_t''^{+}(u)$ and $m_r''^{+}(u)$ can be estimated by $\widehat{\beta}_{0,1}$, $2\widehat{\beta}_{0,2}$ and $2\widehat{\beta}_{2,2}$, respectively from the local quadratic regression

$$\begin{aligned}\big(\widehat{\beta}_{k,j}, k,j=0,1,2\big) \;=\; & \arg\min_{\beta_{k,j},k,j=0,1,2} \sum_{\{i\, R_i \geq r_0\}} K\left(\frac{R_i - r_0}{b_R}\right) K\left(\frac{T_i - \hat{q}^+(u)}{b_T}\right)\\ & \times\left(Y_i - \sum_{j=0}^{2}\sum_{k=0}^{j}\beta_{k,j}(R_i - r_0)^k\left(T_i - \hat{q}^+(u)\right)^{j-k}\right)^2.\end{aligned}$$

Plugging in C_{B} and the estimates of $m_t'^{\pm}(u)$, $q_r''^{\pm}(u)$, $m_r''^{\pm}(u)$, and $m_t''^{\pm}(u)$, one obtains $\widehat{\mathsf{B}}_{R\tau}(u)$ and $\widehat{\mathsf{B}}_{T\tau}(u)$. Then the bias of $\hat{\pi}^*$ is estimated by plugging in these estimates.

C.2 Variances estimation

For the standard error of $\widehat{\tau}(u)$, we estimate $\mathsf{V}_\tau(u)$ in equation (B.1). For the Uniform kernel, $C_{\mathsf{V}} = 4$ and $\lambda_0 = 1/4$. $\Delta q(u)$ is estimated by Step 1 estimation described in the main text. The remaining unknowns are c_R, c_T, $f^{\pm}_{T|R}(u)$, $f_R(r_0)$, and $\sigma^{2\pm}(u)$.

The densities $f^{+}_{T|R}(u)$ and $f_R(r_0)$ can be estimated by the standard Nadaraya-Watson estimator. That is, $\hat{f}^{\pm}_{T|R}(u) = \sum_{i=1}^n K\left(\frac{R_i - r_0}{g_R}\right) K\left(\frac{T_i - q^{\pm}(u)}{g_T}\right) Z_i \Big/ \sum_{i=1}^n K\left(\frac{R_i - r_0}{g_R}\right) Z_i$ and $\hat{f}_R(r_0) = (ng)^{-1} \sum_{i=1}^n K\left(\frac{R_i - r_0}{g}\right)$, where the Silverman-rule-of-thumb bandwidth for the Uniform kernel $g_R = 0.7344 \sigma_R n^{-1/6}$ and $g_T = 0.7344 \sigma_T n^{-1/6}$ for $\hat{f}^{\pm}_{T|R}(u)$ and $g = 1.843 \sigma_R n^{-1/5}$ for $\hat{f}_R(r_0)$. The standard deviations σ_R and σ_T are estimated directly by the sample standard deviations of R and T, respectively.

$\sigma^{2+}(u)$ can be estimated by $\widehat{\theta}_0$ from the local linear regression

$$\begin{aligned}\left(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2\right) = \arg\min_{\theta_0,\theta_1,\theta_2} \sum_{\{i\, R_i \geq r_0\}} & K\left(\frac{T_i - \hat{q}^{+}(u)}{b_T}\right) K\left(\frac{R_i - r_0}{b_R}\right) \\ & \times \left((Y_i - \hat{m}^{+}(u))^2 - \theta_0 - \theta_1 (R_i - r_0) - \theta_2 \left(T_i - \hat{q}^{+}(u)\right) \right)^2,\end{aligned}$$

where $\hat{m}^{+}(u)$ is estimated in Step 2 estimation described in the main text.

Plugging in all the estimates and the constants C_{V} and λ_0, one obtains $\widehat{\mathsf{V}}_\tau(u)$.

Consider next the standard error of the bias-corrected estimator $\hat{\tau}^{bc}(u)$. For the Uniform kernel, $\mathsf{V}_{\mathsf{B}_\tau}(u) = 9.765625 \mathsf{V}_\tau(u)$ by equation (B.2), and $\mathsf{C}_\tau(u;\rho) = 3.125 \rho^3 \mathsf{V}_\tau(u)$ when $\rho \leq 1$, and $\mathsf{C}_\tau(u;\rho) = 37.5(\rho/3 - 1/4)\mathsf{V}_\tau(u)$ when $\rho > 1$ by equation (B.3). Plugging in $\widehat{\mathsf{V}}_\tau(u)$ for a chosen ρ, one obtains $\widehat{\mathsf{V}}^{bc}_{\tau,n}(u)$.

For the standard error of $\hat{\pi}^*$, we estimate V_π in equation (B.4). Estimation of $\Delta q(u)$, $f_R(r_0)$, $f^{\pm}_{T|R}(u)$, and $\sigma^{2\pm}(u)$ is described at the beginning of this section. $\omega^*(u)$ is estimated in Step 4 estimation in the main text. The only unknown involved in V^q_π is $m_t'^{\pm}(u)$, which appears in $\Lambda^{\pm}(u)$ and can be estimated as described in Section C.1. We estimate $\Lambda^{\pm}(u)$ by plugging in the estimates of $\Delta q(u)$, $m_t'^{\pm}(u)$, and $\omega^*(u)$.

To estimate V^m_π, we include a finite-sample adjustment term in (B.18). Suppose $\mathcal{U} = \cup_{j=1}^J \mathcal{U}_j$ is a union of J disjoint intervals, where $\Delta q(u) \neq 0$ for $u \in \mathcal{U}_j \equiv \left[\underline{u}_j, \bar{u}_j\right]$. Let $\bar{q}^+_j \equiv q^+(\bar{u}_j)$, $\underline{q}^+_j \equiv q^+(\underline{u}_j)$, $Q^+_j \equiv \bar{q}^+_j - \underline{q}^+_j$, and the support of T_1 be $\mathcal{T}^+ \equiv \left[\underline{t}^+, \bar{t}^+\right]$. Then $\widehat{\mathsf{V}}^m_\pi$ is the estimate of

$$\sum_{j=1}^J \left(\mathsf{A}^+_j \int_{\mathcal{U}_j} \sigma^{2+}(u) du + \mathsf{A}^-_j \int_{\mathcal{U}_j} \sigma^{2-}(u) du \right) \frac{C_{\mathsf{V}}}{f_R(r_0)(\int_{\mathcal{U}} |\Delta q(u)| du)^2},$$

where

$$\mathsf{A}_j^+ \equiv G\left(\frac{\bar{t}^+ - \bar{q}_j^+}{h_T}\right) - G\left(\frac{\underline{t}^+ - \bar{q}_j^+}{h_T}\right) - \int_{\frac{\bar{q}_j^+ - \underline{t}^+}{h_T}}^{\frac{\bar{q}_j^+ - \underline{t}^+}{h_T}} G\left(s - \frac{Q_j^+}{h_T}\right) K(s)ds - \int_{\frac{\underline{t}^+ - \underline{q}_j^+}{h_T}}^{\frac{\bar{t}^+ - \underline{q}_j^+}{h_T}} G\left(s - \frac{Q_j^+}{h_T}\right) K(s)\, ds,$$

$G(u) \equiv \int_{-\infty}^{u} K(s)ds$, and A_j^- is defined analogously by changing $+$ to $-$. Note that when h_T is small relative to $Q_j^{\pm}$, the last two terms in $\mathsf{A}_j^{\pm}$ are zero. As $h_T \to 0$, $\mathsf{A}_j^{\pm}$ becomes 1 and equation (**??**) becomes V_π^m. The adjustment term is especially relevant when policies target top or bottom of the treatment distribution and hence $\bar{t}^+ - \bar{q}_j^+$ or $\underline{t}^+ - \underline{q}_j^+$ could be small relative to h_T.

Further plugging in the estimates of $\Delta q(u)$, $f_R(r_0)$, $f_{T|R}^{\pm}(u)$, $\Lambda^{\pm}(u)$, $\sigma^{2\pm}(u)$, and the constant C_{V}, and replacing integration by summation, we obtain $\widehat{\mathsf{V}}_\pi = \widehat{\mathsf{V}}_\pi^m + \widehat{\mathsf{V}}_\pi^q$.

Consider lastly the standard error of the bias-corrected estimator $\hat{\pi}^{bc}$ by Theorem 5, For the Uniform kernel, $\mathsf{V}_{\mathsf{B}_\pi} = 1.641\mathsf{V}_\pi^m$ by equation (B.6) and $\mathsf{C}_\pi(\rho) = (3.125\rho - 2.5\rho^3)\mathsf{V}_\pi^m$ when $\rho \leq 1$, and $\mathsf{C}_\pi(\rho) = (2.5 - 1.875/\rho)\mathsf{V}_\pi^m$ when $\rho > 1$ by equation (B.7). Plugging in the estimates of V_π and the constant $\mathsf{C}_\pi(\rho)$, one obtain $\widehat{\mathsf{V}}_{\pi,n}^{bc}$.

C.3 Optimal bandwidths estimation

Given the consistent estimates in the previous sections, by the plug-in rule, one can consistently estimate the AMSE optimal bandwidths for $\hat{\tau}(u)$ by $\hat{h}_{R\tau}^*(u) = \hat{c}_R^*(u)n^{-1/6}$ and $\hat{h}_{T\tau}^*(u) = \hat{c}_T^*(u)n^{-1/6}$, where $\hat{c}_R^*(u) = (\widehat{\mathsf{V}}_\tau(u)/8)^{1/6}(\widehat{\mathsf{B}}_{T\tau}(u)/\widehat{\mathsf{B}}_{R\tau}^5(u))^{1/12}$ and $\hat{c}_T^*(u) = (\widehat{\mathsf{V}}_\tau(u)/8)^{1/6}(\widehat{\mathsf{B}}_{R\tau}(u)/\widehat{\mathsf{B}}_{T\tau}^5(u))^{1/12}$. For $\hat{\pi}^*$, $\hat{h}_{R\pi}^* = (\widehat{\mathsf{V}}_\pi/(4\widehat{\mathsf{B}}_{R\pi}^2))^{1/5}n^{-1/5}$ and $\hat{h}_T^{rot} = \hat{h}_{R\pi}^* n^{-1/30}\hat{\sigma}_T/\hat{\sigma}_R$.

For our empirical analysis, we choose $b = cn^{-1/8}$, where $1/8$ is the optimal rate for estimating the second order derivatives of $m(t,r)$ by local quadratic regressions, and the constant c is set to be a value (4.5) such that the estimates are stable.

D Supplementary empirical analysis

D.1 Data description

Our data are collected from three sources: the annual reports of the Office of the Comptroller of the Currency (OCC), Rand McNally's Bankers Directory, and the United States population census.

The OCC's annual report includes the detailed balance sheet information. We collect balance sheet data on national banks in 1905 and whether they suspended their operation in the following 24 years (up to 1929). Our analysis focuses on national banks that were established after 1900. The minimum capital requirement changed in 1900. Before 1900, national banks were required to have a minimum capital of $50,000 regardless of whether they operated in a town above or below the 3,000 population threshold. National banks established before 1900 might be subject to either the old or new regulatory regime, depending on when they were rechartered. We do not have the information on when they were rechartered.

The OCC's annual report also indicates the town, county, and state in which each bank was located. We match this information with the United States Population Census to determine town populations. Since all banks in our sample were established between 1900 and 1905, their capital requirement in 1905 was determined by their town population in 1900, as reported by the 1900 census. In addition, we gather information on county characteristics that measure their business and agricultural conditions, including the percentage of black population, the percentage of farmland, and manufacturing output per capita per square miles.

Bank capital is the sum of a bank's capital and surplus. Bank assets refers to a bank's total amount of assets. Leverage is defined as the ratio of a bank's total assets to capital. Higher leverage is associated with lower survival rates during financial crises. However, banks generally have an incentive to increase their leverage so they can accumulate higher rates of returns on their capital.

Table D1.1 Sample summary statistics

	Z=0		Z=1			
	N	Mean (SD)	N	Mean (SD)	Difference	(SE)
Log(capital)	717	10.5 (0.40)	105	11.2 (0.39)	0.66	(0.04)***
Log(assets)	717	11.7 (0.53)	105	12.5 (0.54)	0.77	(0.06)***
Log(leverage)	717	1.19 (0.34)	105	1.30 (0.34)	0.11	(0.04)***
Suspension	717	0.10 (0.30)	105	0.06 (0.23)	-0.04	(0.03)
Bank age	717	2.45 (1.07)	105	2.78 (1.03)	0.33	(0.11)**
Black population (%)	674	0.07 (0.16)	101	0.08 (0.15)	0.01	(0.02)
Farmland (%)	674	0.77 (0.25)	101	0.71 (0.27)	-0.06	(0.03)**
Log(manufacturing output)	672	3.73 (1.11)	101	4.39 (0.96)	0.66	(0.12)***

Note: The sample consists of all national banks established between 1900 and 1905 and located in towns with a town population less than 6,000; ***Significant at the 1% level, **Significant at the 5% level

Brief sample summary statistics are provided in Table D1.1. Banks operating in towns with 3,000 people or more have more capital on average; they also hold more assets and have higher measured leverages. However, these simple correlations may not reflect the true causal relationships. As we can see, towns with more than 3,000 people are associated with older banks, a lower percentage of farm land in their counties, and higher manufacturing output per capita. These results highlight the importance to seek for local identification. Causal relationships would be confounded if one compares banks far away from the regulation threshold.

D.2 Additional estimation results

Table D2.1 presents estimates using a bandwidth that satisfies the undersmoothing conditions in Theorems 7 and 8. These estimates remain similar to those bias-corrected estimates reported in the main text. Note that the bias-corrected estimates use larger bandwidths, so there is no loss of precision compared with estimates by undersmoothing.

As a convenient alternative to computing the analytic standard errors, one may use the standard nonparametric bootstrap based on drawing n observations with replacement to obtain standard errors and confidence intervals. The bootstrap is valid for the bias corrected Q-LATE estimator by the Delta method. Tables D2.2 presents estimates with bootstrapped standard errors. These

bootstrapped standard errors are similar to the analytic standard errors reported in the main text. Tables D2.3 further presents estimates with bootstrapped standard errors that are clustered at the town level. Clustering does not have a big impact. Our main conclusions remain the same.

Table D2.1 Impacts of log(capital) on bank outcomes (undersmoothing)

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	0.954	(0.260)***	-0.046	(0.234)	0.018	(0.143)
	0.12	0.922	(0.236)***	-0.078	(0.218)	0.005	(0.125)
	0.14	0.896	(0.225)***	-0.104	(0.209)	0.007	(0.121)
	0.16	0.871	(0.245)***	-0.129	(0.228)	0.007	(0.133)
	0.18	0.760	(0.294)***	-0.240	(0.255)	-0.028	(0.149)
	0.20	0.769	(0.293)***	-0.231	(0.252)	-0.031	(0.149)
	0.22	0.766	(0.295)***	-0.234	(0.254)	-0.030	(0.148)
	0.24	0.805	(0.284)***	-0.195	(0.244)	-0.043	(0.141)
	0.26	0.808	(0.270)***	-0.192	(0.231)	-0.041	(0.132)
	0.28	0.814	(0.272)***	-0.186	(0.233)	-0.048	(0.131)
WQ-LATE		0.836	(0.397)**	-0.164	(0.375)	-0.016	(0.150)

Note: The top panel presents estimated Q-LATEs; The last row presents the estimated WQ-LATEs; $h_R = 1,150$ and $h_T = 0.441$ for all estimation, which satisfy the undersmoothing conditions for the Q-LATE and WQ-LATE estimators in Theorems 7 and 8; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 862.5$; Standard errors are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

Table D2.2 Impacts of log (capital) on bank outcomes with bootstrapped standard errors

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	1.030	(0.376)***	0.030	(0.376)***	-0.131	(0.195)
	0.12	1.062	(0.344)***	0.062	(0.344)***	-0.134	(0.181)
	0.14	1.050	(0.327)***	0.050	(0.327)	-0.140	(0.180)
	0.16	1.035	(0.315)***	0.035	(0.315)	-0.141	(0.181)
	0.18	0.911	(0.304)**	-0.089	(0.304)	-0.161	(0.178)
	0.20	0.955	(0.315)***	-0.045	(0.315)	-0.162	(0.183)
	0.22	0.972	(0.307)***	-0.028	(0.307)	-0.165	(0.186)
	0.24	1.014	(0.310)***	0.014	(0.310)	-0.168	(0.188)
	0.26	1.093	(0.324)***	0.093	(0.324)	-0.167	(0.195)
	0.28	1.087	(0.338)***	0.087	(0.338)	-0.165	(0.206)
	0.30	1.100	(0.363)***	0.100	(0.363)	-0.168	(0.214)
WQ-LATE		1.034	(0.291)***	0.034	(0.291)	-0.155	(0.175)

Note: The top panel presents the bias-corrected estimates of Q-LATEs; The last row presents the bias-corrected estimates of WQ-LATEs; For all estimation, $h_R = 1,462.76$, which is the AMSE optimal bandwidth for the WQ-LATE estimator (The AMSE optimal bandwidth for the Q-LATE estimator h_R ranges from $1,158.15$ to $1,303.90$), $h_T = 0.441$ and $\rho = 0.618$; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 1,097.07$; Bootstrapped standard errors (in the parentheses) are based on 500 replications; ***Significant at the 1% level, **Significant at the 5% level.

Table D2.3 Impacts of log (capital) on bank outcomes with bootstrapped clustered standard errors

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	1.030	(0.358)***	0.030	(0.358)	-0.131	(0.191)
	0.12	1.062	(0.318)***	0.062	(0.318)	-0.134	(0.186)
	0.14	1.050	(0.314)***	0.050	(0.314)	-0.140	(0.186)
	0.16	1.035	(0.307)***	0.035	(0.307)	-0.141	(0.190)
	0.18	0.911	(0.311)**	-0.089	(0.311)	-0.161	(0.191)
	0.20	0.955	(0.321)***	-0.045	(0.321)	-0.162	(0.192)
	0.22	0.972	(0.309)***	-0.028	(0.309)	-0.165	(0.195)
	0.24	1.014	(0.315)***	0.014	(0.315)	-0.168	(0.195)
	0.26	1.093	(0.331)***	0.093	(0.331)	-0.167	(0.199)
	0.28	1.087	(0.353)***	0.087	(0.353)	-0.165	(0.217)
	0.30	1.100	(0.371)***	0.100	(0.371)	-0.168	(0.230)
WQ-LATE		1.034	(0.288)***	0.034	(0.288)	-0.155	(0.185)

Note: The top panel presents the bias-corrected estimates of Q-LATEs; The last row presents the bias-corrected estimates of WQ-LATEs; For all estimation, $h_R = 1,462.76$, $h_T = 0.441$ and $\rho = 0.618$; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 1,097.07$; Bootstrapped standard errors (in the parentheses) are based on 500 replications; ***Significant at the 1% level, **Significant at the 5% level.

D.3 Additional validity checks

Table D3.1 presents the bias-corrected estimates of WQ-LATEs on the first two raw moments of covariates, which serve as joint specification tests. None of these estimates are statistically significant.

Table D3.1 Tests for local rank invariance or rank similarity

	First moment		Second moment	
Bank age	0.879	(0.731)	3.653	(3.860)
Black Population (%)	0.015	(0.128)	0.012	(0.079)
Farmland (%)	-0.010	(0.156)	0.052	(0.195)
Log(manufacturing output)	0.511	(0.669)	3.798	(5.881)

Note: Bias-corrected estimates of WQ-LATEs are reported; For all estimation, $h_R = 1,462.76$, $h_T = 0.441$ and $\rho = 0.618$; The trimming thresholds are determined by using a preliminary bandwidth $3/4h_R = 1097.07$. Standard errors are in the parentheses.

Table D3.2 presents the formal testing results for smoothness of the density of town population and smoothness of the conditional means of pre-determined covariates near the policy threshold. For details of various RD density tests, see, e.g., McCrary (2008), Cattaneo, Frandsen, and Titiunik (2015), and Cattaneo, Jansson, and Ma (2020).

Table D3.2 Tests for smoothness of covariates and density

I: Covariate					
Bank age	0.218	(0.350)	Farmland (%)	0.005	(0.110)
Black Population (%)	-0.096	(0.084)	Log(manufacturing output)	0.342	(0.382)
II: Density of town population					
		-0.592	(0.554)		

Note: Panel I presents the estimated discontinuities in the conditional means of covariate; Robust standard errors are in parentheses; Panel II presents the t statistic of the estimated density discontinuity of town population along with the p-value using rddensity; $h_R = 1,462.76$ for all estimation.

References

[1] Angrist, J. D., G. Imbens, and K. Graddy, (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," The Review of Economic Studies, 67, 499-527.

[2] Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Bias Corrected Inference in Regression Discontinuity Design," Econometrica, 82(6), 2295-2326.

[3] Cattaneo, M. D., B. Frandsen, and R. Titiunik (2015): "Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate, " Journal of Causal Inference, 3(1), 1-24.

[4] Cattaneo, M. D., M. Jansson, and X. Ma (2020): "Simple Local Polynomial Density Estimators." Journal of the American Statistical Association, 115(531), 1449-1455.

[5] Fang, Z. and A. Santos (2019): "Inference on Directionally Differentiable Functions," The Review of Economic Studies, 86(1), 377-412.

[6] Kong, E., O. Linton, and Y. Xia (2010): "Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and its Application to the Additive Model," Econometric Theory, 26(5), 1529-1564.

[7] McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," Journal of Econometrics, 142(2), 698-714.

[8] Qu, Z. and J. Yoon (2015): "Nonparametric Estimation and Inference on Conditional Quantile Processes," Journal of Econometrics, 185(1), 1-19.

[9] Qu, Z. and J. Yoon (2019): "Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs," Journal of Business & Economic Statistics, 4(37), 625-647.