Regression Discontinuity Designs with a Continuous Treatment

Yingying Dong, Ying-Ying Lee, Michael Gou*

First version: April 2017; this version: April 2018

Abstract

This paper provides identification and inference theory for the class of regression discontinuity (RD) designs with a continuous treatment. We identify causal effects of treatment given any discontinuity in the treatment distribution at the RD threshold (including the usual change in mean as a special case). We provide bias-corrected robust inference for the identified local treatment effects, either at a given treatment quantile or averaging over the treatment distribution, and associated asymptotically mean squared error optimal bandwidths. Our model applies to a large class of policies that target parts or features of the treatment distribution other than the mean, such as changing the variance or shifting one or both tails of the distribution. We apply our estimator to investigate the impacts of bank capital on bank stability, taking advantage of the minimum capital requirements in the early 20th century. This policy targets small banks, creating shifts in the lower tail of the capital distribution at certain policy thresholds.

JEL codes: C21, C25, I23
Keywords: Regression discontinuity (RD) design, Continuous treatment, Control variable, Robust inference, Distributional change, Rank invariance, Rank similarity, Capital regulation, Bank stability

1 Introduction

The RD design is originated in the early 1960's (Thistlethwaite and Campbell, 1960). Hahn, Tood, and van der Klaauw (2001) formally establish the identification theory for the RD design. Since then, there has been a large number of studies applying and extending RD designs for treatment effect identification and policy evaluation.[1] The existing RD identification theory assumes a binary treatment (Hahn, Todd, and van der Klaauw 2001; see also discussion in Lee 2008 and Dong 2016).

*University of California Irvine. Yingying Dong, yyd@uci.edu. Ying-Ying Lee, yingying.lee@uci.edu.

[1]Existing theoretical discussion of the standard RD design includes Porter (2003), Lee (2008), Imbens and Kalyanaraman (2012), Frandsen, Frolich, and Melly (2012), Calonico, Cattaneo, and Titiunik (2014), Cattaneo, Frandsen, and Titiunik (2015), Otsu, Xu, and Matsushita (2015), Dong and Lewbel (2015), Angrist and Rokkanen (2015), Feir, Lemieux, and Marmer (2016), Bertanha (2016), Dong (2017), Arai et al. (2017), Chiang, Hsu, and Sasaki (2018), Bugni and Canary (2018), Canay and Kamat (2018), Cattaneo, Jansson, and Ma (2018), and Gerard, Rokkanen, and Rothe (2018) among many others.

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be the outcome of interest, which can be continuous or discrete, and $T \in \mathcal{T} \subset \mathbb{R}$ be a treatment. Let $R \in \mathcal{R} \subset \mathbb{R}$ be the continuous running or forcing variable that partly determines the treatment. Define $Z \equiv 1(R \geq r_0)$ for a known cutoff value of the running variable r_0, where $1(\cdot)$ is an indicator function equal to 1 if the expression in the parentheses is true and 0 otherwise. When the treatment T is binary, let Y_t, $t = 1, 0$, be the potential outcome under treatment or no treatment (Rubin, 1974) and further let C be the set of compliers with $T = Z$. In a seminal paper, Hahn, Todd, and van der Klaauw (2001) show that the following local Wald ratio identifies a local average treatment effect (LATE) for compliers at the RD threshold.

$$\frac{\lim_{r \to r_{+0}} \mathbb{E}[Y|R=r] - \lim_{r \to r_0^-} \mathbb{E}[Y|R=r]}{\lim_{r \to r_{+0}} \mathbb{E}[T|R=r] - \lim_{r \to r_0^-} \mathbb{E}[T|R=r]} = \mathbb{E}[Y_1 - Y_0|R=r_0, C]. \tag{1}$$

In practice, many empirical applications of RD designs involve continuous treatments (see, for recent examples, Oreopoulos, 2006, Card, Chetty, and Weber, 2007, Schmieder, Wachter, and Bender, 2012, Pop-Eleches and Urquiola, 2013, and Clark and Royer, 2013). Empirical researchers typically apply the above standard RD estimand derived for a binary treatment to applications with continuous treatments. Intuitively, with a continuous treatment, under certain conditions, this local Wald ratio might identify a causal effect of the treatment, providing that the average treatment level changes at the RD cutoff, i.e., $\lim_{r \to r_0^+} \mathbb{E}[T|R=r] - \lim_{r \to r_0^-} \mathbb{E}[T|R=r] \neq 0$. Even when it is valid, the interpretation of the above local Wald ratio for a continuous treatment would be more complicated than the case of a binary treatment.[2]

More importantly, many public policies target not necessarily the average units, but only those in some parts (e.g., top or bottom) of the distribution for treatment. Examples include minimum school leaving age, minimum wage, maximum welfare benefits, etc. When there is little or no mean change in the treatment, the standard RD identification may be weak or fail. In this case, can we still identify causal impacts of the treatment?

Consider our empirical scenario for concreteness. We are interested in investigating how banks respond to the minimum capital requirements and further how the mount of capital banks hold affects their short- and long-run outcomes. Capital regulation has been a primary tool used to promote bank stability. We take advantage of the fact that minimum capital requirements change discontinuously at certain policy thresholds in the early 20th century of the United States.

Figure 1 plots banks' capital against the population of the town in which these banks were located. The most prominent feature of this scatter plot is the stair-case shaped bottom contour. This bottom contour, marked by the red line, reveals the minimum capital requirements at different town populations. In towns with a population below 3,000, banks were required to hold a minimum capital of \$25,000. The minimum capital requirements doubled or changed to \$50,000 once a town had a population 3,000 or above. Then at the 6,000 population threshold, the minimum capital requirements

[2]We discuss this in greater detail in Section 3.1.

Figure 1: Minimum capital requirements in 1905

doubled again, changing from $50,000 to $100,000. Finally at at the 50,000 population threshold, the requirements doubled yet another time, changing from $100,000 to $200,000.

Clearly, the minimum capital requirements target banks with low levels of capital, so changes in bank capital occur primarily at the bottom of the capital distribution at each policy threshold. Further empirical analysis in Section 6 shows that there is no significant change in the average capital level around the first threshold, where most of banks are present. Can we still identify the causal effects of increased bank capital?

Write $T = \alpha + \beta Z$, where random variables $\alpha = T_0$ and $\beta = T_1 - T_0$. One may then utilize this first-stage heterogeneity or random coefficient to identify causal effects of the treatment. In particular, following Imbens and Newey (2009), we construct a 'control variable,' which is the conditional cumulative distribution function of T given Z and R. Intuitively, this control variable isolates exogenous distributional changes, instead of mean changes, at the RD cutoff for identification. The defining feature of any 'control variable' is that conditional on this variable, treatment is exogenous to the outcome of interest.

Using this control variable approach, we establish new identification results for the class of RD designs with a continuous treatment. We show that as long as there are any changes in the treatment distribution (including the mean change as a special case) at the RD threshold, one can identify causal effects of the treatment. We provide identification of local treatment effects at different treatment quantiles. These quantile specific treatment effects provide useful information on treatment effect heterogeneity at different treatment levels. We further identify a local (weighted) average effect averaging over the treatment distribution. The estimand for the local (weighted) average effect incorporates the standard RD local Wald ratio as a special case. It works (and is the same) when the standard RD estimand works, and can still work when the standard RD estimand does not. Finally we provide both conventional and bias-corrected robust inference for the identified causal parameters, along with their

associated asymptotic mean squared error (AMSE) optimal bandwidths.

The new theory allows us to quantify how bank capital affects bank short- and long-run outcomes among those banks targeted by the capital regulation. We show that while capital requirements lead to small banks to hold more capital, these banks respond in ways that prevent the regulation from having intended effects. On average a 1% increase in capital leads to almost a 1% increase in assets among banks at lower quantiles of the capital distribution. Leverage is not significantly lowered, leaving their long run (up to 24 years) risk of suspension stays unchanged.

Our paper complements the existing theoretical discussion of the classical RD design with a binary treatment. Our paper is further related to a few other important literatures. For example, our paper aims to discuss causal model identification with a continuous treatment. Existing studies on causal model identification primarily focus on binary treatments. This includes the LATE literature (see, e.g., Imbens and Angirst, 1994, Angrist, Imbens, and Rubin 1996), the local quantile treatment effect (LQTE) literature (see, e.g., Abadie, Angrist, and Imbens, 2002, Abadie, 2003, Frolich and Melly 2013), and the marginal treatment effect (MTE) literature (see, e.g., Vytlacil 2002, Heckman and Vytlacil 2005, 2007, Carneiro, Heckman and Vytlacil, 2010). Continuous treatments have received far less attention. Important work discussing causal identification with a continous treatment includes Angrist, Graddy, and Imbens (2000) and Florens et al., (2008) among others.

More broadly, our paper is related to the non-separable IV literature with continuous endogenous covariates, where identification typically requires a scalar unobservable (rank invariance) in either the first-stage or the outcome equation or both (see, e.g., Chesher, 2003, Horowitz and Lee, 2007, Chernozhukov, Imbens, and Newey, 2007, Florens et al., 2008, Imbens and Newey, 2009, D'haultfoeuille and Février, 2015, and Torgovitsky, 2015). In contrast, we allow for multidimensional unobservables in both the first-stage and outcome equations and exemplify identification with a binary 'IV' in this case. Torgovitsky (2015) particularly discusses the identifying power of imposing restrictions on heterogeneity or the dimensions of unobservables. He shows that by imposing rank invariance in both the first-stage and outcome equations, one can identify an infinite-dimensional object even with a discrete or binary IV. See also similar discussion in D'haultfoeuille and Février (2015).

Our paper is also related to the IV literature using first-stage heterogeneity for identification. See, for recent examples, Brinch et al., (2017) and Caetano and Escanciano (2017). These existing studies consider heterogeneity of treatment responses in covariates. In contrast, we consider heterogeneity of treatment responses along the treatment distribution. Further, the RD QTE model of Frandsen Frolich, and Melly (2012) considers a binary treatment in the RD design, but identifies heterogenous treatment effects along the outcome distribution.

The rest of the paper proceeds as follows. Section 2 provides identification results and discusses testable implications of the identifying assumptions. Section 3 discusses the standard RD estimand and further provides a doubly robust estimand for the local weighted average treatment effect. Section 4 describes our estimators. Section 5 discusses the asymptotic properties of the proposed estimators.

Section 6 presents the empirical analysis. Short concluding remarks are provided in Section 7. All proofs are in the Appendix.

2 Identification

This section discusses causal identification in RD designs with continuous treatments following a control variable approach by Imbens and Newey (2009). Various discussions on the control variable approach to simultaneous equations models include Blundell and Powell (2003), Newey, Powell, and Vella (1999) and Pinkse (2000), and Ma and Koenker (2006).

Assume $Y = G(T, R, \varepsilon)$, where $\varepsilon \in \mathcal{E} \subset \mathbb{R}^{d_\varepsilon}$.[3] We do not restrict d_ε to be finite, so ε is allowed to be of arbitrary dimension. Further assume that the reduced-form equation for the treatment is $T = H_1(R, V_1) Z + H_0(R, V_0)(1 - Z)$.[4] For notational convenience, when $H_z(R, V_z)$ is viewed as a random variable, we further use T_z to denote $H_z(R, V_z)$.

Let $F_{\cdot|\cdot}(\cdot,\cdot)$ and $f_{\cdot|\cdot}(\cdot,\cdot)$ be the conditional cumulative distribution function (CDF) and probability density function (PDF), respectively, and $f_{\cdot}(\cdot)$ be the unconditional PDF. The following discussion assumes $r \in \mathcal{R}$, where $\mathcal{R}$ is an arbitrarily small compact interval around some known cutoff r_0.

Define $U_z \equiv F_{T_z|R}(T_z, R)$, $z = 0, 1$, so U_z is the conditional rank of T_z given R. Assume that for any $r \in \mathcal{R}$, the conditional CDF of T_z is strictly increasing. Thus $U_z|(R = r) \sim Unif(0, 1)$ for any $r \in \mathcal{R}$ and further $U_z \sim Unif(0, 1)$. Rewrite $T_z = q_z(R, U_z)$ for $z = 0, 1$.

Assume that $H_z(r, V_z)$ is strictly monotonic in V_z for all $r \in \mathcal{R}$. This assumption essentially requires that the reduced-form disturbance V_z, $z = 0, 1$, is a scalar, i.e., $V_z \in \mathcal{V}_z \subset \mathbb{R}$. Then $U_z = F_{V_z|R}(V_z, R)$.[5] Normalizing V_z to be U_z requires that given any $r \in \mathcal{R}$, the random variable V_z has a strictly increasing CDF $F_{V_z|R}(V_z, r)$.

Assumption 1 (Smoothness)**.** *$q_z(\cdot, u)$, $z = 0, 1$, is a continuous function for all $u \in (0, 1)$. $G(\cdot,\cdot,\cdot)$ is a continuous function. $f_{\varepsilon|U_z,R}(e, u, r)$ for all $e \in \mathcal{E}$ and $u \in (0, 1)$ is continuous in $r \in \mathcal{R}$. $f_R(r)$ is continuous and strictly positive at $r = r_0$.*

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

[3] For clarity of derivation, we assume that any observable covariates are subsumed in ε. All the derivations can be extended to explicitly condition on covariates.

[4] Assume $T = H(R, V)$ for $V \in \mathcal{V} \subset \mathbb{R}^{d_V}$. Since $Z = 1(R \geq r_0)$ is binary and is a deterministic function of R, then without loss of generality, one can rewrite $T = H_1(R, V_1) Z + H_0(R, V_0)(1 - Z)$.

[5] Under this assumption, $V_z = H_z^{-1}(r, T_z)$, where $H_z^{-1}(r, T_z)$ is the inverse function of $H_z(r, v_z)$ in its second argument. For any t and any $r \in \mathcal{R}$ and $r \geq r_0$,

$$F_{T_1|R}(t, r) = \Pr(H_1(R, V_1) \leq t | R = r) = \Pr(V_1 \leq H_1^{-1}(r, t) | R = r) = F_{V_1|R}(H_1^{-1}(r, t), r) = F_{V_1|R}(V_1, r) \sim Unif(0, 1).$$

The above holds for any $r \in \mathcal{R} \cap [r_0, +\infty)$, so $U_1 = F_{V_1|R}(V_1, R)$. Following the same argument, for $R < r_0$, $U_0 = F_{V_0|R}(V_0, R) \sim Unif(0, 1)$.

Assumption 1 assumes that the running variable has only smooth effects on potential treatments and that the running variable, treatment, 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 at the RD threshold, is a standard assumption that is essentially required for any RD designs. Assumption 1 in practice requires the "no manipulation" condition that units cannot sort to be just above or below the RD threshold (McCrary, 2008).

Assumption 2 imposes local rank restrictions. In particular, rank invariance or rank similarity is required to hold only at the RD cutoff. Assumption 2.1 requires units to stay at the same rank of the potential treatment distribution right above or below the RD threshold. This assumption holds if $V_0 = V_1$, conditional on $R = r_0$, i.e., the disturbance in the reduced-form treatment equation $T = H_1(R, V_1) Z + H_0(R, V_0)(1-Z)$ is a scalar at $R = r_0$. A scalar disturbance in the reduced-form treatment equation is one of the key identifying assumptions in Imbens and Newey (2009). Assumption 2.1 holds more generally if $V_0 \sim V_1$, conditional on $R = r_0$.

Assumption 2.2 assumes local rank similarity, a weaker condition than 2.1. Note that 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 here 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, 2006). Unlike Chernozhukov and Hansen (2005, 2006), we impose the similarity assumption on the ranks of potential treatments, instead of ranks of potential outcomes. In our empirical analysis, Assumption 2 requires that if a bank tends to hold more capital when operating in a town with a population just below 3,000, then it would also tend to hold more capital when operating in a town with a population at or right above 3,000.

These local rank restrictions have readily testable implications (see, e.g., Dong and Shu 2018, for discussion on the testable implications of rank invariance or rank similarity in treatment models). In Section 2.3 we discuss a convenient test for both assumptions in our setting.

Lemma 1. *Let* $U \equiv U_1 1(R \geq r_0) + U_0 1(R < r_0)$. *Under Assumptions 1 and 2,* U *is a control variable conditional on* R, *i.e.,* $T \perp \varepsilon|(U, R)$.

The above lemma follows from Theorem 1 of Imbens and Newey (2009). Imbens and Newey (2009) show that under certain conditions, the conditional CDF of an endogenous variable given the excluded IV (and any other exogenous covariates) is a control variable.

We similarly define a control variable U to be the conditional distribution of T given Z and R in our RD design; however, there are important differences. Imbens and Newey (2009) require a continuous IV. With further a large support assumption, which requires that the instrument for any given endogenous variable value varies a lot, Imbens and Newey (2009) show that one can nonparametrically

identify objects such as the quantile structural function. Here the 'IV' $Z = 1(R \geq r_0)$ is binary and is a deterministic function of a possibly endogenous covariate R. As we show in the proof of Lemma 1, for any $r \in \mathcal{R} \backslash r_0$, conditional on U and R, T is a constant everywhere except at r_0. That is, causal identification with this control variable U is still local to the RD cutoff, which is a generic feature of the RD design.

The following lemma shows that conditioning on U, any changes in the outcome at the RD threshold are causally related to changes in the treatment.

Lemma 2. *Let $\Gamma(Y)$ be an integrable function of Y. Under Assumptions 1–2, for any $u \in (0,1)$,*

$$\begin{aligned} &\lim_{r \to r_0^+} \mathbb{E}[\Gamma(Y) | U = u, R = r] - \lim_{r \to r_0^-} \mathbb{E}[\Gamma(Y) | U = u, R = r] \\ = &\int (\Gamma(G(q_1(r_0, u), r_0, e)) - \Gamma(G(q_0(r_0, u), r_0, e))) \, dF_{\varepsilon|U,R}(e, u, r_0). \end{aligned}$$

Lemma 2 states that conditional on $U = u$, the mean difference in $\Gamma(Y)$ above and below the cutoff represents the impacts of an exogenous change in treatment from $q_0(r_0, u)$ to $q_1(r_0, u)$, where $q_z(r_0, u)$, $z = 0, 1$, is the u quantile of the potential treatment below or above the cutoff. Lemma 2 gives the average effect of the 'IV' $Z = 1(R \geq r_0)$ on the outcome $\Gamma(Y)$ for individuals at the u quantile of the treatment distribution.

Based on Lemma 2, we define our parameters of interest and discuss identification. Let $\mathcal{U} \equiv \{u \in (0,1): |q_1(r_0, u) - q_0(r_0, u)| > 0\}$. For any $u \in \mathcal{U}$, define the treatment quantile specific *LATE (Q-LATE)* as

$$\begin{aligned} \tau(u) &\equiv \int \frac{G(q_1(r_0, u), r_0, e) - G(q_0(r_0, u), r_0, e)}{q_1(r_0, u) - q_0(r_0, u)} dF_{\varepsilon|U,R}(e, u, r_0) \\ &= \frac{\mathbb{E}[Y|U_1 = u, R = r_0] - \mathbb{E}[Y|U_0 = u, R = r_0]}{q_1(r_0, u) - q_0(r_0, u)}. \end{aligned} \tag{2}$$

Further define the weighted average of Q-LATEs or, for short, WQ-LATE as

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

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

Assuming that the function $G(T, r, \varepsilon)$ is continuously differentiable in its first argument, both parameters can be expressed as weighted average derivatives of the outcome $G(T, r, \varepsilon)$ with respect to treatment T. In particular, following Lemma 5 of Angrist, Graddy, and Imbens (2000), one can

write the numerator of (2) as

$$\begin{aligned}
& \mathbb{E}\left[G\left(q_1(r_0,u),r_0,\varepsilon\right)-G\left(q_0(r_0,u),r_0,\varepsilon\right)\Big|U=u,R=r_0\right] \\
= \;& \mathbb{E}\left[\int_{q_0(r_0,u)}^{q_1(r_0,u)}\frac{\partial}{\partial t}G\left(t,r_0,\varepsilon\right)dt\Big|U=u,R=r_0\right] \\
= \;& \int_{q_0(r_0,u)}^{q_1(r_0,u)}\mathbb{E}\left[\frac{\partial}{\partial t}G\left(t,r_0,\varepsilon\right)\Big|U=u,R=r_0\right]dt,
\end{aligned}$$

where the second equality follows from that under standard regularity conditions, one can interchange the order of integration. $\tau(u)$ is then given by

$$\tau(u)=\int_{q_0(r_0,u)}^{q_1(r_0,u)}\mathbb{E}\left[\frac{\partial}{\partial t}G\left(t,r_0,\varepsilon\right)\Big|U=u,R=r_0\right]\varkappa(u)\,dt,$$

for $\varkappa(u)\equiv\frac{1}{q_1(r_0,u)-q_0(r_0,u)}$. That is, $\tau(u)$ is a weighted average derivative, averaging over the change in T at a given quantile u at r_0. It follows that $\pi(\omega)$ is also a weighted average derivative, averaging over both changes in T at a given quantile u and over $\mathcal{U}$ at the RD threshold. The following provides the first-stage assumption for identification of both parameters.

Assumption 3 (First-stage)**.** *$q_1(r_0,u)\neq q_0(r_0,u)$ for at least some $u\in(0,1)$.*

For $z=0,1$, denote the conditional support of T_z given $R=r_0$ as $\mathcal{T}_z$. Assumption 3 requires that $F_{T_1|R}(t,r_0)=F_{T_0|R}(t,r_0)$ does not hold for all $t\in\mathcal{T}_0\cup\mathcal{T}_1$.[6] This is in contrast to the standard RD first-stage assumption requiring a mean change, i.e., $\mathbb{E}[T_1|R=r_0]\neq\mathbb{E}[T_0|R=r_0]$.

For notational convenience, let $T=q(R,U)\equiv q_0(R,U_0)(1-Z)+q_1(R,U_1)Z$. Given smoothness of $q_z(r,U_z)$ by Assumption 1, $q(r,U)$ is right and left continuous in r at $r=r_0$. Then define $q^+(u)\equiv\lim_{r\to r_0^+}q(r,u)$ and $q^-(u)\equiv\lim_{r\to r_0^-}q(r,u)$. Let $m(t,r)\equiv\mathbb{E}[Y|T=t,R=r]$, and similarly 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. The following theorem provides identification of Q-LATE and WQ-LATE.

Theorem 1 (Identification)**.** *Under Assumptions 1–3, for any $u\in\mathcal{U}$, $\tau(u)$ is identified and is given by*

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

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

[6] When either $F_{T_1|R}(t,r_0)$ or $F_{T_0|R}(t,r_0)$ is not defined for $t\in\mathcal{T}_0\cup\mathcal{T}_1$, we assume that $F_{T_1|R}(t,r_0)=F_{T_0|R}(t,r_0)$ does not hold.

To aggregate Q-LATE, one simple weighting function is equal weighting, i.e., $w(u) = w^S(u) \equiv 1/\int_{\mathcal{U}} 1du$. One may choose other properly defined weighting functions, which we disuss in the next section.

Replacing Y by any integrable function $\Gamma(Y)$ in the above, one can readily identify Q-LATE and WQ-LATE on $\Gamma(Y)$. Our identifying Assumption 2 specifies that local rank invariance or rank similarity holds for $U_z \in (0, 1)$, $z = 0, 1$. Theorem 1 in fact only requires that the local rank restrictions hold for $U_z \in \mathcal{U}$, $z = 0, 1$.

In addition to Q-LATE and WQ-LATE, one may identify other parameters. Conditional on $U = u$, there are two potential treatment values at $r = r_0$, in particular, $t_0 \equiv q_0(r_0, u)$ and $t_1 \equiv q_1(r_0, u)$. One can then identify potential outcome distributions at each $u \in (0, 1)$ at the two treatment values. Assume that Y is continuous. Let the potential outcome corresponding to the treatment value $t \in \mathcal{T}$ be $Y_t \equiv G(t, R, \varepsilon)$. Under Assumptions 1-3, $F_{Y_{t_1}|U,R}(y, u, r_0) = \lim_{r \to r_0^+} \mathbb{E}\left[1(Y \leq y) | T = q^+(u), R = r\right]$ for any $u \in (0, 1)$. $F_{Y_{t_0}|U,R}(y, u, r_0)$ can be analogously identified. Further one can identify the LQTE at each $u \in \mathcal{U}$ when treatment changes from $q_0(r_0, u)$ to $q_1(r_0, u)$. It is given by $Q_{Y_{t_1}|U,R}(\upsilon, u, r_0) - Q_{Y_{t_0}|U,R}(\upsilon, u, r_0)$ for any $\upsilon \in (0, 1)$ and $u \in \mathcal{U}$, where $Q_{Y_{t_z}|U,R}(\upsilon, u, r_0) \equiv F^{-1}_{Y_{t_z}|U,R}(\upsilon, u, r_0)$.

3 Discussion and a Doubly Robust Estimand

In this section, we briefly discuss the standard RD estimand and show that it can be expressed as a weighted average of Q-LATEs, using a particular weighting function. We then seek to provide a doubly robust estimand that incorporates the standard RD estimand as a special case. That is, it works and is equivalent to the standard RD estimand when the standard RD estimand works and continues to work under our assumptions when the standard RD estimand does not work.

3.1 Standard RD Estimand

Consider the standard RD estimand in the form of the local Wald ratio,

$$\pi^{RD} \equiv \frac{\lim_{r \to r_0^+} \mathbb{E}[Y|R = r] - \lim_{r \to r_0^-} \mathbb{E}[Y|R = r]}{\lim_{r \to r_0^+} \mathbb{E}[T|R = r] - \lim_{r \to r_0^-} \mathbb{E}[T|R = r]}. \tag{4}$$

Rewrite equation (4) as follows

$$\pi^{RD} = \frac{\int_0^1 \{\mathbb{E}[Y|U_1 = u, R = r_0] - E[Y|U_0 = u, R = r_0]\} du}{\int_0^1 (q_1(r_0, u) - q_0(r_0, u)) du} \quad (5)$$

$$= \int_0^1 \frac{\mathbb{E}[Y|U_1 = u, R = r_0] - \mathbb{E}[Y|U_0 = u, R = r_0]}{\Delta q(u)} \frac{\Delta q(u)}{\int_0^1 \Delta q(u) du} du \quad (6)$$

$$= \int_{\mathcal{U}} \tau(u) \frac{\Delta q(u)}{\int_{\mathcal{U}} \Delta q(u) du} du, \quad (7)$$

where $\Delta q(u) \equiv q_1(r_0, u) - q_0(r_0, u)$, equation (5) follows from the smoothness conditions in Assumption 1, (6) follows from re-arranging, and (7) follows from Assumption 2 and the definition of Q-LATE $\tau(u)$. That is, under our assumptions, the standard RD estimand identifies a weighted average of Q-LATEs, using as weights $w^{RD}(u) \equiv \Delta q(u) / \int_{\mathcal{U}} \Delta q(u) du$. Note that the weighting function $w^{RD}(u)$ is required to be positive over $\mathcal{U}$, i.e., $\Delta q(u) > 0$ or $\Delta q(u) < 0$ for all $u \in \mathcal{U}$. The following assumption is sufficient for $w^{RD}(u)$ for $u \in \mathcal{U}$ to be a well-defined weighting function.

Assumption 2′ (Monotonicity)**.** $\Pr(T_1 \geq T_0 | R = r_0) = 1$ *or* $\Pr(T_1 \leq T_0 | R = r_0) = 1$.

Assumption 2′ requires that treatment T is weakly increasing or weakly decreasing almost surely in response to Z changes. Assumption 2′ implies $\Delta q(u) \geq 0$ or $\Delta q(u) \leq 0$ for all $u \in (0, 1)$. Otherwise, $\Delta q(u)$ and hence the weighting function $w^{RD}(u)$ can be positive or negative. The resulting weighted average in equation (7) would then be a weighted difference of average treatment effects for units with positive treatment changes and units with negative treatment changes if $\int_{\mathcal{U}} \Delta q(u) du \neq 0$, and would be undefined if $\int_{\mathcal{U}} \Delta q(u) du = 0$.

Unlike Assumption 2, which imposes rank restrictions, Assumption 2′ imposes a sign restriction on the treatment changes at the RD threshold. A similar assumption has been made by Angrist, Graddy, and Imbens (2000, Assumption 4) in identifying a general simultaneous equations system with binary IVs.

When Assumption 2 local rank invariance or rank similarity does not hold, $\frac{\mathbb{E}[Y|U_1=u,R=r_0]-\mathbb{E}[Y|U_0=u,R=r_0]}{\Delta q(u)}$ involved in equation (6) does not have a causal interpretation. However, the RD estimand may still identify a causal parameter under Assumption 2′ monotonicity. We formally state this result in the following Lemma 3.

Lemma 3. *Let Assumptions 1, 2′, and 3 hold. Then* π^{RD} *identifies a local causal effect* T *on* Y *at* $R = r_0$.

The proof of Lemma 3 shows that the identified local causal effect is a weighted average of individual treatment effects among those individuals who change their treatment intensity in response to

crossing the RD threshold. When further $G(T, R, \varepsilon)$ is continuously differentiable in its first argument, the identified causal parameter can be expressed as a weighted average derivative of Y w.r.t. T. The exact form of the weighted average derivative is provided in the proof of Lemma 3 in the Appendix.

3.2 Doubly Robust Estimand

The discussion so far suggests that regardless of whether the rank restriction holds or not, the standard RD estimand in general requires monotonicity in order to be a causal estimand. Monotonicity does not always hold in practice. For example, Kitagawa's (2015) test casts doubts on this assumption, when college proximity is used as an IV for college attendance.[7]

Theorem 2 below provides a 'doubly robust' estimand that is valid under either monotonicity or local rank invariance or rank similarity.

Theorem 2 (Doubly Robust Estimand)**.** *Let Assumptions 1 and 3 hold. Then under either Assumption 2 or 2′,*

$$\pi^* = \int_{\mathcal{U}} \frac{m^+(u) - m^-(u)}{q^+(u) - q^-(u)} \frac{|q^+(u) - q^-(u)|}{\int_{\mathcal{U}} |q^+(u) - q^-(u)| du} du \tag{8}$$

identifies a local causal effect of Y on T at $R = r_0$.

When monotonicity holds, $\pi^* = \pi^{RD}$, which identifies a causal parameter by Lemma 3; otherwise, when monotonicity does not hold, but our rank assumption holds, π^* identifies $\pi(w^*) \equiv \int_{\mathcal{U}} \tau(u) w^*(u) du$ for $w^*(u) \equiv \frac{|\Delta q(u)|}{\int_{\mathcal{U}} |\Delta q(u)| du}$, which is also a causal parameter based on Theorem 1. Note that our doubly robust estimand merely provides a robust way to aggregate individual causal effects. In either case, the estimand identifies a weighted average effect of the treatment among those individuals who change their treatment intensity in response to crossing the RD threshold. The two alternative identifying assumptions lay out different ways in which individuals can respond to crossing the policy threshold.[8] When further $Y = G(T, R, \varepsilon)$ is continuously differentiable in T, in either case, the identified causal parameter can be re-written as a weighted average derivative of Y w.r.t. T.

In contrast, when monotonicity does not hold, π^{RD} in general does not identify a causal parameter. In the special case when treatment effect is locally constant, the weighting function does not matter. The estimands for $\pi(w)$ with any valid weighting function would identify the same constant treatment effect.

[7] Kitagawa's (2015) test is for the joint LATE assumptions, including monotonicity and independence. Rejecting the null can mean that either one of these two assumptions fails.

[8] Under monotonicity, π^* identifies a weighted average of individual causal effects $\frac{G(q_1(u_1), r_0, \varepsilon) - G(q_0(u_0), r_0, \varepsilon)}{q_1(u_1) - q_0(u_0)}$ among those having $q_1(u_1) - q_0(u_0) > 0$ or < 0 for any $(u_0, u_1) \in (0,1) \times (0,1)$. On the other hand, under the rank restriction, π^* identifies a weighted average of $\frac{G(q_1(u), r_0, \varepsilon) - G(q_0(u), r_0, \varepsilon)}{q_1(u) - q_0(u)}$ among those having $\varepsilon | (U_1 = u, R = r_0) \sim \varepsilon | (U_0 = u, R = r_0)$ and $|q_1(u) - q_0(u)| \neq 0$ for any $u \in \mathcal{U}$.

Further, one can relax Assumption 2 or 2′ in Theorem 2 to allow either assumption to hold only among the subsample having $q_1(r_0, u_1) - q_0(r_0, u_0) \neq 0, \forall\, (u_0, u_1) \in \mathcal{U} \times \mathcal{U}$. The estimand in equation (8) would still identify a causal effect for this subsample.

Note that monotonicity imposes a sign restriction on $T_1 - T_0$ at r_0, while the local rank invariance or rank similarity imposes a rank restriction on the joint distribution of T_1 and T_0 at r_0. Both assumptions impose restrictions on the first-stage heterogeneity, but neither assumption implies the other. It is therefore useful to have an estimand that is valid under either assumption.

3.3 Testing for the identifying assumptions

Assumptions 1 and 2 impose local smoothness conditions and rank restrictions for identification. These smoothness conditions can be examined using the standard RD validity tests. In particular, one can test smoothness of the conditional means of covariates or perform the McCrary (2008)'s density test. See, also Otsu, Xu, and Matsushita (2013), Bugni and Canary (2018), Canay and Kamat (2018), and Cattaneo, Jansson, and Ma (2018). These tests are rather standard, so in the following we instead focus on testing the assumption of local rank invariance or rank similarity. We briefly discuss their testable implications and then propose convenient tests for these implications. The proposed tests follow the the general approach of Dong and Shu (2018), which leverages covariates for testing, but are adapted to our particular setup.

It is worth noting that we discuss testing for local rank invariance or local rank similarity by assuming that smoothness conditions hold; otherwise what we propose are joint tests for the implications of both sets of assumptions.

Recall $Y = G(T, R, \varepsilon)$, where ε contains any other (observable and unobservable) covariates of Y other than R. Let $X \in \mathcal{X} \subset \mathbb{R}$ be some observable component of ε. Under either local rank invariance or local rank similarity, $U_0 | (\varepsilon, R = r_0) \sim U_1 | (\varepsilon, R = r_0)$. Further by Bayes' theorem, $\varepsilon | (U_0 = u, R = r_0) \sim \varepsilon | (U_1 = u, R = r_0)$, for any $u \in (0, 1)$, and hence $F_{X|U_1,R}(x, u, r_0) = F_{X|U_0,R}(x, u, r_0)$ for all $x \in \mathcal{X}$ and any $u \in (0, 1)$. As mentioned, our identification of Q-LATE or WQ-LATE only requires local rank invariance or rank similarity to hold for all $u \in \mathcal{U}$. Therefore, to test the implications of these local rank restrictions, one can test the following null hypothesis

$$H_0\colon \lim_{r \to r_0^+} \mathbb{E}[1(X \leq x) | U = u, R = r] - \lim_{r \to r_0^-} \mathbb{E}[1(X \leq x) | U = u, R = r] = 0, \forall\, x \in \mathcal{X}, u \in \mathcal{U}, \tag{9}$$

Note that by Assumption 1, $F_{\varepsilon|U_z,R}(e, u, r)$ and hence $F_{X|U_z,R}(x, u, r)$, $z = 0, 1$, is continuous at $r = r_0$. It follows that $F_{X|U,R}(x, u, r)$ and $F_{X|U,R}(x, u, r)$ are right and left continuous at $r = r_0$, respectively. That is, the above limits exist.

The left-hand side of equation (9) corresponds to the numerator of equation (3) with Y being replaced by $1(X \leq x)$. Testing for these local rank restrictions then amounts to testing that Q-LATEs

or WQ-LATEs on the covariate distribution are zero. Such tests are essentially falsification tests to show that treatment has no false significant impacts on covariates at any treatment quantiles.[9] In practice, instead of testing the entire conditional distribution of X, one may test more conveniently the conditional low order (raw) moments of X.

Lastly, it might be interesting to report the covariate distributions at each u quantile, which characterizes units at the u quantile of the treatment distribution. The covariate distribution is simply given by $\lim_{r\to r_0^+}\mathbb{E}\left[1\left(X\leq x\right)|U=u,\,R=r\right]=\lim_{r\to r_0^-}\mathbb{E}\left[1\left(X\leq x\right)|U=u,\,R=r\right],\forall\, x\in\mathcal{X},\,u\in(0,1)$.

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 employ the same kernel function $K(\cdot)$ for all estimation. Let the bandwidths for T and R be h_T and h_R, respectively. Define $h_T\equiv h\sigma_T$ and $h_R\equiv h\sigma_R$, where σ_T and σ_R are the standard deviations of T and R, respectively. 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 and WQ-LATE by the following procedure.

Step 1: Partition the unit interval $(0,\,1)$ into a grid of equally spaced quantiles $\mathsf{U}^{(l)}\equiv\{u_1,u_2,...,u_l\}$. For $u\in\mathsf{U}^{(l)}$, estimate $\widehat{q}^{+}(u)$ and $\widehat{q}^{-}(u)$ by $\widehat{a}_0^{+}(u)$ and $\widehat{a}_0^{-}(u)$, respectively, from the following local linear quantile regressions

$$\begin{aligned}\left(\widehat{a}_0^{+}(u),\widehat{a}_1^{+}(u)\right) &= \arg\min_{a_0^+,a_1^+}\sum_{\{i:\,R_i\geq r_0\}}K\left(\frac{R_i-r_0}{h_R}\right)\rho_u\left(T_i-a_0^+-a_1^+\left(R_i-r_0\right)\right),\\ \left(\widehat{a}_0^{-}(u),\widehat{a}_1^{-}(u)\right) &= \arg\min_{a_0^-,a_1^-}\sum_{\{i:\,R_i<r_0\}}K\left(\frac{R_i-r_0}{h_R}\right)\rho_u\left(T_i-a_0^--a_1^-\left(R_i-r_0\right)\right),\end{aligned}$$

where $\rho_u(\alpha)=\alpha(u-1(\alpha<0))$ is the standard check function.[10]

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 $\epsilon_n\to 0$ is a positive sequence satisfying $\epsilon_n^{-1}\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)|-|\Delta q(u)|\right|=o_p(1)$ and $\epsilon_n^2\left(\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)|-|\Delta q(u)|\right|\right)^{-1}=o_p(1)$.[11] Estimate $\widehat{m}^{+}(u)$ and $\widehat{m}^{-}(u)$ by $\widehat{b}_0^{+}(u)$ and $\widehat{b}_0^{-}(u)$, respectively, for all $u\in\widetilde{\mathcal{U}}$ from the

[9] In contrast, Dong and Shu (2018) essentially test

$$\lim_{r\to r_0^+}\mathbb{E}\left[1\left(T\leq q_1(u,r)\right)|X,\,R=r\right]-\lim_{r\to r_0^-}\mathbb{E}\left[1\left(T\leq q_0(u,r)\right)|X,\,R=r\right]=0.$$

[10] If desired, one could monotonize $\widehat{q}^{\pm}(u)$ using the inequality constraints or rearrangement methods in Chernozhukov, Fernández-Val, and Galichon (2010) or Qu and Yoon (2015a). Both papers show that the monotonized estimators share the same first-order limiting distribution with the initial local linear estimator.

[11] If one wishes to focus on quantiles such that $|\Delta q(u)|>c$ for some small $c>0$, then one can define the trimming parameter to be $c_n=c+\epsilon_n$, where ϵ_n is defined the same way.

following local linear regressions

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

Step 3: Estimate $\tau(u)$ for all $u\in\widetilde{\mathcal{U}}$ by plugging the above estimates into equation (3), i.e., $\widehat{\tau}(u)=\frac{\hat{m}^+(u)-\hat{m}^-(u)}{\hat{q}^+(u)-\hat{q}^-(u)}$.

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

Our identification theory requires trimming out treatment quantiles where there are no changes, i.e., $\Delta q(u)=0$, whereas in practice we do not know the true $\Delta q(u)$. To avoid any pre-testing problem, we trim out all quantiles having $|\Delta\hat{q}(u)|\leq\epsilon_n$ for some chosen ϵ_n. Lemma 6 in Appendix B shows that when ϵ_n satisfies the above listed conditions, this trimming procedure is asymptotically equivalent to trimming out those treatment quantiles where $\Delta q(u)=0$ and preserves the asymptotic properties of our estimator.

In practice, one can choose $\epsilon_n=\epsilon_n(u)\equiv se\left(\Delta\tilde{q}(u)\right)\times 1.96$, where $\Delta\tilde{q}(u)$ is a preliminary Step 1 estimator of the treatment quantile change, using the bandwidth $\tilde{h}$ such that $\tilde{h}/h\to 0$ and $n\tilde{h}^2/h\to\infty$.[12] The associated standard errors satisfy $se(\Delta\tilde{q}(u))=O_p\left(\left(n\tilde{h}\right)^{-1/2}\right)>se(\Delta\hat{q}(u))=O_p\left((nh)^{-1/2}\right)$. It follows that $\epsilon_n^{-1}\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)|-|\Delta q(u)|\right|=se\left(\Delta\tilde{q}(u)\right)^{-1}se\left(\Delta\hat{q}(u)\right)=o_p(1)$ and $\epsilon_n^2\left(\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)|-|\Delta q(u)|\right|\right)^{-1}=se\left(\Delta\tilde{q}(u)\right)^2se\left(\Delta\hat{q}(u)\right)^{-1}=o_p(1)$. 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 and the asymptotic behavior of our estimator is not affected.

The above describes estimation of average treatment effects. To estimate LQTEs conditional on $U=u$ described in Section 2.1, one may simply replace the local linear mean regressions in Step 2 by local linear quantile regressions. Other steps remain the same.

[12] Consider the bandwidth sequences $h=cn^{-a}$ and $\tilde{h}=cn^{-b}$ for some constants $0<a,b<1$ and $c>0$. The required conditions for ϵ_n are satisfied when choosing b such that $a<b<(a+1)/2$.

Intuitively, there is a trade-off on the convergence rate of ϵ_n: on the one hand, we need ϵ_n converge to zero not too fast compared with $|\Delta\hat{q}|-|\Delta q|$, so that the sampling variation of $\Delta\hat{q}$ in the trimming procedure is asymptotically ignorable; on the other hand, ϵ_n needs to converge to zero fast enough in order to keep all the quantiles in U.

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 appear 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 $\Delta\hat{q}(u)$. To deal with these complications, we build on the results of Kong, Linton, and Xia (2010) and Qu and Yoon (2015a). Qu and Yoon (2015a) provide uniform convergence results for local linear quantile regressions, while Kong, Linton, and Xia (2010) establish strong uniform convergence results for local polynomial estimators.

We first present conventional inference assuming undersmoothing, i.e., applying a bandwidth sequence such that the leading bias is asymptotically first-order negligible. We then discuss the leading bias under more general bandwidth conditions. Following the popular approach of Calonico, Cattaneo, and Titiunik (2014), we bias-correct the main estimators and then develop robust inference for the bias-corrected estimators. The robust inference takes into account the added variability due to bias correction.[13] Additionally, we present the asymptotically mean squared error (AMSE) optimal bandwidths for both the Q-LATE and WQ-LATE estimators by minimizing the asymptotic mean squared errors (AMSE). Imbens and Kalyanaraman (2012) propose the AMSE optimal bandwidth for the standard RD estimator. The robust confidence intervals deliver valid inference when the AMSE optimal bandwidths are used.

We impose the following assumptions for asymptotics.

Assumption 4 (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 $\mathbb{V}\left[G(T_z, R, \varepsilon)|T_z = t, R = r\right]$ is continuous and bounded away from zero; the conditional density $f_{T_z R|Y}(t, r, y)$ is bounded for any $y \in \mathcal{Y}$. $\mathbb{E}\left[\left|Y - \mathbb{E}\left[Y|T_z, R\right]\right|^3\right] < \infty$ for $z = 0, 1$.

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

Assumption 4.1 imposes sufficient smoothness conditions to derive the asymptotic linear representations of $\hat{q}^{\pm}(u)$. In particular, the bounded joint density implies a compact support where the

[13] See Cattaneo, Titiuni, and Vazquez-Bare (2016) for a comparison of different inference approaches for the standard RD design.

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 4.2 imposes additional conditions to derive the asymptotic linear representation of $\hat{\mathbb{E}}[Y|T,R]$ and asymptotic normality of our estimators. Assumption 4.3 presents the standard regularity conditions for the kernel function.

5.1 Asymptotic distributions under undersmoothing

Theorem 3 below presents the asymptotic distribution of $\hat{\tau}(u)$ under a bandwidth sequence $h = h_n$ that goes to zero fast enough with the sample size n (i.e., satisfying $nh^6 \to 0$ instead of $nh^6 \to c \in (0,+\infty)$), so that the bias is asymptotically negligible.

Theorem 3 (Asymptotic distribution of $\hat{\tau}(u)$)**.** *Let Assumptions 1-4 hold. If $h = h_n \to 0$, $nh^3 \to \infty$, and $nh^6 \to 0$, then for $u \in \mathcal{U}$*

$$\frac{\hat{\tau}(u) - \tau(u)}{\sqrt{\mathsf{V}_{\tau,n}(u)}} \longrightarrow_d \mathcal{N}(0,1), \text{ where } \mathsf{V}_{\tau,n}(u) \equiv \frac{\mathsf{V}_\tau(u)}{nh^2}.$$

The exact form of $\mathsf{V}_\tau(u)$ is given by equation (11) of Lemma 5 in Appendix B.

The bandwidth conditions in Theorem 3 imply a bandwidth choice $h = h_n = C_\tau n^{-a}$ for some constant $a \in (1/6, 1/3)$ and $C_\tau \in (0,\infty)$. Theorem 3 implies $\sqrt{nh^2}\left(\hat{\tau}(u) - \tau(u)\right) \to_d \mathcal{N}(0, \mathsf{V}_\tau(u))$, where $\mathsf{V}_\tau(u)$ is the asymptotic variance of $\sqrt{nh^2}\hat{\tau}(u)$. The $100(1-\alpha)\%$ confidence interval for $\tau(u)$ is then given by $\left[\hat{\tau}(u) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}_\tau(u)/(nh^2)}\right]$, where $\Phi^{-1}_{1-\alpha/2}$ 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 4 below similarly presents the asymptotic distribution of $\hat{\pi}^*$ using a bandwidth sequence that goes to zero fast enough with the sample size (i.e., satisfying $nh^5 \to 0$ instead of $nh^5 \to c \in (0,+\infty)$), so that the bias is asymptotically negligible.

Theorem 4 (Asymptotic distribution of $\hat{\pi}^*$)**.** *Let Assumptions 1-4 hold. If $h = h_n \to 0$, $nh^4 \to \infty$, and $nh^5 \to 0$, then*

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

The exact form of V_π is given by equation (13) of Lemma 6 in Appendix B.

The bandwidth conditions in Theorem 4 imply a bandwidth choice $h = h_n = C_\pi n^{-a}$ for $a \in (1/5, 1/4)$ and $C_\pi \in (0,\infty)$. Based on Theorem 4, $\sqrt{nh}\left(\hat{\pi}^* - \pi^*\right) \to_d \mathcal{N}(0, \mathsf{V}_\pi)$, where V_π is the asymptotic variance of $\sqrt{nh}\hat{\pi}^*$. The $100(1-\alpha)\%$ confidence interval for π^* is then given by $\left[\hat{\pi}^* \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}_\pi/(nh)}\right]$. V_π can be estimated by the usual plug-in estimator.

Note that the standard nonparametric bootstrap is valid for our estimators, so in practice, one may apply the usual bootstrap based on drawing n observations with replacement to conveniently obtain standard errors and confidence intervals. The bootstrap is known to be valid for the local linear mean and quantile estimators for $\Delta\hat{m}(u)$ and $\Delta\widehat{q}(u)$. $\hat{\tau}(u)$ is a differentiable function of $\Delta\hat{m}(u)$ and $\Delta\widehat{q}(u)$. The bootstrap is then valid for $\hat{\tau}(u)$ by the standard delta method.[14] Further, $\hat{\pi}^*$ is a differentiable function of $\hat{\tau}(u)$ and $\widehat{w}^*(u)$, while $\widehat{w}^*(u)$ is Hadamard differentiable in $\Delta\widehat{q}(u)$. By the functional delta method (Theorem 23.9 in van der Vaart, 2000), bootstrap is valid for $\hat{\pi}^*$.

5.2 Bias-corrected robust inference

The asymptotic distributions of $\hat{\tau}(u)$ and $\hat{\pi}^*$ presented in the previous section 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. In particular, when $nh^6 \to c \in (0, +\infty)$ or $nh^5 \to c \in (0, +\infty)$, a leading bias term appears in the asymptotic distribution of $\hat{\tau}(u)$ or $\hat{\pi}^*$ (see for details Lemma 5 and Lemma 6 in Appendix B). Denote the bias for $\hat{\tau}(u)$ as $h^2\mathsf{B}_\tau(u)$, and that for $\hat{\pi}^*$ as $h^2\mathsf{B}_\pi$. The exact forms of $\mathsf{B}_\tau(u)$ and B_π are presented, respectively, in equation (10) of Lemma 5 and equation (12) of Lemma 6 in Appendix B. These biases depend on changes in the curvatures of the conditional quantile and mean functions in Step 1 and Step 2 estimation. We propose the following bias-corrected estimator for $\tau(u)$

$$\hat{\tau}^{bc}(u) \equiv \hat{\tau}(u) - h^2\hat{\mathsf{B}}_\tau(u),$$

where $\hat{\mathsf{B}}_\tau(u)$ is a consistent estimator for $\mathsf{B}_\tau(u)$. We similarly propose the bias-corrected estimator for π^*

$$\hat{\pi}^{bc} \equiv \hat{\pi}^* - h^2\hat{\mathsf{B}}_\pi,$$

where $\hat{\mathsf{B}}_\pi$ is a consistent estimator of B_π.

Bias correction reduces biases, but also introduces variability. When the added variability is not accounted for, the empirical coverage of the resulting confidence intervals can be well below their nominal target, which implies that conventional confidence intervals may substantially over-reject the null hypothesis of no treatment effect. Following the robust inference approach of Calonico, Cattaneo, and Titiunik (2014), we present the asymptotic distributions of the bias-corrected estimators $\hat{\tau}^{bc}(u)$ and $\hat{\pi}^{bc}$ by taking into account the sampling variation induced by bias correction.

Theorem 5 (Asymptotic distribution of $\hat{\tau}^{bc}(u)$)**.** *Let Assumptions 1-4 hold. If* $h = h_n \to 0$, $b = b_n \to 0$, $h/b \to \rho \in [0, \infty]$, $n\min\{h^6, b^6\}\max\{h^2, b^2\} \to 0$, $n\min\{h^2, b^6h^{-4}\} \to \infty$, *and*

[14] More specifically, by replacing the observation (Y_i, T_i, R_i) with the bootstrap data (Y_i^*, T_i^*, R_i^*) and replacing the probability measure p with p^* implied by bootstrap sampling, the asymptotic linear representations in Lemma 4 hold for bootstrap estimators $\Delta\widehat{q}^*(u)$ and $\Delta\hat{m}^*(u)$. Bootstrap validity then follows from the fact that $\tau(u)$ is a differentiable function of $\Delta m(u)$ and $\Delta q(u)$.

$nh^3 \max\{1, h^6/b^6\} \to \infty$, *then for any* $u \in \mathcal{U}$,

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

$\mathsf{V}_\tau(u)$ *is defined in Theorem 3. The exact forms of* $\mathsf{V}_{\mathsf{B}_\tau}(u)$ *and* $\mathsf{C}_\tau(u;\rho)$ *are given in equations (21) and (22), respectively, in Appendix B.*

$\mathsf{V}^{bc}_{\tau,n}(u)$ consists of three terms. $\mathsf{V}_\tau(u)$ comes from the variance of $\hat{\tau}(u)$, $\mathsf{V}_{\mathsf{B}_\tau}(u)$ comes from the variance of $\hat{\mathsf{B}}_\tau$, and $\mathsf{C}_\tau(u;\rho)$ comes from the covariance between $\hat{\tau}(u)$ and $\hat{\mathsf{B}}_\tau$.[15] Theorem 5 incorporates three limiting cases depending on $h/b \to \rho \in [0,\infty]$. When $h/b \to 0$, the actual estimator $\hat{\tau}(u)$ is first-order while the bias estimator $\widehat{\mathsf{B}}_\tau(u)$ is of smaller order, i.e., $\mathsf{V}_{\mathsf{B}_\tau}(u)/(nb^6h^{-4}) + \mathsf{C}_\tau(u;\rho)/(nhb) = o_p\left(\mathsf{V}_\tau(u)/\left(nh^2\right)\right)$, and hence the variance reduces to $\mathsf{V}^{bc}_{\tau,n}(u) = \mathsf{V}_\tau(u)/(nh^2)$. When $h/b \to \rho \in (0,\infty)$, then both $\hat{\tau}(u)$ and $\widehat{\mathsf{B}}_\tau(u)$ contribute to the asymptotic variance. When $h/b \to \infty$, the bias estimator $\widehat{\mathsf{B}}_\tau(u)$ is first-order while the actual estimator $\hat{\tau}(u)$ is of smaller order and hence $\mathsf{V}^{bc}_{\tau,n}(u) = \mathsf{V}_{\mathsf{B}_\tau}(u)/(nb^6h^{-4})$.

Note that the additional terms due to bias correction $\mathsf{V}_{\mathsf{B}_\tau}(u)$ and $\mathsf{C}_\tau(u;\rho)$ depend on $\mathsf{V}_\tau(u)$ and some constants determined by the kernel function (see the proof of Theorem 5 in Appendix B for details). As a result, $\mathsf{V}^{bc}_{\tau,n}(u)$ only depends on $\mathsf{V}_\tau(u)$ and some constants, which implies that estimating the robust variance is not computationally more demanding than estimating the conventional variance provided in the previous section.[16] Details for bias and variance estimation are provided in Appendix C. Based on Theorem 5, the $100(1-\alpha)\%$ confidence interval for $\tau(u)$ is $\left[\hat{\tau}^{bc}(u) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}^{bc}_{\tau,n}(u)}\right]$.

Theorem 6 (Asymptotic distribution of $\hat{\pi}^{bc}$)**.** *Let Assumptions 1-4 hold. If* $h = h_n \to 0$, $b = b_n \to 0$, $h/b \to \rho \in [0,\infty]$, $n\min\{h^5, b^5\}\max\{h^2, b^2\} \to 0$, $n\min\{h, b^5h^{-4}\} \to \infty$, *and* $nh^4\max\{1, h^5b^{-5}\} \to \infty$, *then*

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

V_π *is defined in Theorem 4. The exact forms of* $\mathsf{V}_{\mathsf{B}_\pi}$ *and* C_π *are given in equations (25) and (26), respectively, in Appendix B.*

$\mathsf{V}^{bc}_{\pi,n}$ consists of three terms. V_π comes from the variance of $\hat{\pi}^*$, $\mathsf{V}_{\mathsf{B}_\pi}$ comes from the variance of $\hat{\mathsf{B}}_\pi$, and C_π comes from the covariance between $\hat{\pi}^*$ and $\hat{\mathsf{B}}_\pi$. Similar to Theorem 5, Theorem 6 also incorporates three limiting cases depending on $h/b \to \rho \in [0,\infty]$. When $h/b \to 0$, the actual estimator $\hat{\pi}^*$ is first-order while the bias estimator $\widehat{\mathsf{B}}_\tau$ is of smaller order, and hence $\mathsf{V}^{bc}_{\pi,n} \equiv \mathsf{V}_\pi/(nh)$.

[15] For the standard RD design with a binary treatment, Calonico, Cattaneo, and Titiunik (2014) derive the conditional variance given the sample data. In contrast, we derive the asymptotic unconditional variance. These two approaches are asymptotically 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 asymptotic unconditional variance.

[16] For an example of the Uniform kernel and $\rho = 1$, $\mathsf{V}^{bc}_{\tau,n}(u) = 13.89\mathsf{V}_\tau(u)/(nh^2)$

When $h/b \to \rho \in (0,\infty)$, then both $\hat{\pi}^*$ and $\widehat{\mathsf{B}}_\pi$ contribute to the asymptotic variance. When $h/b \to \infty$, the bias estimator $\widehat{\mathsf{B}}_\tau$ is first-order while the actual estimator $\hat{\pi}^*$ is of smaller order, and hence $\mathsf{V}^{bc}_{\pi,n} \equiv \mathsf{V}_{\mathsf{B}_\pi}/(nb\rho^{-4})$. Based on Theorem 6, the $100(1-\alpha)\%$ confidence interval for π (ω^*) is $\left[\hat{\pi}^{bc} \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\mathsf{V}^{bc}_{\pi,n}}\right]$.

Given our results, one can estimate the robust variances by the plug-in estimators. Alternatively, one can bootstrap the standard errors or confidence intervals for the bias-corrected estimators $\hat{\tau}^{bc}(u)$ and $\hat{\pi}^{bc}$. Bootstrap validity follows the same arguments as those for $\hat{\tau}(u)$ and $\hat{\pi}^*$ in Section 5.1.

5.3 AMSE optimal bandwidth

Choosing a bandwidth is known to be a delicate task in nonparametric estimation. Following Imbens and Kalyanaraman (2012), we derive the bandwidths for $\hat{\tau}(u)$ and $\hat{\pi}^*$ that minimize the AMSE. These results are presented in Theorem 7 and Theorem 8 below. Further details for estimating these AMSE optimal bandwidths are provided in Appendix C.

Theorem 7 (AMSE optimal bandwidth for $\hat{\tau}(u)$)**.** *Let Assumptions 1-4 hold . If $h = h_n \to 0$ and $nh^2 \to \infty$, then the mean squared error of $\hat{\tau}(u)$ is $\mathbb{E}\left[\left(\hat{\tau}(u)-\tau(u)\right)^2\right] = h^4\mathsf{B}_\tau(u)^2 + \left(nh^2\right)^{-1}\mathsf{V}_\tau(u) + o\left(h^4+\left(nh^2\right)^{-1}\right)$. If $\mathsf{B}_\tau(u) \neq 0$, the bandwidth that minimizes the asymptotic mean squared error is $h^*_\tau = \left(\mathsf{V}_\tau(u)/\left(2\mathsf{B}^2_\tau(u)\right)\right)^{1/6} n^{-1/6}$.*

The AMSE optimal bandwidth for $\hat{\tau}(u)$ is of the form $C_\tau n^{-1/6}$ for some constant $C_\tau > 0$, which satisfies the bandwidth conditions specified in Theorem 5. Therefore, one can apply the above AMSE optimal bandwidth and then conduct the bias-corrected robust inference provided in Theorem 5.

Theorem 8 (AMSE optimal bandwidth for $\hat{\pi}^*$)**.** *Let Assumptions 1-4 hold. If $h = h_n \to 0$ and $nh \to \infty$, then the mean squared error of $\hat{\pi}^*$ is $\mathbb{E}\left[\left(\hat{\pi}^*-\pi^*\right)^2\right] = h^4\mathsf{B}^2_\pi + (nh)^{-1}\mathsf{V}_\pi + o\left(h^4+(nh)^{-1}\right)$. If $\mathsf{B}_\pi \neq 0$, then the bandwidth that minimizes the asymptotic mean squared error is $h^*_\pi = \left(\mathsf{V}_\pi/\left(4\mathsf{B}^2_\pi\right)\right)^{1/5} n^{-1/5}$.*

The AMSE optimal bandwidth for $\hat{\pi}^*$ is of the form $C_\pi n^{-1/5}$ for some constant $C_\pi > 0$, which satisfies the bandwidth conditions in Theorem 6. These AMSE optimal bandwidths trade off squared biases with variances, so when the biases are small, the AMSE optimal bandwidths can be large.

6 Empirical Analysis

Since the establishment of the national banking system in 1863-1864, capital regulation has been a primary tool used to promote bank stability.[17] Capital regulation is motivated by the concern that a

[17] Earlier capital regulations stipulated minimum capital requirements, while modern capital regulations focus on capital ratios (the percentage of a bank's capital to its risk-weighted assets).

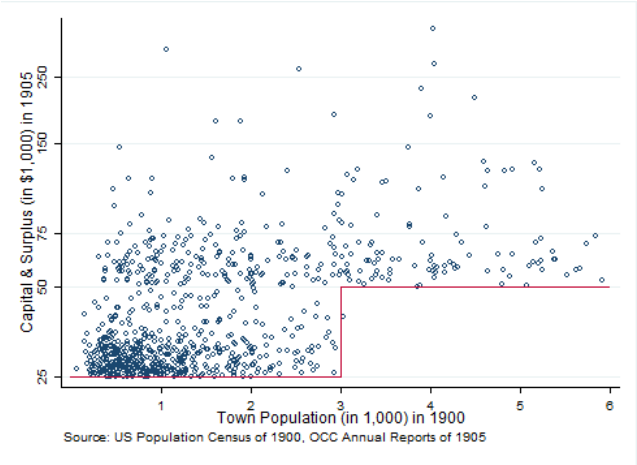

Figure 2: Minimum capital requirements around the town population 3,000 in 1905

bank may hold less capital than is socially optimal, as negative externalities resulting from bank default are not reflected in the market value of bank capital.

Evaluating the true causal impacts of capital requirements on bank behavior faces a number of challenges. The most important challenge is to find exogenous variation in capital requirements. Changes in capital requirements are often responses to ongoing economic events, particularly financial crises. This endogeneity generates a correlation between higher capital requirements and financial instability, which complicates the task of isolating their true causal relationship. Further, modern regulatory regimes often impose the same capital requirements on all banks, leaving no cross-sectional variation to exploit.

The regulation regime in the early 20th century United States provides a unique quasi-experimental setting that allows one to nonparametrically identify the true causal impacts of capital requirements on bank outcomes. As shown previously, the minimum capital requirements were graded according to the population size of the town a bank operated in. The requirement changes abruptly at certain population thresholds.

Figure 2 presents a close-up of banks operating in towns with populations around 3,000, the first regulatory threshold for the minimum capital requirements. Over 80% of the towns in our sample have a population less than 6,000. We therefore focus on the first threshold and exploit the exogenous distributional change in capital for identification. These towns represent rural farming regions where "low population density required, numerous widely, dispersed banking offices" (White, 1983). Arguably these small banks are the right target of the capital regulation. As is clear from Figure 2, the bottom of the capital distribution shifts up at the population threshold 3,000.

We estimate the impacts of increased capital requirements on bank capital (i.e., the first stage impact of Z on T), and further the causal relationships between bank capital and three outcomes of interest (i.e., the impacts of T on Y), total assets, leverage, and the suspension probability in the long run. Our proposed approach makes it possible to quantify banks' responses to increased capital

requirements at different low levels of capital.

6.1 Data Description

We gather first-hand data 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 balance sheet information for all nationally chartered banks.[18] On the asset side, this information includes loans, discounts, investments in securities and bonds, holdings of real estate, cash on hand, deposits in other banks, and overdrafts. On the liability side, this information includes capital, surplus and undivided profits, circulation, and deposits. We collect detailed balance sheet data on individual national banks in 1905, and their suspension outcome in the following 24 years (up to 1929). To avoid any confounding impacts of earlier regulation regimes, we focus on national banks that were established after 1900.[19]

In our analysis, bank assets are defined as the sum of a bank's total amount of assets, and capital is the sum of a bank's capital and surplus. We further define (accounting) leverage as the ratio of a bank's total assets to capital, or the amount of assets a bank holds for each dollar of capital they own.[20] This leverage is a measure of the amount of risk a bank engages in. Higher leverage is associated with lower survival rates during financial crises (Berger and Bouwman, 2013). However, banks generally have an incentive to increase their leverage so they can accumulate higher rates of returns on their capital. We use logged values for all three variables since they have rather skewed distributions.

The OCC's annual report also indicates the town, county, and state in which each bank 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 requirements in 1905 were determined by their town population population in 1900, as reported by the 1900 census. Our final sample consists of 822 banks in 45 towns, among which 717 had a population below 3,000 and 105 had a population at or above 3,000 (but below 6,000). 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.

Brief sample summary statistics are provided in Table 1. Banks operating in towns with more than 3,000 people have more capital on average; they also hold more assets, and have higher measured

[18] Commercial banks can be chartered in the United States at either the federal or state level. National banks are regulated by the Office of the Comptroller of the Currency, while state banks are regulated by their state banking authority.

[19] There was no difference in the minimum capital requirements at the 3,000 population threshold before 1900 and after 1933. Banks were required to have a minimum capital of $50,000 regardless. The requirements changed in 1900, when the Gold Standard Act of 1900 lowered the minimum capital required for banks operating in towns with a population less than 3,000 from $50,000 to $25,000. This was in response to state bank regulation setting their capital requirements lower than national bank capital requirements. Then in 1933, in response to the banking runs, the Banking Act of 1933 raised the minimum capital required for banks operating in towns with a population less than 3,000 back to $50,000.

[20] This is different from various leverage ratios used in the bank regulation, which are defined as the ratio of a bank's capital to its (possibly risk-adjusted) assets.

Table 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

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 suggest that identification using banks far away from the threshold would be confounded, and hence local identification comparing towns near the threshold is crucial for causal conclusions.

6.2 Main Results

We first investigate distributional changes in log capital around the population threshold 3,000. Table 2 reports the estimated changes from 0.10 to 0.90 quantiles and the estimated mean change. Figure 3 plots the estimated quantile curves of log capital right above or below the population threshold 3,000 (left) and the estimated quantile changes (right) along with their 95% confidence bands. These first-stage estimates are based on a bandwidth implied by the undersmoothing conditions in Theorems 3 and 4. In particular, $h_R = 4\sigma_R n^{-0.23} = 1039.5$. Further analysis using other bandwidths suggests that the estimated quantile changes are robust to different bandwidth choices.[21]

Consistent with the visual evidence in Figure 2, results in Table 2 and Figure 3 suggest 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. No significant change is found in the average level of log capital. At the same time, there are visible quantile crossings at the high end of the quantiles in Figure 3. The estimated changes at the related quantiles are negative, even though they are not statistically significant. Given that there is no mean change in log capital at the policy threshold, we cannot apply the standard RD design. Nevertheless, our new approach takes advantage of any changes in the treatment distribution, making it possible to estimate the causal impacts of bank capital in this case.

Tables 3 and 4 present the estimated impacts ($\widehat{\tau}(u)$ and $\hat{\pi}^*$) of log capital on the outcomes of

[21]For our main analysis, we conduct both conventional inference (under undersmoothing) and bias-corrected robust inference using the AMSE optimal bandwidths.

Table 2 Changes in log(capital) at the population threshold 3,000

Quantile			Quantile		
0.10	0.648	(0.113)***	0.550	0.122	(0.378)
0.15	0.575	(0.128)***	0.600	0.063	(0.390)
0.20	0.540	(0.142)***	0.650	0.095	(0.340)
0.25	0.534	(0.193)***	0.700	-0.017	(0.343)
0.30	0.542	(0.230)**	0.750	-0.076	(0.334)
0.35	0.386	(0.337)	0.800	-0.046	(0.371)
0.40	0.313	(0.364)	0.850	-0.044	(0.425)
0.45	0.151	(0.381)	0.900	0.105	(0.650)
0.50	0.065	(0.393)			
Average	0.169	(0.175)			

Note: The top panel presents estimated changes in log capital at different quantiles, while the bottom row reports the estimated average change; The bandwidth is set to be $h_R = 4\sigma_R n^{-0.23} = 1039.5$, which satisfies the undersmoothing conditions for the Q-LATE or WQ-LATE estimator in Theorems 3 and 4; Standard errors are in parentheses; *** Significant at the 1% level; ** Significant at the 5% level.

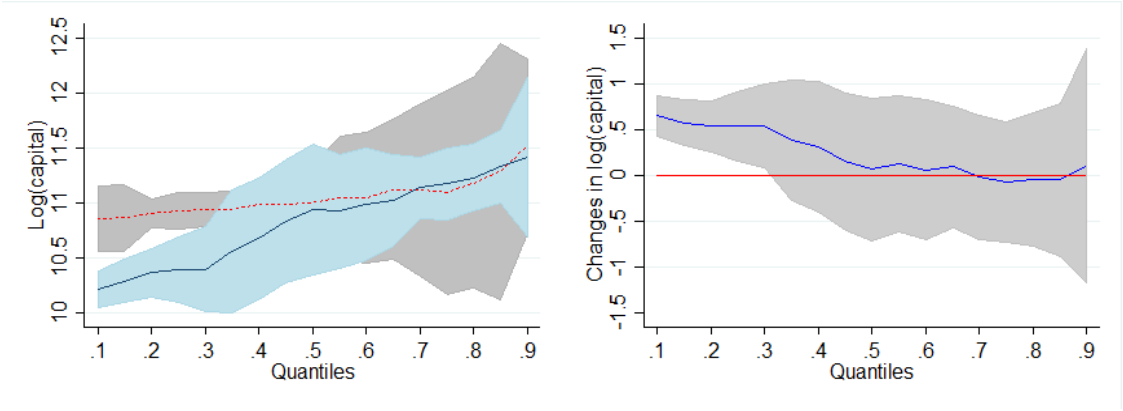

Figure 3: Estimated quantile curves above and below the population threshold 3,000 (left) and the estimated changes at different quantiles (right).

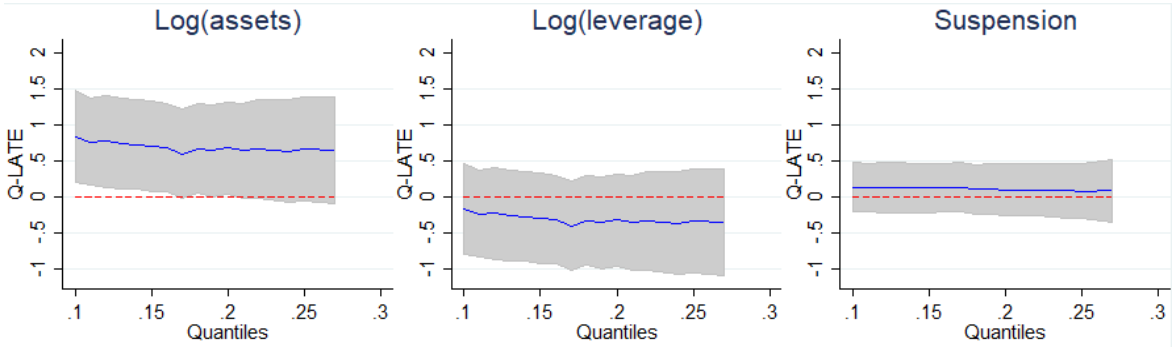

Figure 4: Estimates of Q-LATEs at different quantiles

Table 3 Impacts of log(capital) on bank outcomes (undersmoothing)

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	0.843	(0.322)***	-0.157	(0.322)	0.136	(0.176)
	0.12	0.777	(0.327)**	-0.223	(0.327)	0.133	(0.176)
	0.14	0.734	(0.321)**	-0.266	(0.321)	0.125	(0.176)
	0.16	0.687	(0.312)**	-0.313	(0.312)	0.132	(0.174)
	0.18	0.677	(0.316)**	-0.323	(0.316)	0.107	(0.176)
	0.20	0.681	(0.327)**	-0.319	(0.327)	0.103	(0.181)
	0.22	0.665	(0.348)*	-0.335	(0.348)	0.102	(0.183)
	0.24	0.639	(0.366)*	-0.361	(0.366)	0.088	(0.197)
WQ-LATE		0.700	(0.287)**	-0.300	(0.287)	0.116	(0.172)

Note: The first panel presents estimated Q-LATEs at equally spaced quantiles; The last row presents the estimated WQ-LATEs; The bandwidths are set to be $h_R = 4\sigma_R n^{-0.23} = 1039.5$ and $h_T = 4\sigma_T n^{-0.23} = 0.3905$, which satisfy the undersmoothing conditions in Theorems 3 and 4; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R$, or 779.6; Bootstrapped standard errors are clustered at the town level and are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

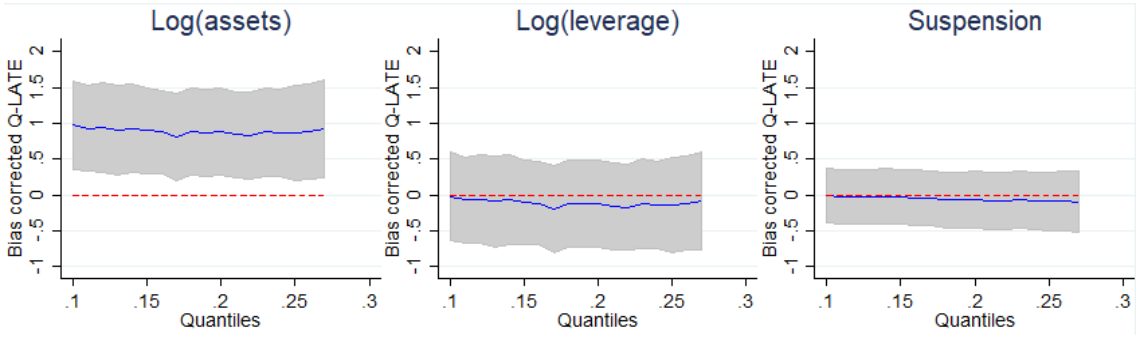

Figure 5: Bias-corrected estimates of Q-LATEs at different quantiles

interest. Estimates in Table 3 use bandwidths consistent with undersmoothing and so no bias correction is made. Bias-corrected counterparts are presented in Table 4. These bias-corrected estimates use the AMSE optimal bandwidth given in Theorem 8.[22] We report bootstrapped standard errors. Since capital regulation varies at the town level, bootstrapping facilitates clustering the standard errors at the town level. Results with analytical standard errors (without clustering) are presented in Appendix D. For brevity, Tables 3 and 4 present the estimated Q-LATEs at selected quantiles. Figures 4 and 5 illustrate the estimated Q-LATEs at a finer grid of quantiles along with the 95% confidence intervals.

Overall, estimates by undersmoothing and those by explicit bias correction show similar patterns. For example, the estimated impacts on log assets are significant for all banks at the low quantiles of log capital, while the estimated impacts on log leverage and suspension are all insignificant. In addition, the estimates for log assets are slightly larger at for banks at the lower quantiles of log capital. Note that the bias-corrected estimates use larger bandwidths and hence there is no loss of

[22]Note that the AMSE optimal bandwidth does not take into account the clustering nature of the error, so they are not necessarily AMSE optimal in this particular empirical application. Rather we use it as a reference point and later present estimates with a larger range of bandwidths.

Table 4 Impacts of log(capital) on bank outcomes (bias-corrected estimates)

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	0.977	(0.314)***	-0.023	(0.314)	-0.001	(0.194)
	0.12	0.945	(0.317)***	-0.055	(0.317)	-0.024	(0.196)
	0.14	0.930	(0.318)***	-0.070	(0.318)	-0.023	(0.199)
	0.16	0.881	(0.295)***	-0.119	(0.295)	-0.036	(0.195)
	0.18	0.880	(0.309)***	-0.120	(0.309)	-0.067	(0.198)
	0.20	0.881	(0.311)***	-0.119	(0.311)	-0.070	(0.202)
	0.22	0.829	(0.307)***	-0.171	(0.307)	-0.078	(0.204)
	0.24	0.864	(0.311)***	-0.136	(0.311)	-0.090	(0.204)
	0.26	0.885	(0.336)***	-0.115	(0.336)	-0.088	(0.212)
WQ-LATE		0.873	(0.298)**	-0.127	(0.298)	-0.051	(0.199)

Note: The first panel presents the bias-corrected estimates of Q-LATEs at equally spaced quantiles; The last row presents the bias-corrected estimates of WQ-LATEs; The standardized AMSE optimal bandwidth for the WQ-LATE estimator is $h_\pi^* = 0.91$ (the standardized AMSE optimal bandwidth for the Q-LATE estimator h_τ^* ranges from 0.72 to 1.1); The bandwidths in the estimation are then set to be $h_R = h_\pi^* \sigma_R = 1108.0$ and $h_T = h_\pi^* \sigma_T = 0.4173$; The bandwidths used to estimate the biases are 2 times of the main bandwidths; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 831.0$; Bootstrapped standard errors are clustered at the town level and are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

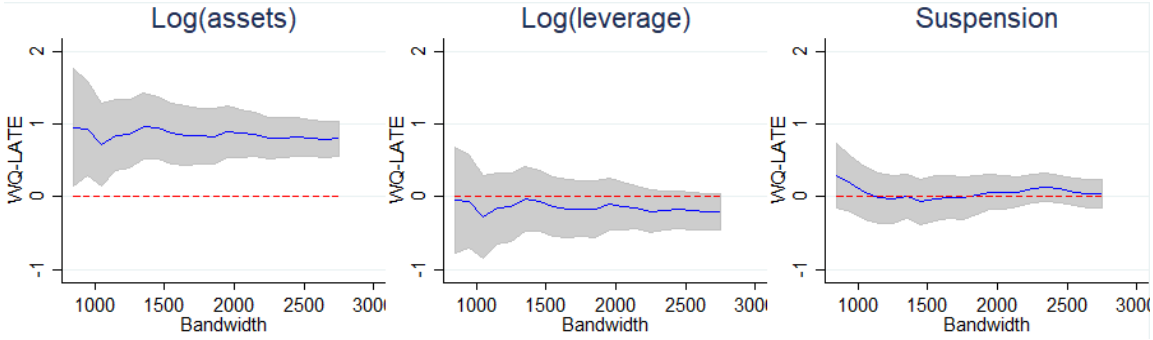

Figure 6: Estimated WQ-LATEs by different bandwidths

precision compared with estimates by undersmoothing. In the following, we focus on interpreting the bias-corrected estimates.

The bias-corrected estimates for log assets in Table 4 range from 0.829 to almost 0.977 at various low quantiles of log capital. All estimates are significant at the 1% level. That is, a 1% increase in bank capital leads to an increase of 0.829% - 0.977% in total assets for banks at the bottom of the capital distribution. The corresponding weighted average is estimated to be 0.873, which is also significant at the 1% level. On average, a 1% increase in bank capital leads to a 0.873% increase in a bank's total assets among all the banks that are affected by the minimum capital requirement. As a result, the estimated decreases in log leverage are all small and insignificant, so the increased minimum capital requirements do not significantly lower leverage among those affected small banks. Not surprisingly, the estimated impacts of bank capital on their long-run suspension probability are

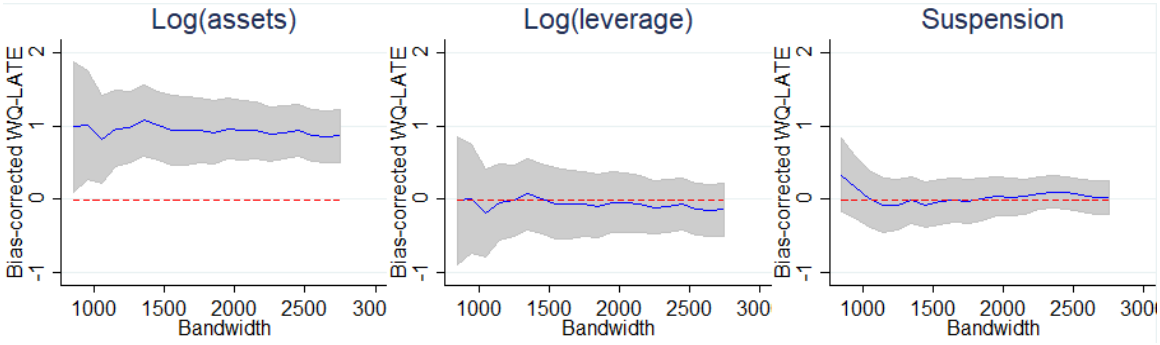

Figure 7: Bias-corrected estimates of WQ-LATEs by different bandwidths

small and insignificant.

Figures 6 and 7 plot sensitivity to bandwidth choices of both types of estimates. Both types of estimates are robust to a wide range of bandwidths. Not surprisingly, with the same bandwidth, bias correction leads to wider 95% confidence intervals.

6.3 Validity Checks

We have estimated the impacts of increased bank capital on various outcomes of interest for banks with different low levels of capital. Validity of these estimates requires smoothness conditions as well as local rank invariance or rank similarity to hold. In the following, we evaluate these assumptions in our empirical scenario.

We first check the smoothness conditions. These smoothness conditions are imposed to ensure that banks as well as their associated business and agricultural conditions above and below the policy threshold are comparable. Given the differential capital requirements, one may be concerned that banks took advantage of the lower capital requirements and hence were more likely to operate in towns with populations just under 3,000, which is evident in the sample summary statistics for the full sample.

We follow the standard practice of the RD literature to test smoothness of the density of town population near the threshold. We also test smoothness of the conditional means of pre-determined covariates. These covariates include bank age and county characteristics, particularly percentage of black population, percentage of farmland and log manufacturing output per capita per square miles.

Figures 8 and 9 provide visual evidence of smoothness of the density of town population and smoothness of the conditional means of these covariates. In particular, the left graph in Figure 8 presents the histogram of the town population, while the right graph plots the log frequency of the town population within each bin of 200 population. Superimposed on the right graph is the estimated log density along with the 95% confidence interval. Formal test results are reported in Table 5. No significant discontinuities are found in the conditional means of these covariates or in the density of town population. Therefore, smoothness conditions are plausible in our empirical setting.

We next test the assumption of local rank invariance or rank similarity. As discussed in Section

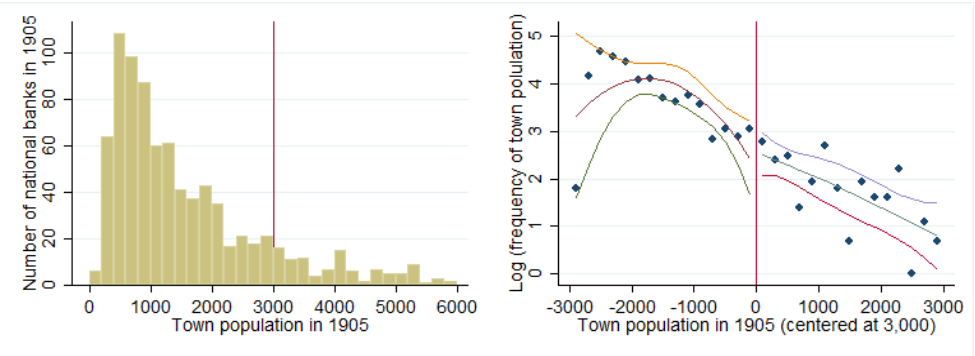

Figure 8: Histogram and the empirical density of town population

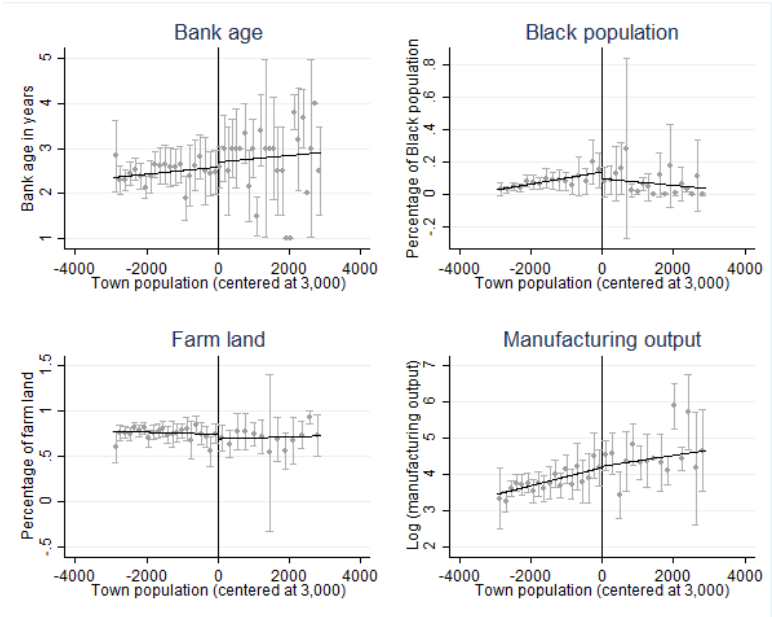

Figure 9: Conditional means of covariates conditional on town population

Table 5 Tests for smoothness of covariates and density

		I: Covariate			
Bank age	0.276	(0.385)	Farmland (%)	-0.013	(0.119)
Black Population (%)	-0.087	(0.093)	Log(manufacturing output)	0.431	(0.429)
		II: Density of town population			
		-0.429	(0.668)		

Note: Panel I presents the estimated discontinuities in the conditional means of covariate; Robust standard errors are clustered at the town level and are in parentheses; Panel II presents the t statistic of the estimated density discontinuity of town population along with the p-value using the Stata command rddensity; $h_R = h_\pi^* \sigma_R = 1108.0$ for all estimation.

Table 6 Tests for local rank invariance or rank similarity

	First moment		Second moment	
Bank age	0.880	(0.764)	3.710	(3.932)
Black Population (%)	-0.035	(0.160)	-0.031	(0.097)
Farmland (%)	0.066	(0.224)	0.170	(0.260)
Log(manufacturing output)	0.068	(0.899)	0.096	(7.710)

Note: Bias-corrected estimates of WQ-LATEs are reported; The standardized AMSE optimal bandwidth for the WQ-LATE estimator is $h_\pi^* = 0.91$, so the bandwidths for estimation are set to be $h_R = 1108.0$ and $h_T = 0.4173$ for R and T; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 831.0$. The bandwidths used to estimate the biases are 2 times of the main bandwidths; Bootstrapped standard errors are in the parentheses.

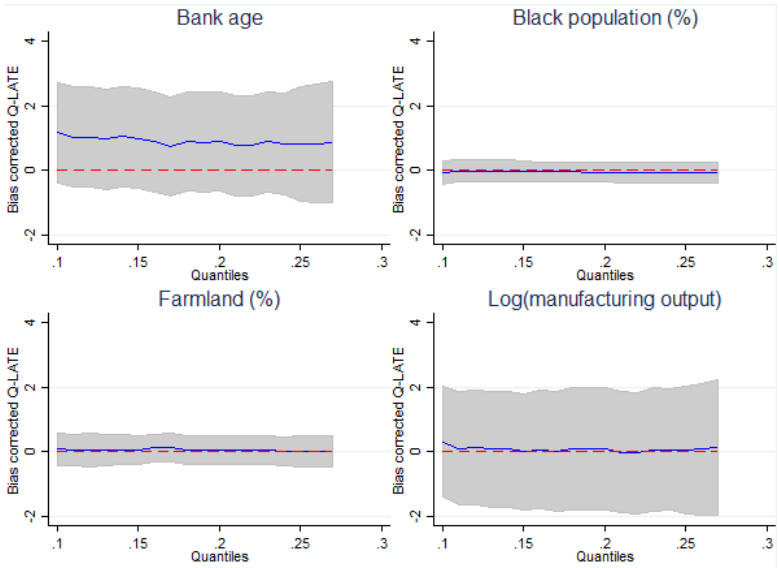

Figure 10: Bias-corrected estimates of Q-LATEs on covariates (first moments)

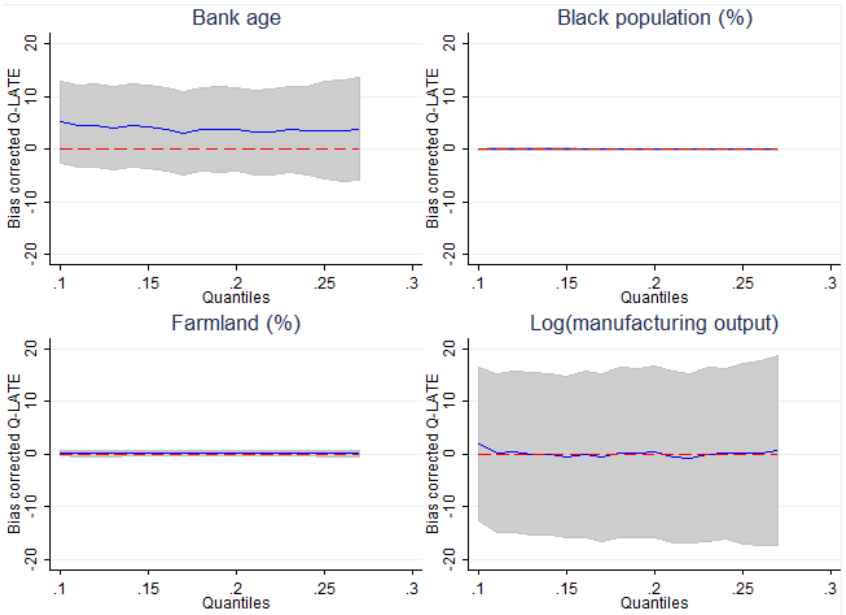

Figure 11: Bias-corrected estimates of Q-LATEs on covariates (second moments)

2.3, assuming that smoothness conditions are satisfied, we can test these local rank assumptions by a falsification test. We replace the outcome variable 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 manufacturing output per capita) and re-estimate Q-LATEs and Q-LATEs. We use the same bandwidth and specification as those used to produce the main estimates in Table 4. Results of these falsification tests are presented in Table 6. Figures 10 and 11 further visualize the results. None of these estimates are significant, so we do not find evidence against local rank invariance or rank similarity. Overall these test results provide empirical evidence supporting plausibility of our identifying assumptions.

6.4 Policy Implications

Our empirical analysis shows that while higher capital requirements indeed lead to banks with low levels of capital to hold more capital, these banks also respond in ways that prevent the regulation from having their intended effects. In particular, as bank capital increase, banks increase their assets by almost the same percentage, leaving their leverage and long-run risk of failure roughly unchanged. This analysis helps shed light on the U.S. banking crisis in the early twentieth century, when bank runs and bank panics occurred often. For example, 29 banking panics occurred from 1865 to 1933. Our analysis also provides empirical evidence supporting the regime shift of capital regulation. Earlier capital regulations stipulated minimum capital requirements, while modern capital regulations focus on capital ratios (the percentage of a bank's capital to its risk-weighted assets).

7 Conclusion

An empirically important class of RD designs involve continuous treatments. The standard RD design identification theory focuses on a binary treatment. This is the first paper to formally consider causal identification and inference in an RD design with a continuous treatment. For identification, we utilize any discontinuous changes in the distribution of treatment that occurs at the RD threshold. Treatment changes are generally responses to relevant policies, and such policies may target features of the treatment distribution other than the mean. By focusing on where the true changes occur in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects.

In particular, we identify not only the local (weighted) average treatment effect but also the treatment effects at different levels of the treatment. When the goal is to identify an aggregate weighted average effect at the policy threshold, our WQ-LATE estimator has an appealing double robustness property – it is valid under either the local rank restriction or local monotonicity assumption. In particular, it is valid and is the same as the standard RD LATE estimator when monotonicity holds, and continues to be valid when monotonicity does not hold, but local rank invariance or rank similarity holds. Our results highlight the complementary nature of these two alternative identifying assump-

tions. Another proposed estimator, the Q-LATE estimator, can be used to estimate quantile specific treatment effects at various treatment quantiles. The Q-LATE estimator provides useful information on treatment effect heterogeneity at different levels of treatment.

We further provide both conventional inference (assuming undersmoothing) and bias-corrected robust inference for the identified treatment effects. We also provide their associated AMSE optimal bandwidths. The new identification and inference theory is relevant to a large class of policies that target a certain part (e.g., top or bottom) or features (e.g., variance) of the distribution for treatment.

In our empirical scenario, the minimum capital regulation leads to the bottom of the capital distribution to shift up at the policy threshold. Our proposed approach utilizes this feature of the capital regulation for causal identification. Causal identification would be difficult by just applying the standard RD design, since there are no significant changes in the average capital level.

We show that while the capital requirement in the early 20th century did lead to small banks holding more capital, these banks responded in ways that prevented the regulation from having its desired effects. On average a 1% increase in capital led to a close to 1% increase in assets among all banks at the lower quantiles of the capital distribution. Leverage was not significantly lowered, and as a result the banks' long-run (up to 24 years, from 1905 to 1929) risk of suspension stayed the same. These results have important implications for understanding the bank runs that were prevalent during the studied time period. Our empirical results also provide additional support for the establishment of the Federal Deposit Insurance Corporation (FDIC) in 1933, for reducing the frequency of bank runs.

Appendix

The Appendix is organized as follows. Section A provides proofs for the lemmas, theorem, and corollary 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 in Section 5 Inference. Section C describes how to estimate the biases, variances, and AMSE optimal bandwidths presented in Sections 5.2 and 5.3. Section D provides additional results for Section 6 Empirical Analysis.

A Proofs for Section 5 Identification

Proof of Lemma 1 First we show that $T \perp \varepsilon | (U, R)$ holds trivially for $R \in \mathcal{R} \backslash r_0$. Next we show $T \perp \varepsilon | (U, R = r_0)$.

For any $R = r \in \mathcal{R}\backslash r_0$ and a bounded function $\eta(T)$,

$$\begin{aligned} & \mathbb{E}\left[\eta\left(T\right)|U=u,\,R=r\right] \\ = \;& \mathbb{E}\left[\eta\left(T_1\right)|U_1=u,\,R=r\right]1\left(r\geq r_0\right)+\mathbb{E}\left[\eta\left(T_0\right)|U_0=u,\,R=r\right]1\left(r<r_0\right) \\ = \;& \eta\left(q_1\left(r,u\right)1\left(r\geq r_0\right)+q_1\left(r,u\right)1\left(r<r_0\right)\right). \end{aligned}$$

That is, $\eta\left(T\right)$ is constant conditional on $U=u$ and $R=r$. Then for any bounded function $\gamma\left(\varepsilon\right)$, we have

$$\mathbb{E}\left[\eta\left(T\right)\gamma\left(\varepsilon\right)|U=u,\,R=r\right]=\mathbb{E}\left[\eta\left(T\right)|U=u,\,R=r\right]\mathbb{E}\left[\gamma\left(\varepsilon\right)|U=u,\,R=r\right].$$

Therefore, $T\perp\varepsilon|\left(U,\,R\right)$ holds trivially for $R\in\mathcal{R}\backslash r_0$.

Next consider the case $R=r_0$. Under either local rank invariance or local rank similarity in Assumption 2, $U_0|\left(\varepsilon,\,R=r_0\right)\sim U_1|\left(\varepsilon,\,R=r_0\right)$. Further by Bayes' Theorem, $\varepsilon|\left(U_0,=u,\,R=r_0\right)\sim\varepsilon|\left(U_1=u,\,R=r_0\right)$ for $u=(0,1)$. Then

$$\begin{aligned} f_{\varepsilon|U_1,R}\left(e,u,r_0\right) &= f_{\varepsilon|U_0,R}\left(e,u,r_0\right)\overset{(1)}{\Longleftrightarrow} \\ \lim_{r\to r_0^+}f_{\varepsilon|U_1,R}\left(e,u,r\right) &= \lim_{r\to r_0^-}f_{\varepsilon|U_0,R}\left(e,u,r\right)\overset{(2)}{\Longleftrightarrow} \\ \lim_{r\to r_0^+}f_{\varepsilon|T,U,R,}\left(e,q_1\left(r,u\right),u,r\right) &= \lim_{r\to r_0^-}f_{\varepsilon|U,R,T}\left(e,q_0\left(r,u\right),u,r\right)\overset{(3)}{\Longleftrightarrow} \\ f_{\varepsilon|T,U,R}\left(e,q_1\left(r_0,u\right),u,r_0\right) &= f_{\varepsilon|T,U,R}\left(e,q_0\left(r_0,u\right),u,r_0\right), \end{aligned}$$

where equivalence (1) follows from smoothness of $f_{\varepsilon|U_z,R}\left(e,u,r\right)$ in Assumption 1, (2) follows from the definition $U=U_1 1\left(R\geq r_0\right)+U_0 1\left(R<r_0\right)$ and the fact that conditional on $U=u$ and $R=r$, T is deterministic, and (3) follows again from smoothness of $f_{\varepsilon|U_z,R}\left(e,u,r\right)$ and $q_z\left(r,u\right)$, $z=0,1$.

Let $t_z\equiv q_z\left(r_0,u\right)$, $z=0,1$. The above shows $f_{\varepsilon|T,U,R}\left(e,t_1,u,r_0\right)=f_{\varepsilon|T,U,R}\left(e,t_0,u,r_0\right)=f_{\varepsilon|U,R}\left(e,u,r_0\right)$ for any $u\in(0,1)$, so $T\perp\varepsilon|U,\,R=r_0$.

Therefore, $T\perp\varepsilon|U,\,R$ for $R\in\mathcal{R}$.

Proof of Lemma 2 For simplicity, the following assumes that Γ is an identity mapping, i.e., $\Gamma\left(Y\right)=Y$. The derivation can be readily extended to any integrable functional Γ.

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

where the first equality follows from the definition $U\equiv U_1 1\left(R\geq r_0\right)+U_0 1\left(R<r_0\right)$ and the fact that conditional on $U=u$, $R=r$, T is deterministic; the second equality follows from the fact

$Y = G(T, R, \varepsilon)$; the third equality follows from the smoothness conditions of Assumption 1, and the last equality follows from the fact that Assumption 2 implies $F_{\varepsilon|U_1,R}(e, u, r_0) = F_{\varepsilon|U_0,R}(e, u, r_0) = F_{\varepsilon|U,R}(e, u, r_0)$.

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

Proof of Lemma 3 For notational convenience, the following derivation uses $q_z(U_z)$ to denote $q_z(r_0, U_z)$, $z = 0, 1$. Assumption 2′ monotonicity states $\Pr(q_1(U_1) \geq q_0(U_0)) = 1$ or $\Pr(q_1(U_1) \leq q_0(U_0)) = 1$. Without loss of generality, we assume the former is true. Given the smoothness conditons in Assumption 1, we have

$$
\begin{aligned}
\pi^{RD} &= \frac{\mathbb{E}\left[G(q_1(U_1), r_0, \varepsilon) | R = r_0\right] - \mathbb{E}\left[G(q_0(U_0)), r_0, \varepsilon) | R = r_0\right]}{\mathbb{E}\left[q_1(U_1) | R = r_0\right] - \mathbb{E}\left[q_0(U_0) | R = r_0\right]} \\
&= \frac{\mathbb{E}\left[G(q_1(U_1), r_0, \varepsilon) - G(q_0(U_0), r_0, \varepsilon) | R = r_0\right]}{\mathbb{E}\left[q_1(U_1) - q_0(U_0) | R = r_0\right]} \\
&= \frac{\iint\int (G(q_1(u_1), r_0, \varepsilon) - G(q_0(u_0), r_0, \varepsilon))\, F_{\varepsilon|U_0,U_1,R=r_0}(de, u_0, u_1)\, F_{U_0,U_1|R=r_0}(du_0, du_1)}{\iint (q_1(u_1) - q_0(u_0)) F_{U_0,U_1|R=r_0}(du_0, du_1)} \\
&= \iint_{\mathcal{I}^2}\int \frac{G(q_1(u_1), r_0, \varepsilon) - G(q_0(u_0), r_0, \varepsilon)}{q_1(u_1) - q_0(u_0)} \widetilde{\omega}^{RD} F_{\varepsilon|U_0,U_1,R=r_0}(de, u_0, u_1)\, F_{U_0,U_1|R=r_0}(du_0, du_1),
\end{aligned}
$$

where $\mathcal{I}^2 \subseteq (0, 1) \times (0, 1)$ denotes the sub support of (U_0, U_1) given $R = r_0$ such that $q_1(U_1) - q_0(U_0) > 0$, and $\widetilde{\omega}^{RD} \equiv \frac{q_1(u_1) - q_0(u_0)}{\iint_{\mathcal{I}^2} (q_1(u_1) - q_0(u_0)) F_{U_0,U_1|R=r_0}(du_0, du_1)}$. When monotonicity holds, $\widetilde{\omega}^{RD} > 0$ and $\iint_{\mathcal{I}^2} \widetilde{\omega}^{RD} F_{U_0,U_1|R=r_0}(du_0, du_1) = 1$. That is, under Assumptions 1, 2′, and 3, π^{RD} identifies a weighted average of individual causal effects $\frac{G(q_1(u_1), r_0, \varepsilon) - G(q_0(u_0), r_0, \varepsilon)}{q_1(u_1) - q_0(u_0)}$ among those having $q_1(u_1) - q_0(u_0) > 0$ for any $(u_0, u_1) \in \mathcal{I}^2$.

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

$$\begin{aligned}
\pi^{RD} &= \frac{\mathbb{E}\left[\int_{q_0(U_0)}^{q_1(U_1)} \frac{\partial G(t,r_0,\varepsilon)}{\partial t} dt \,\middle|\, R = r_0\right]}{\mathbb{E}\left[\int_{q_0(U_0)}^{q_1(U_1)} 1 dt \,\middle|\, R = r_0\right]} \\
&= \frac{\mathbb{E}\left[\int \frac{\partial G(t,r_0,\varepsilon)}{\partial t} 1\left(q_0(U_0) \le t \le q_1(U_1)\right) dt \,\middle|\, R = r_0\right]}{\mathbb{E}\left[\int 1\left(q_0(U_0) \le t \le q_1(U_1)\right) dt \,\middle|\, R = r_0\right]} \\
&= \frac{\int\int\int \mathbb{E}\left[\frac{\partial G(t,r_0,\varepsilon)}{\partial t} \,\middle|\, R = r_0, q_0(u_0) \le t \le q_1(u_1)\right] \Pr\left(q_0(u_0) \le t \le q_1(u_1) | R = r_0\right) du_0 du_1 dt}{\int_{\mathcal{T}}\int\int \Pr\left(q_0(u_0) \le t \le q_1(u_1) | R = r_0\right) du_0 du_1 dt} \\
&= \int\int\int \mathbb{E}\left[\frac{\partial G\left(t, r_0, \varepsilon\right)}{\partial t} \,\middle|\, R = r_0, q_0(u_0) \le t \le q_1(u_1)\right] \overline{w}^{RD} du_0 du_1 dt,
\end{aligned}$$

where $\overline{w}^{RD} \equiv \frac{\Pr(q_0(u_0) \le t \le q_1(u_1) | R = r_0)}{\int\int\int \Pr(q_0(u_0) \le t \le q_1(u_1)) du_0 du_1 dt}$, the second to the last equality follows from the law of iterated expectations and interchanging the order of integration when standard regularity conditions hold.

Proof of Theorem 2 When Assumption 2 local rank invariance or local rank similarity holds, under Assumptions 1 and 3, $\pi^* \equiv \int_{\mathcal{U}} \frac{m^+(u) - m^-(u)}{q^+(u) - q^-(u)} \frac{|q^+(u) - q^-(u)|}{\int_{\mathcal{U}} |q^+(u) - q^-(u)| du} du$ 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 causal parameter by Theorem 1 and the discussion in the main text.

Alternatively, when Assumption 2′ monotonicity holds, under Assumptions 1 and 3, $\pi^* = \pi^{RD}$. Then π^{RD} identifies a causal parameter by Lemma 3.

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. These lemmas can also be of independent interest. Section B.2 collects the proofs of the lemmas in Section B.1. Section B.3 provides the proofs of Theorem 3, Theorem 5, and Theorem 7 in Section 5, which pertain to $\widehat{\tau}(u)$. Section B.4 presents the proofs of Theorem 4, Theorem 6, and Theorem 8 in Section 5, which pertain to $\hat{\pi}^*$.

Notation Let $f_{T|R}^{\pm}(u) \equiv \lim_{r \to r_0^{\pm}} f_{T|R}(q^{\pm}(u), r)$, $\sigma^{2\pm}(u) \equiv \lim_{r \to r_0^{\pm}} \mathbb{V}\left[Y \middle| T = q^{\pm}(u), R = r\right]$, and $q_r''^{\pm}(u) \equiv \lim_{r \to r_0^{\pm}} \partial^2 q(r,u)/\partial r^2$. Let $m_t'^{\pm}(u) \equiv \lim_{r \to r_0^{\pm}} \partial \mathbb{E}\left[Y | T = t, R = r\right]/\partial t|_{t=q^{\pm}(u)}$, $m_t''^{\pm}(u) \equiv \lim_{r \to r_0^{\pm}} \partial^2 \mathbb{E}\left[Y | T = t, R = r\right]/\partial t^2|_{t=q^{\pm}(u)}$, and $m_r''^{\pm}(u) \equiv \lim_{r \to r_0^{\pm}} \partial^2 \mathbb{E}\left[Y | T = q^{\pm}(u), R = r\right]/\partial r^2$.

Define the following constants, which depend entirely on the kernel function. $\kappa_j \equiv \int_0^\infty \nu^j K(\nu) d\nu$, $\lambda_j \equiv \int_0^\infty \nu^j K^2(\nu) d\nu$, $C_{\mathsf{V}} \equiv 4\left(\kappa_2^2 \lambda_0 - 2\kappa_1 \kappa_2 \lambda_1 + \kappa_1^2 \lambda_2\right)\left(\kappa_2 - 2\kappa_1^2\right)^{-2}$, $C_{\mathsf{B}} \equiv \left(\kappa_2^2 - \kappa_1 \kappa_3\right)\left(\kappa_2 - 2\kappa_1^2\right)^{-1}$, and $C_{\mathsf{C}} \equiv \int_0^\infty K(\nu/\rho) K(\nu) d\nu \left(\rho \kappa_2 \int_0^\infty K(\nu/\rho) K(\nu) d\nu - \kappa_1 \int_0^\infty \nu K(\nu/\rho) K(\nu) d\nu\right)$.[23]

[23] For the Uniform kernel, $\lambda_0 = 1/4$, $C_{\mathsf{V}} = 4$, $C_{\mathsf{B}} = -1/12$, $C_{\mathsf{C}} = \rho^3/384$ if $\rho \le 1$, and $C_{\mathsf{C}} = 0.03125(\rho/3 - 0.25)$ if

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

B.1 Preliminary asymptotic results

In the following, Lemma 4 presents the asymptotic linear representations for $\Delta\hat{q}(u)$ and $\Delta\hat{m}(u)$. Lemma 5(I) and Lemma 6(I) present the asymptotic linear representations for $\hat{\tau}(u)$ and $\hat{\pi}^*$, respectively. Lemma 5(D) and Lemma 6(D) present the asymptotic distributions of $\hat{\tau}(u)$ and $\hat{\pi}^*$, respectively.

Lemma 4. *Let Assumptions 1-4 hold. Then uniformly in $u \in \mathcal{U}$,*

(Q) $\Delta\hat{q}(u) - \Delta q(u) - h^2\big(\mathsf{B}_1^+(u) - \mathsf{B}_1^-(u)\big) = n^{-1}\sum_{i=1}^n Z_i\Phi_{1i}^+(u) - (1-Z_i)\Phi_{1i}^-(u) + O_p(h^3) + o_p\left((nh)^{-1/2}\right)$, *where* $\mathsf{B}_1^\pm(u) \equiv C_\mathsf{B} q_r''^\pm(u)\sigma_R^2$,

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

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

(M) $\Delta\hat{m}(u) - \Delta m(u) - h^2\left(\mathsf{B}_2(u) + \mathsf{B}_1^+(u)\,m_t'^+(u) - \mathsf{B}_1^-(u)\,m_t'^-(u)\right) = n^{-1}\sum_{i=1}^n Z_i\big(\phi_{2i}^+(u) + \Phi_{1i}^+(u)m_t'^+(u)\big) - (1-Z_i)\big(\phi_{2i}^-(u) + \Phi_{1i}^-(u)m_t'^-(u)\big) + Rem$, *where* $\mathsf{B}_2(u) \equiv C_\mathsf{B}\big(m_r''^+(u) - m_r''^-(u)\big)\sigma_R^2 + \kappa_2\big(m_t''^+(u) - m_t''^-(u)\big)\sigma_T^2$,

$$\begin{aligned}\phi_{2i}^\pm(u) \;\equiv\; & \big(Y_i - \big(m^\pm(u) + m_r'^\pm(u)(R_i - r_0) + m_t'^\pm(u)\left(T_i - q^\pm(u)\right)\big)\big) \\ & \times \frac{2(\kappa_2 - \kappa_1(R_i - r_0)/(h\sigma_R))}{f_{TR}^\pm(u)(\kappa_2 - 2\kappa_1^2)}\frac{1}{h\sigma_T}K\left(\frac{T_i - q^\pm(u)}{h\sigma_T}\right)\frac{1}{h\sigma_R}K\left(\frac{R_i - r_0}{h\sigma_R}\right)\end{aligned}$$

and the remainder term $Rem = O_p\left(\left(\log n/\left(nh^2\right)\right)^{3/4} + h^4 + n^{-1}h^{-5/2} + h^3\right)$.

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

(I) *Then uniformly in $u \in \mathcal{U}$, $\hat{\tau}(u) - \tau(u) - h^2\mathsf{B}_\tau(u) = n^{-1}\sum_{i=1}^n IF_{\tau i}(u) + Rem$, where*

$$\mathsf{B}_\tau(u) \equiv \left(\mathsf{B}_2(u) + \mathsf{B}_1^+(u)\left(m_t'^+(u) - \tau(u)\right) - \mathsf{B}_1^-(u)\left(m_t'^-(u) - \tau(u)\right)\right)\frac{1}{\Delta q(u)}, \tag{10}$$

$\rho > 1$. For the Epanechnikov kernel, $\lambda_0 = 0.3$, $C_\mathsf{V} = 0.243$, $C_\mathsf{B} = 0.07414$, $C_\mathsf{C} = 0$.if $\rho = 0$, and $C_\mathsf{C} = \lambda_0(\kappa_2\lambda_0 - \kappa_1\lambda_1)$ if $\rho = 1$.

$\mathsf{B}_1^{\pm}(u)$ *and* $\mathsf{B}_2(u)$ *are given in Lemma 4, the influence function* $IF_{\tau i}(u) \equiv \big(Z_i(\phi_{2i}^{+}(u) + \Phi_{1i}^{+}(u)(m_t'^{+}(u) - \tau(u))) - (1 - Z_i)(\phi_{2i}^{-}(u) + \Phi_{1i}^{-}(u)(m_t'^{-}(u) - \tau(u)))\big)(\Delta q(u))^{-1}$, *and* $\Phi_{1i}^{\pm}(u)$, $\phi_{2i}^{\pm}(u)$, *and Rem are given in Lemma 4.*

(D) *If* $h = h_n \to 0$, $nh^3 \to \infty$, *and* $nh^6 \to c \in [0, \infty)$, *then for* $u \in \mathcal{U}$, $\sqrt{nh^2}\big(\hat{\tau}(u) - \tau(u) - h^2\mathsf{B}_\tau(u)\big) \to_d \mathcal{N}\big(0, \mathsf{V}_\tau(u)\big)$, *where*

$$\mathsf{V}_\tau(u) \equiv \frac{2\lambda_0 C_{\mathsf{V}}}{\sigma_T \sigma_R (\Delta q(u))^2 f_R(r_0)} \left(\frac{\sigma^{2+}(u)}{f_{T|R}^{+}(u)} + \frac{\sigma^{2-}(u)}{f_{T|R}^{-}(u)} \right). \tag{11}$$

Lemma 6. *Let Assumptions 1-4 hold.*

(I) *Then* $\hat{\pi}^* - \pi^* - h^2\mathsf{B}_\pi = n^{-1}\sum_{i=1}^{n} IF_{\pi i} + Rem$, *where*

$$\mathsf{B}_\pi \equiv \int_{\mathcal{U}} \mathsf{B}_\tau(u) w^*(u)\, du + \int_{\mathcal{U}} \big(\mathsf{B}_1^{+}(u) - \mathsf{B}_1^{-}(u)\big)\big(\tau(u) - \pi^*\big) \frac{w^*(u)}{\Delta q(u)} du, \tag{12}$$

the influence function $IF_{\pi i} \equiv \big(Z_i\Phi_{21i}^{+} - (1 - Z_i)\Phi_{21i}^{-}\big)\mathbf{1}(U_i \in \mathcal{U}) + \int_{\mathcal{U}}\big(Z_i\Phi_{1i}^{+}(u)\Lambda^{+}(u) - (1 - Z_i)\Phi_{1i}^{-}(u)\Lambda^{-}(u)\big)du$,

$$\Phi_{21i}^{\pm} \equiv \big(Y_i - \big(m^{\pm}(U_i) + m_r'^{\pm}(U_i)(R_i - r_0)\big)\big) \frac{w^*(U_i)}{\Delta q(U_i)} \frac{2\left(\kappa_2 - \kappa_1(R_i - r_0)/(h\sigma_R)\right)}{f_R(r_0)\left(\kappa_2 - 2\kappa_1^2\right)} \frac{1}{h\sigma_R} K\left(\frac{R_i - r_0}{h\sigma_R}\right),$$

$U_i = Z_i F_{T_1|R}(T_{1i}, r_0) + (1 - Z_i)F_{T_0|R}(T_{0i}, r_0)$,[24] $\Lambda^{\pm}(u) \equiv \big(m_t'^{\pm}(u) - \pi^*\big)w^*(u)/\Delta q(u)$, *and* $\Phi_{1i}^{\pm}(u)$ *and Rem are given in Lemma 4.*

(D) *If* $h = h_n \to 0$, $nh^4 \to \infty$, *and* $nh^5 \to c \in [0, \infty)$, *then* $\sqrt{nh}\left(\hat{\pi}^* - \pi^* - h^2\mathsf{B}_\pi\right) \to_d \mathcal{N}(0, \mathsf{V}_\pi)$, *where*

$$\mathsf{V}_\pi \equiv \mathsf{V}_\pi^m + \mathsf{V}_\pi^q. \tag{13}$$

V_π^m *is due to estimation of* $\Delta\hat{m}(u)$ *in Step 2 and*

$$\mathsf{V}_\pi^m \equiv \frac{C_{\mathsf{V}} \int_{\mathcal{U}} \left(\sigma^{2+}(u) + \sigma^{2-}(u)\right) du}{\sigma_R f_R(r_0) \left(\int_{\mathcal{U}} |\Delta q(u)| du\right)^2}. \tag{14}$$

V_π^q *is due to estimation of* $\Delta\hat{q}(u)$ *in Step 1 and*

$$\mathsf{V}_\pi^q \equiv \frac{C_{\mathsf{V}}}{\sigma_R f_R(r_0)} \int_{\mathcal{U}}\int_{\mathcal{U}} (\min\{u, v\} - vu) \left(\frac{\Lambda^{+}(u)\Lambda^{+}(v)}{f_{T|R}^{+}(u) f_{T|R}^{+}(v)} + \frac{\Lambda^{-}(u)\Lambda^{-}(v)}{f_{T|R}^{-}(u) f_{T|R}^{-}(v)} \right) dv du. \tag{15}$$

[24] Note that $\frac{w^*(U_i)}{\Delta q(U_i)} = \frac{2\mathbf{1}\left(q^{+}(U_i) > q^{-}(U_i)\right) - 1}{\int_{\mathcal{U}} |\Delta q(u)| du}$ and $T_i = Z_i T_{1i} + (1 - Z_i)T_{0i}$. Thus an alternative expression is

$$\begin{aligned} \Phi_{21i}^{\pm} \;=\; & \big(Y_i - \big(m(T_i, r_0^{\pm}) + m_r'(T_i, r_0^{\pm})(R_i - r_0)\big)\big) \frac{2\mathbf{1}\left(F_{T_0|R}(T_i, r_0) > F_{T_1|R}(T_i, r_0)\right) - 1}{\int_{\mathcal{U}} |\Delta q(u)| du} \\ & \times \frac{2\left(\kappa_2 - \kappa_1(R_i - r_0)/(h\sigma_R)\right)}{f_R(r_0)\left(\kappa_2 - 2\kappa_1^2\right)} \frac{1}{h\sigma_R} K\left(\frac{R_i - r_0}{h\sigma_R}\right), \end{aligned}$$

where $m_r'(T_i, r_0^{+}) \equiv \lim_{r \to r_0^{+}} \partial \mathbb{E}[Y|T = T_{1i}, R = r]/\partial r$ and $m_r'(T_i, r_0^{-}) \equiv \lim_{r \to r_0^{-}} \partial \mathbb{E}[Y|T = T_{0i}, R = r]/\partial r$.

Define $\chi(u) = \mathbf{1}(|\Delta q(u)| > 0)$. Rewrite $\pi^* = \int_{\mathcal{U}} \tau(u) w^*(u)\, du = \int_0^1 \tau(u) w^*(u)\, \chi(u) du$. In estimation, we replace $\chi(u)$ by $\hat{\chi}(u) = \mathbf{1}\left(|\Delta \hat{q}(u)| > \epsilon_n\right)$. Lemma 7 below shows that using $\hat{\chi}(u)$ is asymptotically equivalent to using $\chi(u)$.

Lemma 7. *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 $\int_0^1 \Delta\hat{q}(u)\big(\hat{\chi}(u) - \chi(u)\big) du = o_p\big(\sup_{u\in\mathcal{U}} \left||\Delta\hat{q}(u)| - |\Delta q(u)|\right|\big)$.*

Given the above Lemma 7, in the following proofs for $\hat{\pi}$ and $\hat{\pi}^{bc}$ we focus on estimators using the infeasible trimming function $\chi(u)$.

B.2 Proofs for Section B.1

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

Proof of Lemma 4

(Q) Proof for $\Delta\hat{q}(u)$ By Theorem 1.2 of Qu and Yoon (2015a), 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 $\mathcal{D}_{h_R}^+ \equiv \big[0, (\bar{r} - r_0)/h_R\big) \cap Supp(K)$ 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 $\mathsf{B}_1^+(u) \equiv C_{\mathsf{B}} q_r''^+(u) \sigma_R^2$. Similarly we can show $\mathsf{B}_1^-(u) \equiv C_{\mathsf{B}} q_r''^-(u) \sigma_R^2$.

In addition, by the Taylor expansion in Step 3 of the proof of Theorem 1 in Qu and Yoon (2015a), by their notation, $e_i(u) = -h_R^2 \frac{1}{2} \left(\frac{R_i - r_0}{h_R}\right)^2 \partial^2 q(u,r)/\partial r^2 - h_R^3 \frac{1}{3} \left(\frac{R_i - r_0}{h_R}\right)^3 \partial^3 q(u,r)/\partial r^3 + o(h_R^3)$. 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 asymptotic linear representation holds:[25] $\tilde{m}^+(u) - m^+(u) - \mathbb{B}\left[\tilde{m}^+(u)\right] =$

[25] Note that Kong, Linton, and Xia (2010) use the same bandwidths for all (standardized) regressors.

$n^{-1}\sum_{i=1}^{n} Z_i \phi_{2i}^{+}(u) + O_p\left(\left(\log n/(nh^2)\right)^{3/4}\right)$ uniformly over $u \in \mathcal{U}$, where the bias

$$\mathbb{B}\left[\tilde{m}^{+}(u)\right] = h^2(1,0,0)S_1^{-1}Q_1\left(\frac{\sigma_R^2}{2}m_r''^{+}(u), \sigma_R\sigma_T \lim_{r\to r_0^{+}} \left.\frac{\partial^2 m(t,r)}{\partial r\,\partial t}\right|_{t=q^{+}(u)}, \frac{\sigma_T^2}{2}m_t''^{+}(u)\right)^{\top} + O_p(h^3),$$

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

Note that $(1,0,0)S_1^{-1} = (2\kappa_2, -2\kappa_1, 0)\,/\left(\kappa_2 - 2\kappa_1^2\right)$ and $(1,0,0)S_1^{-1}Q_1 = 2(C_{\mathsf{B}}, 0, \kappa_2)$. Then

$$\mathsf{B}_2(u) \equiv \mathbb{B}\left[\tilde{m}^{+}(u)\right] - \mathbb{B}\left[\tilde{m}^{-}(u)\right] = C_{\mathsf{B}}\sigma_R^2\left(m_r''^{+}(u) - m_r''^{-}(u)\right) + \kappa_2\sigma_T^2\left(m_t''^{+}(u) - m_t''^{-}(u)\right).$$

Applying Corollary 1 of Kong, Linton, and Xia (2010) and Lemma 4(Q), we have

$$\begin{aligned}
&\sup_{u\in\mathcal{U}}\left|\hat{m}^{+}(u) - \tilde{m}^{+}(u) - \left(\hat{q}^{+}(u) - q^{+}(u)\right)\frac{\partial}{\partial t}\mathbb{E}\left[Y|T=t, R=r_0\right]\Big|_{t=q^{+}(u)}\right| \\
= \;& O_p\left(\left(\sup_{u\in\mathcal{U}}\left|\hat{q}^{+}(u) - q^{+}(u)\right|\right)^2\right) \\
&+ O_p\left(\sup_{t\in\mathcal{T}_0}\left|\frac{\partial}{\partial t}\hat{\mathbb{E}}\left[Y|T=t, R=r_0\right] - \frac{\partial}{\partial t}\mathbb{E}\left[Y|T=t, R=r_0\right]\right| \sup_{u\in\mathcal{U}}\left|\hat{q}^{+}(u) - q^{+}(u)\right|\right) \\
= \;& O_p\left((nh)^{-1} + h^4 + \left(\left(nh^4\right)^{-1/2} + h\right)\left((nh)^{-1/2} + h^2\right)\right) \\
= \;& O_p\left(\left(n^{-1}h^{-5/2}\right) + h^3\right),
\end{aligned}$$

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

Proof of Lemma 5

(I) Lemma 4 implies $\|\Delta\hat{q} - \Delta q\|_\infty = O_p\left((nh)^{-1/2} + h^2\right)$, $\|\Delta\hat{m} - \Delta m\|_\infty = O_p\left(\left(nh^2\right)^{-1/2} + h^2\right)$, and uniformly over $u \in \mathcal{U}$,

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

By Lemma 4, we obtain the influence function $IF_{\tau i}(u)$ and the bias.

(D) Now consider the asymptotic variance $\mathsf{V}_\tau(u)$. Since $\mathbb{E}\left[Z\left(u - \mathbf{1}\left(T \le q_1(R,u)\right)\right)\middle|R\right] = 0$, $\mathbb{E}\left[Z_i\Phi_{1i}^{+}\right] = 0$. Since $\lim_{r\to r_0^{+}}\mathbb{E}\left[Y - \left(m^{+}(u) + m_r'^{+}(u)\left(R - r_0\right) + m_t'^{+}(u)\left(T - q^{+}(u)\right)\right)\middle|T = q^{+}(u), R =\right.$

$r\big] = 0$, we can show $\mathbb{E}\left[Z_i \phi_{2i}^{+}\right] = O(h)$. Then the sampling variation from $\hat{m}(u)$ in Step 2 contributes

$$\begin{aligned} h^2 \mathbb{V}\left[Z_i \phi_{2i}^{+}(u)\right] &= h^2 \mathbb{E}\Bigg[Z \mathbb{E}\Bigg[\left(Y - \left(m^{+}(u) + m_r'^{+}(u)\left(R - r_0\right) + m_t'^{+}(u)\left(T - q^{+}(u)\right)\right)\right)^2 \Bigg| T, R \Bigg] \\ &\quad \times \left(\frac{2\left(\kappa_2 - \kappa_1 (R - r_0)/h\right)}{f_{TR}^{+}(u)\left(\kappa_2 - 2\kappa_1^2\right)} \right)^2 \frac{1}{h^2 \sigma_T^2} K^2 \left(\frac{T - q^{+}(u)}{h \sigma_T} \right) \frac{1}{h^2 \sigma_R^2} K^2 \left(\frac{R - r_0}{h \sigma_R} \right) \Bigg] + o(1) \\ &= \frac{2 \lambda_0 C_V}{\sigma_T \sigma_R} \frac{\sigma^{2+}(u)}{f_{TR}^{+}(u)} + o(1), \end{aligned}$$

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

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

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 $\mathsf{V}_\tau(u)$.

To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. The Lyapounov condition $\left(\sum_{i=1}^n \mathbb{V}\left[IF_{\tau i}(u)\right]\right)^{-3/2} \sum_{i=1}^n \mathbb{E}\left[\left|IF_{\tau i}(u)\right|^3\right] = O\left((nh^{-2})^{-3/2}\right) \sum_{i=1}^n \mathbb{E}\left[\left|IF_{\tau i}(u)\right|^3\right] = O\left((nh^2)^{-1/2}\right) = o(1)$ holds.

Proof of Lemma 6

(I) The proof is for the estimator using the infeasible trimming, i.e., we use $\tilde{\omega}^*(u) \equiv \frac{|\Delta \hat{q}(u)|}{\int_{\mathcal{U}} |\Delta \hat{q}(u)| du}$ for $\mathcal{U} = \{u \in (0,1) : |\Delta q(u)| > 0\}$. Denote this infeasible estimator as $\tilde{\pi} \equiv \int_{\mathcal{U}} \hat{\tau}(u) \tilde{\omega}^*(u) du$. We show $l \to \infty$ and $n \to \infty$, $\hat{\pi}^* - \tilde{\pi} = o_p((nh)^{-1/2})$ 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 $\tilde{\omega}^*(u) \equiv \frac{|\Delta \hat{q}(u)|}{\int_{\mathcal{U}} |\Delta \hat{q}(u)| du} \equiv \frac{\hat{A}(u_j)}{\hat{B}}$. A linear expansion $\tilde{\omega}^*(u) - w^*(u) = \frac{\hat{A}(u) - A(u)}{B} - \frac{w^*(u)}{B}\left(\hat{B} - B\right) + O_p\left(\left\|\hat{A} - A\right\|_\infty \left|\hat{B} - B\right|\right) = O_p\left(\|\hat{q} - q\|_\infty\right) = O_p\left((nh)^{-1/2} + h^2\right)$. Then

$$\begin{aligned} \tilde{\pi} - \pi^* &= \int_{\mathcal{U}} \hat{\tau}(u) \hat{\omega}^*(u) du - \int_{\mathcal{U}} \tau(u) w^*(u) du \\ &= \int \left(\hat{\tau}(u) - \tau(u)\right) w^*(u) du + \int \tau(u) \left(\tilde{\omega}^*(u) - w^*(u)\right) du \\ &\quad + \int \left(\hat{\tau}(u) - \tau(u)\right)\left(\tilde{\omega}^*(u) - w^*(u)\right) du, \end{aligned} \tag{16}$$

where the last term is $O_p\left(\left((nh^2)^{-1/2}+h^2\right)\left((nh)^{-1/2}+h^2\right)\right)$ by Lemma 5.

First consider the estimation error in the estimated weighting function $\tilde{\omega}^*(u)$ in equation (16). Let $\phi_{1i}(u) \equiv \phi_{1i}^+(u) - \phi_{1i}^-(u)$, where $\phi_{1i}^+(u) \equiv Z_i\Phi_{1i}^+(u) + h^2\mathsf{B}_1^+(u)$ and $\phi_{1i}^-(u) \equiv (1-Z_i)\Phi_{1i}^-(u) + h^2\mathsf{B}_1^-(u)$, so $\Delta\hat{q}(u) - \Delta q(u) = n^{-1}\sum_{i=1}^n \phi_{1i}(u) + O_p(h^3) + o_p((nh)^{-1/2})$. The absolute value function is Hadamard directionally differentiable. By the delta method in Example 2.1 of Fang and Santos (2015), $\hat{A}(u) - A(u) = |\Delta\hat{q}(u)| - |\Delta q(u)| = n^{-1}\sum_{i=1}^n \phi_{1i}(u)\left(\mathbf{1}\left(\Delta q(u) > 0\right) - \mathbf{1}\left(\Delta q(u) < 0\right)\right) + O_p(h^3) + o_p\left((nh)^{-1/2}\right) = O_p\left((nh)^{-1/2}+h^2\right)$, since $\mathbf{1}\left(\Delta q(u) = 0\right) = 0$ for $u \in \mathcal{U}$. It follows that $\hat{B} - B = \int_{\mathcal{U}}\left(\hat{A}(u) - A(u)\right)du + o(1) = n^{-1}\sum_{i=1}^n \int_{\mathcal{U}} \phi_{1i}(u)\left(\mathbf{1}\left(\Delta q(u) > 0\right) - \mathbf{1}\left(\Delta q(u) < 0\right)\right)du + o_p\left((nh)^{-1/2}\right) = O_p\left((nh)^{-1/2}+h^2\right)$. Then

$$
\begin{aligned}
&\int_{\mathcal{U}} \tau(u)\left(\tilde{\omega}^*(u) - \omega^*(u)\right)du \\
&= \int_{\mathcal{U}} \frac{\tau(u)}{B}\left(\hat{A}(u) - A(u)\right)du - \frac{\pi^*}{B}\left(\hat{B} - B\right) + O_p\left(\int_{\mathcal{U}} |\tau(u)|du \left\|\hat{A} - A\right\|_\infty \left|\hat{B} - B\right|\right) \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{U}} \left(\frac{\tau(u)}{B} - \frac{\pi^*}{B}\right)\phi_{1i}(u)\left(\mathbf{1}\left(\Delta q(u) > 0\right) - \mathbf{1}\left(\Delta q(u) < 0\right)\right)du \\
&\quad + O_p\left((nh)^{-1} + h^4\right) + o_p\left((nh)^{-1/2}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{U}} \left(\tau(u) - \pi^*\right)\phi_{1i}(u)\frac{\omega^*(u)}{\Delta q(u)}du + O_p\left((nh)^{-1} + h^4\right) + o_p\left((nh)^{-1/2}\right). \qquad (17)
\end{aligned}
$$

Next consider the first term in equation (16). 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 4 becomes

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

Let $\mathcal{U}^+ \equiv \left[\underline{u}, \bar{u}\right] \subseteq \mathcal{U}$ such that $\Delta q(u) > 0$ for all $u \in \mathcal{U}^+$. Note $\omega^*(u)/\Delta q(u) = \big(\mathbf{1}(\Delta q(u) >$

$0) - \mathbf{1}(\Delta q(u) < 0)\big)/B$. Then

$$
\begin{aligned}
& \int_{\mathcal{U}^+} \phi_{2i}^+(u) \frac{w(u)}{\Delta q(u)} du \\
= & \int_{q^+(\underline{u})}^{q^+(\bar{u})} \left(Y_i - \left(\mathsf{m}^+(v) + \mathsf{m}_r'^+(v)(R_i - r_0) + \mathsf{m}_t'^+(v)(T_i - v)\right)\right) \\
& \times \frac{2(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_{TR}^+(F_{T_1|R}(v, r_0))\left(\kappa_2 - 2\kappa_1^2\right)} \frac{1}{h_T} K\left(\frac{T_i - v}{h_T}\right) K_{h_R}(R_i - r_0) \frac{f_{TR}^+(F_{T_1|R}(v, r_0))}{f_R(r_0) B} dv \\
= & \int_{\frac{q^+(\underline{u}) - T_i}{h_T}}^{\frac{q^+(\bar{u}) - T_i}{h_T}} \left(Y_i - \left(\mathsf{m}^+(T_i + h_T s) + \mathsf{m}_r'^+(T_i + h_T s)(R_i - r_0) - \mathsf{m}_t'^+(T_i + h_T s)(h_T s)\right)\right) K(s) ds \\
& \times \frac{2(\kappa_2 - \kappa_1(R_i - r_0)/h_R)}{f_R(r_0) B\left(\kappa_2 - 2\kappa_1^2\right)} K_{h_R}(R_i - r_0) \\
= & \Phi_{21i}^+ \mathbf{1}\left(T_{1i} \in \left[q^+(\underline{u}), q^+(\bar{u})\right]\right)\left(1 + O_p\left(h^2\right)\right).
\end{aligned}
$$

The last equality follows by letting $U_{zi} \equiv F_{T_z|R}(T_{zi}, r_0) \sim Unif(0,1)$ for $z \in \{0,1\}$. Thus $T_{1i} = q^+(U_{1i})$ and $\mathsf{m}^+(T_{1i}) = m^+(U_{1i})$. 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 5, the first term in equation (16) is given by

$$
\begin{aligned}
& \int_{\mathcal{U}} \left(\hat{\tau}(u) - \tau(u)\right) w^*(u) du \\
= & \frac{1}{n} \sum_{i=1}^n \left(Z_i \Phi_{21i}^+ - (1 - Z_i)\Phi_{2i}^-\right) \mathbf{1}(U_i \in \mathcal{U}) + \int_{\mathcal{U}} \Big(\phi_{1i}^+ \left(m_t'^+(u) - \tau(u)\right) - \phi_{1i}^- \left(m_t'^-(u) - \tau(u)\right) \\
& + h^2 \mathsf{B}_2(u)\Big) \frac{w^*(u)}{\Delta q(u)} du + Rem.
\end{aligned}
$$

Together with (17), we obtain the asymptotic linear representation for $\hat{\pi}^*$.

(D) The asymptotic variance V_π is derived using the influence function in Lemma 6(I),

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

$\lim_{r \to r_0^+} \mathbb{E}\Big[\left(Y - \left(m^+(U) + m_r^+(U)(R - r_0)\right)\right) w^*(u)(U)/\Delta q(U) \Big| U = F_{T_1|R}(T_1, r_0), R = r\Big] = 0$, so we can show $\mathbb{E}\left[Z_i \Phi_{2i}^+ \mathbf{1}(U_i \in \mathcal{U})\right] = O(h)$ and $\mathbb{E}\left[\mathbf{1}(U_i \in \mathcal{U}) Z_i \Phi_{21i}^+ \int_{\mathcal{U}} Z_i \Phi_{1i}^+(u) \Lambda^+(u) du\right] =$

$O(h)$. Then

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

Note that $(w^{*}(U_i)/\Delta q(U_i))^2=\left(\int_{\mathcal{U}}|\Delta q(u)|du\right)^{-2}$, so $\lim_{n\to\infty} h\mathbb{V}\left[\mathbf{1}\left(U_i\in\mathcal{U}\right)Z_i\Phi_{21i}^{+}\right]$ is

$$\begin{aligned}
&\lim_{n\to\infty} h\mathbb{E}\Bigg[Z\mathbb{E}\Bigg[\left(Y-\left(m^{+}(U)+m_r'^{+}(U)\left(R-r_0\right)\right)\right)^2\Bigg|U,R\Bigg]\left(\frac{w^{*}(U)}{\Delta q(U)}\right)^2\mathbf{1}\left(U\in\mathcal{U}\right) \\
&\quad\times\frac{4(\kappa_2-\kappa_1\frac{R-r_0}{h_R})^2}{f_R^2(r_0)(\kappa_2-2\kappa_1^2)^2}\frac{1}{h^2\sigma_R^2}K^2\left(\frac{R-r_0}{h\sigma_R}\right)\Bigg] \\
&= \frac{C_{\mathsf{V}}}{\sigma_R f_R(r_0)}\frac{\lim_{r\to r_0^{+}}\mathbb{E}\left[\mathbb{V}\left[Y\middle|U,R\right]\mathbf{1}\left(U_1\in\mathcal{U}\right)\middle|R=r\right]}{\left(\int_{\mathcal{U}}|\Delta q(u)|du\right)^2}. \qquad (18)
\end{aligned}$$

Note that $\lim_{r\to r_0^{+}}\mathbb{V}\big[Y\big|U_1=F_{T_1|R}(T_1|r_0),R=r\big]=\lim_{r\to r_0^{+}}\mathbb{V}\big[Y\big|T_1=q^{+}(U_1),R=r\big]=\sigma^{2+}(U_1)$ and $U_1\sim Unif(0,1)$. Thus in (18), $\lim_{r\to r_0^{+}}\mathbb{E}\big[\mathbb{V}\big[Y\big|U,R\big]\mathbf{1}\big(U_1\in\mathcal{U}\big)\big|R=r\big]=\int_{\mathcal{U}}\sigma^{2+}(u)du$.

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

Finally, we argue that as the number of grid points l arbitrarily goes to infinity, we can work with $\tilde{\pi}$ in the above proof by showing that $\hat{\pi}^{*}-\tilde{\pi}=o_p((nh)^{-1/2})$. Since $\lim_{l\to\infty}\mathcal{U}^{(l)}=(0,1)$, $\lim_{l\to\infty}\widetilde{\mathcal{U}}=\widehat{\mathcal{U}}\equiv\{u\in(0,1)\ \ |\Delta\hat{q}(u)|>\epsilon_n\}$ for any n. It follows that $\lim_{l\to\infty}l^{-1}\sum_{j=1}^{l}|\Delta\hat{q}(u_j)|=\int_{\widehat{\mathcal{U}}}|\Delta\hat{q}(u)|du$ and $\lim_{l\to\infty}\hat{\pi}^{*}=\int_{\widehat{\mathcal{U}}}\hat{\tau}(u)\tilde{w}^{*}(u)du$ for any n. Next we argue that using the estimated trimming $\widehat{\mathcal{U}}$ is asymptotically equivalent to using the unknown $\mathcal{U}$. Lemma 7 implies $\int_{\widehat{\mathcal{U}}}|\Delta\hat{q}(u)|du-\int_{\mathcal{U}}|\Delta\hat{q}(u)|du=\int_0^1|\Delta\hat{q}(u)|(\hat{\chi}(u)-\chi(u))du=o_p((nh)^{-1/2})$. The smoothness condition in Assumption 4.2 implies Lipschitz continuity $\tau(u)w^{*}(u)\times\int_{\mathcal{U}}|\Delta q(u)|du=O(|\Delta q(u)|)$. Thus $\left|\int_{\widehat{\mathcal{U}}}\hat{\tau}(u)\tilde{w}^{*}(u)du-\int_{\mathcal{U}}\hat{\tau}(u)\tilde{w}^{*}(u)du\right|=O_p\left(\int_0^1|\Delta\hat{q}(u)|(\hat{\chi}(u)-\chi(u))du\right)=o_p((nh)^{-1/2})$ by Lemma 7. Therefore as $l\to\infty$ and $n\to\infty$, $\hat{\pi}^{*}-\tilde{\pi}=o_p((nh)^{-1/2})$.

Proof of Lemma 7 Rewrite

$$\begin{aligned}
\hat{\chi}(u)-\chi(u) &= \mathbf{1}\left(|\Delta\hat{q}(u)|>\epsilon_n,|\Delta q(u)|\le 0\right)-\mathbf{1}\left(|\Delta\hat{q}(u)|\le\epsilon_n,|\Delta q(u)|>0\right) \\
&= \mathbf{1}\left(|\Delta\hat{q}(u)|>\epsilon_n,|\Delta q(u)|\le 0\right)-\mathbf{1}\left(|\Delta\hat{q}(u)|\le\epsilon_n<2\epsilon_n<|\Delta q(u)|\right) \qquad (19) \\
&\quad-\mathbf{1}\left(|\Delta\hat{q}(u)|\le\epsilon_n,0<|\Delta q(u)|\le 2\epsilon_n\right) \qquad (20)
\end{aligned}$$

By the condition $\epsilon_n^{-1}\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)|-|\Delta q(u)|\right|=o_p(1)$, the first term in (19) $\mathbf{1}\big(|\Delta\hat{q}(u)|>$

$\epsilon_n, |\Delta q(u)| \leq 0\big) \leq \mathbf{1}\big(\big||\Delta\hat{q}(u)| - |\Delta q(u)|\big| > \epsilon_n\big) = 0$ 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 (19) and implies $\int_0^1 |\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)| \leq \epsilon_n < 2\epsilon_n < |\Delta q(u)|\big)du = o_p\big(\sup_{u\in\mathcal{U}}\big||\Delta\hat{q}(u)| - |\Delta q(u)|\big|\big)$.

For the term in (20), note that $\int_0^1 \mathbf{1}\left(0 < |\Delta q(u)| \leq 2\epsilon_n\right)du = F(2\epsilon_n)$ denotes the CDF of $|\Delta q(U)|$ with $U \sim Unif(0,1)$. By the smoothness Assumption 4.1, we can apply a Taylor series expansion $F(2\epsilon_n) = F'(0)2\epsilon_n + o(\epsilon_n) = O(\epsilon_n)$. Therefore

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

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

B.3 Proofs of Theorem 3, Theorem 5, and Theorem 7 for $\tau(u)$

Proof of Theorem 3 Lemma 5 implies Theorem 3 by letting the bias be of smaller order, i.e., $\sqrt{nh^2}h^2\mathsf{B}_\tau(u) = o(1)$.

Proof of Theorem 5 The following derives the terms $\mathsf{V}_{\mathsf{B}_\tau}(u)$ and $\mathsf{C}_\tau(u;\rho)$ in the asymptotic variance of $\sqrt{nh^2}\hat{\tau}^{bc}(u)$, which are due to bias-correction. They are defined as follows.

$$\mathsf{V}_{\mathsf{B}_\tau}(u) \equiv \mathsf{V}_\tau(u)\frac{4\lambda_0}{C_\mathsf{V}}\left(C_\mathsf{B}e_4 + \kappa_2 e_6\right)^\top S_2^{-1}e_1e_1^\top S_2^{-1}\left(C_\mathsf{B}e_4 + \kappa_2 e_6\right) \text{ and} \tag{21}$$

$$\mathsf{C}_\tau(u;\rho) \equiv -\mathsf{V}_\tau(u)\frac{8\left(C_\mathsf{B}e_4 + \kappa_2 e_6\right)^\top S_2^{-1}e_1}{\lambda_0 C_\mathsf{V}(\kappa_2 - 2\kappa_1^2)}C_\mathsf{C}. \tag{22}$$

For notational simplicity, we suppress the notation for u in the functions of u. Let $\widehat{\mathsf{B}}_\tau - \mathsf{B}_\tau \equiv \widehat{\mathsf{B}}_\tau^+ - \mathsf{B}_\tau^+ - \left(\widehat{\mathsf{B}}_\tau^- - \mathsf{B}_\tau^-\right)$. We linearize the estimator and focus on the part above the cutoff: $\widehat{\mathsf{B}}_\tau^+ - \mathsf{B}_\tau^+ = \left\{\widehat{\mathsf{B}}_2^+ - \mathsf{B}_2^+ - \mathsf{B}_1^+\left(\hat{\tau} - \tau\right) + \left(\hat{\mathsf{B}}_1^+ - \mathsf{B}_1^+\right)(m_t'^+ - \tau)\right\}/\Delta q + Rem_\tau$. Corollary 1 of Kong, Linton, and Xia (2010) for the local quadratic estimator implies the asymptotic linear representation for $\widehat{\mathsf{B}}_2^+ - \mathsf{B}_2^+$ in (23) below and the convergence rates of the derivatives in $\widehat{\mathsf{B}}_2^+$: $\|\hat{m}_r''^+ - m_r''^+\|_\infty = O_p\big(\left(nb^6\right)^{-1/2} + b\big)$, $\|\hat{m}_t''^+ - m_t''^+\|_\infty = O_p\big(\left(nb^6\right)^{-1/2} + b\big)$, and $\|\hat{m}_t'^+ - m_t'^+\|_\infty = O_p\big(\left(nb^4\right)^{-1/2} + b^2\big)$. Lemma 3 in Qu and Yoon (2015b) suggests $\|\hat{q}_r''^+ - q_r''^+\|_\infty = O_p\big(\left(nb^5\right)^{-1/2} + b\big)$. Thus it can be shown that the

term associated with $\hat{q}_r^{\prime\prime+}$ in $\widehat{\mathsf{B}}_1^+$ and the remainder terms Rem_τ are of smaller order.

$$\begin{aligned}
&\widehat{\mathsf{B}}_\tau^+ - \mathsf{B}_\tau^+ \\
= \;& \frac{1}{\Delta q}\left\{b^{-2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \beta_{n2}^{*+} + \mathbb{B}\left[\hat{\mathsf{B}}_2^+\right]\right\} + O_p\left(\frac{1}{b^2}\left(\frac{\log n}{nb^2}\right)^{3/4}\right) \\
& -\frac{\mathsf{B}_1^+}{\Delta q}\left(\hat{\tau}-\tau\right) + \frac{C_{\mathsf{B}}\sigma_R^2}{\Delta q}\left(\hat{q}_r^{\prime\prime+} - q_r^{\prime\prime+}\right)\left(m_t^{\prime+}-\tau\right) + Rem_\tau \\
= \;& O_p\left(\left(nb^6\right)^{-1/2} + b + \left(nh^2\right)^{-1/2} + h^2\right),
\end{aligned} \tag{23}$$

where $\mathbb{B}\left[\hat{\mathsf{B}}_2^+\right] = O(b)$ 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 $K_b(\underline{X}_i - \underline{x}) \equiv (b^2\sigma_R\sigma_T)^{-1}K\left(\frac{T_i - q^+(u)}{b\sigma_T}\right)K\left(\frac{R_i - r_0}{b\sigma_R}\right)$, $W_2 \equiv diag\{1,1,1,2,1,2\}$, $B_n = diag\{1,b,b,b^2,b^2,b^2\}$, $\underline{X}_i \equiv (T_i/\sigma_T, R_i/\sigma_R)^\top$, $\underline{x} \equiv (q^+(u)/\sigma_T, r_0/\sigma_R)^\top$,
$\mu\left(\underline{X}\right) \equiv \left(1, R/\sigma_R, T/\sigma_T, R^2/\sigma_R^2, RT/(\sigma_R\sigma_T), T^2/\sigma_T^2\right)^\top$, and
$\beta_2(\underline{x}) \equiv \left(m^+, m_r^{\prime+}\sigma_R, m_t^{\prime+}\sigma_T, m_r^{\prime\prime+}\sigma_R^2, \lim_{r\to r_0^+}\frac{\partial^2 m(t,r)}{\partial r\partial t}\big|_{t=q^+(u)}\sigma_R\sigma_T, m_t^{\prime\prime+}\sigma_T^2\right)^\top$. β_{n2}^{*-} is defined as β_{n2}^{*+} by replacing Z_i with $1-Z_i$ and $+$ with $-$.

Together with Lemma 5, the asymptotic linear representation for $\hat{\tau}^{bc}$ is

$$\begin{aligned}
\hat{\tau}^{bc} - \tau &= \hat{\tau} - \tau - h^2\left(\hat{\mathsf{B}}_\tau - \mathsf{B}_\tau\right) - h^2\mathsf{B}_\tau \\
&= \frac{1}{n}\sum_{i=1}^n IF_{\tau^{bc}i} - h^2\Bigg(\frac{\mathbb{B}\left[\hat{\mathsf{B}}_2^+ - \hat{\mathsf{B}}_2^-\right]}{\Delta q} - \frac{\mathsf{B}_1^+ - \mathsf{B}_1^-}{\Delta q}\left(\hat{\tau}-\tau\right) \\
&\quad + \left(\hat{q}_r^{\prime\prime+} - q_r^{\prime\prime+}\right)\frac{C_{\mathsf{B}}\sigma_R^2}{\Delta q}\left(m_t^{\prime+}-\tau\right) - \left(\hat{q}_r^{\prime\prime-} - q_r^{\prime\prime-}\right)\frac{C_{\mathsf{B}}\sigma_R^2}{\Delta q}\left(m_t^{\prime-}-\tau\right)\Bigg) \\
&\quad + O_p\left(\frac{h^2}{b^2}\left(\frac{\log n}{nb^2}\right)^{3/4}\right) + Rem \\
&= \frac{1}{n}\sum_{i=1}^n IF_{\tau^{bc}i} + O_p\left(h^2 b + h^3 + \frac{h^2}{\sqrt{nb^5}} + \frac{h}{\sqrt{n}} + \frac{1}{\sqrt{n^2h^5}} + (1+\rho^2)\left(\frac{\log n}{nb^2}\right)^{3/4}\right),
\end{aligned}$$

where the influence function

$$\begin{aligned}
IF_{\tau^{bc}i} \equiv\; & \frac{1}{\Delta q}\Bigg\{Z_i\left(\phi_{2i}^+ + \Phi_{1i}^+\left(m_t^{\prime+}-\tau\right)\right) - \frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top\beta_{n2}^{*+} \\
& -(1-Z_i)\left(\phi_{2i}^- + \Phi_{1i}^-\left(m_t^{\prime-}-\tau\right)\right) + \frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top\beta_{n2}^{*-}\Bigg\}.
\end{aligned} \tag{24}$$

Next we derive the asymptotic variance

$$\mathbb{V}\left[\beta_{n2}^{*+}\right] = \frac{W_2 S_2^{-1} B_n^{-1}}{n f_{TR}^{+2}} \mathbb{V}\left[K_b(\underline{X}-\underline{x})\left(Y_i - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X}-\underline{x}\right)Z_i\right] B_n^{-1} S_2^{-1} W_2,$$

where the second moment term

$$\begin{aligned}
&\mathbb{V}\left[K_b(\underline{X}-\underline{x})\left(Y - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X}_i-\underline{x}\right)Z_i\right] \\
=&\int_{\mathcal{T}}\int_{r_0}^{\infty} K_b^2(\underline{X}-\underline{x})\mathbb{E}\left[\left(Y - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)^2 \middle| R, T\right]\mu\left(\underline{X}-\underline{x}\right)\mu\left(\underline{X}-\underline{x}\right)^\top f_{TR}(T,R)dTdR \\
=&\frac{1}{b^2\sigma_T\sigma_R}\int_{-\infty}^{\infty}\int_0^{\infty} K^2(\upsilon)K^2(s)\mathbb{E}\left[\left(Y - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)^2 \middle| T = q^+ + bs, R = r_0 + b\upsilon\right] \\
&\times \mu((bs,b\upsilon)^\top)\mu((bs,b\upsilon)^\top)^\top f_{TR}(q^+ + bs, r_0 + b\upsilon)d\upsilon ds \\
=&\frac{2\lambda_0^2}{b^2\sigma_T\sigma_R}\mathbb{V}\left[Y\middle|T = q^+, R = r_0\right] f_{TR}(q^+, r_0)e_1e_1^\top + O\left(b^{-1}\right).
\end{aligned}$$

Therefore

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

Thus the variance of $\hat{\mathsf{B}}_\tau$ 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 $(C_{\mathsf{B}}e_4 + \kappa_2 e_6)^\top W_2 = 2\left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^\top$, we obtain $\mathsf{V}_{\mathsf{B}_\tau}(u)$ defined in (21) by showing that the sample above the cutoff contributes

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

For the covariance term,

$$\begin{aligned}
\mathbb{C}\left[Z_i\phi_{2i}^+, \beta_{n2}^{*+}\right] \;=\; & \frac{1}{n}\frac{2W_2 S_2^{-1} B_n^{-1}}{f_{TR}^{+2}\left(\kappa_2 - 2\kappa_1^2\right)}\mathbb{E}\Big[K_h\left(T - q^+, R - r_0\right)K_b\left(T - q^+, R - r_0\right) \\
& \times\left(Y - \left(m^+ + m_r'^+\left(R - r_0\right) + m_t'^+\left(T - q^+\right)\right)\right)\left(Y - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right) \\
& \times\left(\kappa_2 - \kappa_1(R - r_0)/h\right)\mu\left(\underline{X}-\underline{x}\right)Z\Big],
\end{aligned}$$

where the expectation term is

$$\begin{aligned}
&\int_{\mathcal{T}}\int_{r_0}^{\infty}\mathbb{E}\Big[\left(Y - \left(m^+ + m_r'^+\left(R - r_0\right) + m_t'^+\left(T - q^+\right)\right)\right)\left(Y - \mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\Big|T, R\Big] \\
&\times K_h\left(T - q^+, R - r_0\right)K_b\left(T - q^+, R - r_0\right)\left(\kappa_2 - \kappa_1(R - r_0)/h\right)\mu\left(\underline{X}-\underline{x}\right) f_{TR}(T,R)dRdT \\
=\;& \frac{2\sigma^{2+}}{h^2\sigma_T\sigma_R}\left(\kappa_2\left(\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\right)^2 - \frac{\kappa_1}{\rho}\int_0^{\infty}\upsilon K(\upsilon/\rho)K(\upsilon)d\upsilon\int_0^{\infty}K(\upsilon/\rho)K(\upsilon)d\upsilon\right)e_1 f_{TR}^+ \\
&+ O\left(bh^{-2}\right).
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{C}\left[Z_i\phi_{2i}^+, -\frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \beta_{n2}^{*+}\right]\frac{1}{(\Delta q)^2}\\
= \;& -\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \frac{1}{(\Delta q)^2}\rho^2\mathbb{C}\left[Z_i\phi_{2i}^+, \beta_{n2}^{*+}\right]\\
= \;& -\frac{1}{nb^2\sigma_T\sigma_R}\frac{\sigma^{2+}}{f_{TR}^+}\frac{8\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}e_1}{\left(\kappa_2-2\kappa_1^2\right)(\Delta q)^2}\left\{\kappa_2\left(\int_0^\infty K(\upsilon/\rho)K(\upsilon)d\upsilon\right)^2\right.\\
& \left.-\frac{\kappa_1}{\rho}\int_0^\infty \upsilon K(\upsilon/\rho)K(\upsilon)d\upsilon\int_0^\infty K(\upsilon/\rho)K(\upsilon)d\upsilon\right\}+O\left((nb)^{-1}\right).
\end{aligned}$$

A similar derivation yields

$$\begin{aligned}
&\mathbb{C}\left[Z_i\Phi_{1i}^+\left(m_t'^+-\tau\right), -\frac{h^2}{b^2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \beta_{n2}^{*+}\right]\frac{1}{(\Delta q)^2}\\
= \;& -\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \frac{\left(m_t'^+-\tau\right)}{(\Delta q)^2}\rho^2\mathbb{C}\left[Z_i\Phi_{1i}^+, \beta_{n2}^{*+}\right]=o\left((nb^2)^{-1}\right).
\end{aligned}$$

Thus the covariance between the $\hat{\mathsf{B}}_\tau$ 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 (22) by showing that the sample above the cutoffs contributes

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

Therefore $\mathsf{V}_\tau^{bc} = O\left(\left(nh^2\right)^{-1}+h^4\left(nb^6\right)^{-1}\right)$ and $\mathbb{B}\left[\hat{\tau}^{bc}\right] = -h^2\mathbb{B}\left[\hat{\mathsf{B}}_\tau\right]+O(h^3) = O\left(h^3+h^2b\right)$ is of smaller order by the bandwidth conditions $n\min\{h^6,b^6\}\max\{h^2,b^2\}\to 0$. We have the asymptotic linear representation in (24), $\hat{\tau}^{bc}-\tau = n^{-1}\sum_{i=1}^n IF_{\tau^{bc}i}+o_p\left((nh^2)^{-1/2}+h^2(nb^6)^{-1/2}\right)$. To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. When $h/b\to\rho\in[0,\infty)$, (24) implies $\sqrt{nh^2}\left(\hat{\tau}^{bc}-\tau-\mathbb{B}\left[\hat{\tau}^{bc}\right]\right) = \sqrt{nh^2}n^{-1}\sum_{i=1}^n IF_{\tau^{bc}i}+o_p(1)$. The Lyapounov condition $\left(\sum_{i=1}^n\mathbb{V}\left[IF_{\tau^{bc}i}\right]\right)^{-3/2}\sum_{i=1}^n\mathbb{E}\left[|IF_{\tau^{bc}i}|^3\right] = O\left((nh^{-2})^{-3/2}\right)\sum_{i=1}^n\mathbb{E}\left[|IF_{\tau^{bc}i}|^3\right] = O\left(n^{-1/2}h^3\left(h^{-4}+\rho^6b^{-4}\right)\right) = O\left(\left(nh^2\right)^{-1/2}\right) = o(1)$ holds. Then $\sqrt{nh^2}\left(\hat{\tau}^{bc}(u;h,b)-\tau(u)\right)\to_d\mathcal{N}\left(0,\mathsf{V}_\tau^{bc}(u)\right)$.

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

Proof of Theorem 7 Theorem 7 follows by minimizing the AMSE implied by Lemma 5. The asymptotic distribution becomes $n^{1/3}\left(\hat{\tau}(u)-\tau(u)\right)\to_d\mathcal{N}\left(c_u^2\mathsf{B}_\tau(u), c_u^{-2}\mathsf{V}_\tau(u)\right)$, where $c_u\equiv\left(\mathsf{V}_\tau(u)/\left(2\mathsf{B}_\tau^2(u)\right)\right)^{1/6}$.

B.4 Proofs of Theorem 4, Theorem 6, and Theorem 8 for π^*

Proof of Theorem 4 Lemma 6 implies Theorem 4 by letting the bias be of smaller order, i.e., $\sqrt{nh}h^2\mathsf{B}_\pi = o(1)$.

Proof of Theorem 6 The following derives the terms $\mathsf{V}_{\mathsf{B}_\pi}$ and C_π in the asymptotic variance of $\sqrt{nh}\hat{\pi}^{bc}$, which are due to bias correction. They are defined as follows.

$$\mathsf{V}_{\mathsf{B}_\pi} \equiv \mathsf{V}_\pi^m C_\mathsf{V}^{-1} 4 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top S_2^{-1} \Lambda_2 S_2^{-1} \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right) \text{ and} \tag{25}$$

$$\mathsf{C}_\pi \equiv -\mathsf{V}_\pi^m \frac{8 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top S_2^{-1}}{C_\mathsf{V}\left(\kappa_2 - 2\kappa_1^2\right)} \int_0^\infty K(\upsilon) K(\upsilon/\rho) \mathbf{v}_2 (\kappa_2 - \kappa_1 \upsilon/\rho) d\upsilon, \tag{26}$$

where $\mathbf{v}_2 \equiv (1, \upsilon, 0, \upsilon^2, 0, 0)^\top$. For $\rho = 1$, the integration in C_π becomes $\left(\kappa_2\lambda_0 - \kappa_1\lambda_1, \kappa_2\lambda_1 - \kappa_1\lambda_2, 0, \kappa_2\lambda_2 - \kappa_1\lambda_3, 0, 0\right)^\top$.

Similar to the proof of Lemma 6, the proof below is for the estimator using the infeasible trimming function $\chi(u)$, denoted by $\tilde{\mathsf{B}}_\pi \equiv \int_\mathcal{U} \hat{\mathsf{B}}_\tau(u)\tilde{\omega}^*(u)\, du + \int_\mathcal{U} \left(\hat{\mathsf{B}}_1^+(u) - \hat{\mathsf{B}}_1^-(u)\right)\left(\hat{\tau}(u) - \hat{\pi}\right)\tilde{\omega}^*(u) / \Delta\hat{q}(u) du$. Following the same arguments as in Lemma 7, we have $\tilde{\pi}^{bc} - \hat{\pi}^{bc} = o_p((nh)^{-1/2})$.

First derive the asymptotic 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

$$\begin{aligned} IF_{\pi^{bc}i} \equiv Z_i \Bigg\{ & \Phi_{21i}^+(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right) + \int_\mathcal{U} \Phi_{1i}^+(u)\Lambda^+(u) du - \rho^2 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top \Phi_{22i}^+(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right) \Bigg\} \\ & - (1 - Z_i)\Bigg\{ \Phi_{21i}^-(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right) + \int_\mathcal{U} \Phi_{1i}^-(u)\Lambda^-(u) du \\ & - \rho^2 \left(C_\mathsf{B} e_4 + \kappa_2 e_6\right)^\top \Phi_{22i}^-(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right) \Bigg\} \end{aligned} \tag{27}$$

with $\Phi_{21i}^\pm(h)$ defined in Lemma 6 and

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

To derive $\Phi_{22i}^\pm(b)$, linearize the bias estimator $\hat{\mathsf{B}}_\pi - \mathsf{B}_\pi$ to be

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

where the leading term in the remainder terms Rem_π is $O_p(\|\hat{\mathsf{B}}_\tau - \mathsf{B}_\tau\|_\infty \|\Delta\hat{q} - \Delta q\|_\infty) = O_p\big(((nb^6)^{-1/2} +$

$b + (nh)^{-1/2} + h^2\big)\big((nh)^{-1/2} + h^2\big)\big)$. And the terms associated with the cross products of $\hat{\mathsf{B}}_1^+ - \mathsf{B}_1^+$, $\Delta\hat{q} - \Delta q$, $\hat{\tau} - \tau$, and $\hat{\pi}^* - \pi^*$ in Rem_π are of smaller order. Together with Lemma 5 and Lemma 6,

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

By the same argument in the proof of Lemma 6, the third term associated with $IF_{\tau i}(u)$ is $O_p\big(h^2\big((nh)^{-1/2} + h^2\big)\big)$, which is of smaller order. We focus on the second term $\int_{\mathcal{U}}\big(\hat{\mathsf{B}}_\tau(u) - \mathsf{B}_\tau(u)\big)w(u)du$ using the expansion in (23). One can show that

$$\begin{aligned}
& \int_{\mathcal{U}}\frac{w^*(u)}{\Delta q(u)}\left\{b^{-2}\left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^\top\beta_{n2}^{*+}(u) + \mathbb{B}\left[\hat{\mathsf{B}}_2^+\right] - \mathsf{B}_1^+(u)\left(\hat{\tau}(u) - \tau(u)\right)\right\}du \qquad (28) \\
= \;& O_p\left((nb^5)^{-1/2} + b + (nh)^{-1/2} + h^2\right).
\end{aligned}$$

To see why, the second term associated with $\mathbb{B}\left[\hat{\mathsf{B}}_2^+\right]$ is $O(b)$ and the third term associated with $\hat{\tau} - \tau$ is $O_p\left((nh)^{-1/2} + h^2\right)$ by the proof of Lemma 6 with the additional weight $\mathsf{B}_1^+(U_i)/\Delta q(U_i)$. For the first term in (28), we use the same arguments as those in deriving (18) in the proof of Lemma 4. By change of variable $v = q^+(u)$ and $s = (v - T_i)/b_T$, we have

$$\begin{aligned}
& \int_{\mathcal{U}}\frac{w^*(u)}{\Delta q(u)}\beta_{n2}^{*+}(u)du \\
= \;& \frac{W_2 S_2^{-1}B_n^{-1}}{n}\sum_{i=1}^n\int_{\mathcal{U}}\frac{w^*(u)}{f_{TR}^+(u)\Delta q(u)}K_b(\underline{X}_i - \underline{x})\left(Y_i - \mu\left(\underline{X}_i - \underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left(\underline{X}_i - \underline{x}\right)du\,Z_i \\
= \;& \frac{W_2 S_2^{-1}B_n^{-1}}{n}\sum_{i=1}^n K_b(R_i - r_0)\int_{-\infty}^{\infty}\frac{w^*\left(F_{T_1|R}(T_i + b_T s, r_0)\right)}{\Delta q\left(F_{T|R}(T_i + b_T s, r_0)\right)}\frac{K(s)}{f_R(r_0)}\mathbf{1}\left(F_{T_1|R}(T_i + b_T s, r_0) \in \mathcal{U}\right) \\
& \times\left(Y_i - \mu\left((R_i - r_0, b_T s)^\top\right)^\top W_2^{-1}\beta_2(\underline{x})\right)\mu\left((R_i - r_0, b_T s)^\top\right)ds\,Z_i \\
= \;& \frac{1}{n}\sum_{i=1}^n Z_i\Phi_{22i}^+(b)\mathbf{1}\left(U_i \in \mathcal{U}\right)(1 + O_p(b^2)).
\end{aligned}$$

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

$$
\begin{aligned}
&\mathbb{E}\left[\Phi_{22i}^{+2}(b)\mathbf{1}\left(U_i \in \mathcal{U}\right) Z_i\right] \\
=&W_2 S_2^{-1}\mathbb{E}\Bigg[\left(Y-\left(m^+(U)+m_r'^+(U)\left(R-r_0\right)+\frac{1}{2}m_r''^+(U)\left(R-r_0\right)^2\right)\right)^2\left(\frac{w^*(U)}{\Delta q(U)}\right)^2\mathbf{1}\left(U\in\mathcal{U}\right) \\
&\times\left(1,\frac{R-r_0}{b},0,\left(\frac{R-r_0}{b}\right)^2,0,0\right)^\top\left(1,\frac{R-r_0}{b},0,\left(\frac{R-r_0}{b}\right)^2,0,0\right)K_b^2\left(R-r_0\right)Z\Bigg]\frac{S_2^{-1}W_2}{f_R^2(r_0)} \\
=&W_2 S_2^{-1}\int_0^\infty\int_{\mathcal{T}}\mathbf{1}\left(F_{T_1|R}(T|r_0)\in\mathcal{U}\right)\mathbf{v}^\top\mathbf{v}K^2\left(\upsilon\right)f_{TR}(T,r_0+b\upsilon) \\
&\times\mathbb{E}\left[\left(Y-\left(m^+(U)+m_r'^+(U)\left(b\upsilon\right)+\frac{1}{2}m_r''^+(U)\left(b\upsilon\right)^2\right)\right)^2\middle|U=F_{T_1|R}(T|r_0),R=r_0+b\upsilon\right]dTd\upsilon \\
&\times\frac{S_2^{-1}W_2}{b\sigma_R B^2 f_R^2(r_0)} \\
=&\frac{\mathbb{E}\left[\mathbb{V}\left[Y\middle|U,R\right]\mathbf{1}\left(U\in\mathcal{U}\right)\middle|R=r_0^+\right]}{b\sigma_R B^2 f_R(r_0)}W_2S_2^{-1}\Lambda_2S_2^{-1}W_2+o(b^{-1})=O(b^{-1}).
\end{aligned}
$$

Thus the first term in (28) is $O_p((nb^5)^{-1/2})$. Then $\rho^2\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top\Phi_{22i}^+(b)$ 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 (25) by showing that the sample above the cutoff contributes

$$
\frac{4\int_{\mathcal{U}}\sigma^{2+}(u)du}{\sigma_R B^2 f_R(r_0)}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top S_2^{-1}\Lambda_2S_2^{-1}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right).
$$

The asymptotic covariance is $\lim_{n\to\infty}-2h\rho^2\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top\mathbb{C}\left[Z_i\Phi_{21i}^+(h)\mathbf{1}\left(U_i\in\mathcal{U}\right),Z_i\Phi_{22i}^+(b)\mathbf{1}\left(U_i\in\mathcal{U}\right)\right]=\lim_{n\to\infty}-2h\rho^2\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top\mathbb{E}\left[Z_i\Phi_{21i}^+(h)\Phi_{22i}^+(b)\mathbf{1}\left(U_i\in\mathcal{U}\right)\right]$, where

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

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

$$
\begin{aligned}
&\frac{1}{\sigma_R\rho b}\int_0^\infty\int_{\mathcal{T}}K(\upsilon)K(\upsilon/\rho)\mathbb{V}\left[Y|U=F_{T_1|R}(T,r_0),R=r_0+\upsilon b\right]\mathbf{v}_2\left(\kappa_2-\kappa_1\upsilon/\rho\right) \\
&\times\mathbf{1}\left(F_{T_1|R}(T,r_0)\in\mathcal{U}\right)f_{TR}(T,r_0+\upsilon b)dTd\upsilon=O\left((\rho b)^{-1}\right).
\end{aligned}
$$

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 C_π defined in (26) by showing that the sample above the cutoff contributes

$$-\frac{8\int_{\mathcal{U}} \sigma^{2+}(u)du}{\sigma_R B^2 f_R(r_0)\left(\kappa_2 - 2\kappa_1^2\right)} \left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^\top S_2^{-1} \int_0^\infty K(v)K(v/\rho)\mathbf{v}_2(\kappa_2 - \kappa_1 v/\rho)dv.$$

Therefore $\mathbb{V}\left[\hat{\pi}^{bc}\right] = O\left((nh)^{-1} + (nb^5h^{-4})^{-1}\right)$ and $\mathbb{B}\left[\hat{\pi}^{bc}\right] = O\left(h^2(h+b)\right)$ 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 $h/b \to \rho \in [0, \infty)$, (27) implies $\sqrt{nh}\left(\hat{\pi}^{bc} - \pi^* - \mathbb{B}\left[\hat{\pi}^{bc}\right]\right) = \sqrt{nh}n^{-1}\sum_{i=1}^n IF_{\pi^{bc}i} + o_p(1)$. The Lyapounov condition $\left(\sum_{i=1}^n \mathbb{V}\left[IF_{\pi^{bc}i}\right]\right)^{-3/2} \sum_{i=1}^n \mathbb{E}\left[|IF_{\pi^{bc}i}|^3\right] = O\left((nh^{-1})^{-3/2}\right)\sum_{i=1}^n \mathbb{E}\left[|IF_{\pi^{bc}i}|^3\right] = O\left(n^{-1/2}h^{3/2}h^{-2}\right) = O\left((nh)^{-1/2}\right) = o(1)$ holds. Then $\sqrt{nh}\left(\hat{\pi}^{bc}(h,b) - \pi^*\right) \to_d \mathcal{N}\left(0, \mathsf{V}_\pi^{bc}\right)$.

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

Proof of Theorem 8 Theorem 8 follows by minimizing the AMSE implied by Lemma 6. The asymptotic distribution becomes $n^{2/5}\left(\hat{\pi}^* - \pi^*\right) \to_d \mathcal{N}\left(c_\pi^2 \mathsf{B}_\pi, c_\pi^{-1}\mathsf{V}_\pi\right)$, where $c_\pi \equiv \left(\mathsf{V}_\pi / \left(4\mathsf{B}_\pi^2\right)\right)^{1/5}$.

C Bias, variance, and AMSE optimal bandwidth estimation

This section briefly 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 $h_\tau^*(u)$ and h_π^*. All discussion focuses on estimation of the unknown parameters defined above the cutoff. Those parameters below the cutoff can be estimated analogously.

C.1 Bias estimation

Consider the bias of $\widehat{\tau}(u)$. $\mathsf{B}_\tau(u) \equiv \left(\mathsf{B}_2(u) + \mathsf{B}_1^+(u)\left(m_t'^+(u) - \tau(u)\right) - \mathsf{B}_1^-(u)\left(m_t'^-(u) - \tau(u)\right)\right)\frac{1}{\Delta q(u)}$, where $\mathsf{B}_1^\pm(u) \equiv C_{\mathsf{B}} q_r''^\pm(u)\sigma_R^2$ and $\mathsf{B}_2(u) \equiv C_{\mathsf{B}}\left(m_r''^+(u) - m_r''^-(u)\right)\sigma_R^2 + \kappa_2\left(m_t''^+(u) - m_t''^-(u)\right)\sigma_T^2$.

C_{B} is a constant depending on the kernel function. For the Uniform kernel, $C_{\mathsf{B}} = -1/12$. $\Delta q(u)$ is the denominator of $\tau(u)$, which is estimated in Step 1 of the estimation procedure described in the main text. $m_t'^+(u)$ can be estimated by $\widehat{b}_2^+(u)$ in Step 2 of the local linear estimation described in the main text.

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

following local quadratic quantile regression with a chosen bandwidth b,

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

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

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

Plugging in C_{B} and estimates of $m_t'^{\pm}(u)$, $q_r''^{\pm}(u)$, $m_r''^{\pm}(u)$, and $m_t''^{\pm}(u)$, one obtains $\widehat{\mathsf{B}}_\tau(u)$.

Consider next the bias of $\hat{\pi}^*$. $\mathsf{B}_\pi\equiv\int_{\mathcal{U}}\mathsf{B}_\tau(u)w^*(u)\,du+\int_{\mathcal{U}}\left(\mathsf{B}_1^+(u)-\mathsf{B}_1^-(u)\right)(\tau(u)-\pi^*)\frac{w^*(u)}{\Delta q(u)}du$. $\mathsf{B}_\tau(u)$ and $\mathsf{B}_1^{\pm}(u)$ are estimated in the above. $\Delta q(u)$ is estimated in Step 1 estimation described in the main text. The weighting function $w^*(u)$ is estimated in Step 4. Plugging in these estimates, one obtain $\widehat{\mathsf{B}}_\pi$.

C.2 Variance estimation

Now turn to the standard error of $\widehat{\tau}(u)$. By Theorem 3, $\mathsf{V}_\tau(u)\equiv\frac{2\lambda_0 C_{\mathsf{V}}}{(\Delta q(u))^2 f_R(r_0)\sigma_R\sigma_T}\left(\frac{\sigma^{2+}(u)}{f_{T|R}^+(u)}+\frac{\sigma^{2-}(u)}{f_{T|R}^-(u)}\right)$.[26] 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 $f_R(r_0)$, σ_R, σ_T, $f_{T|R}^{\pm}(u)$, and $\sigma^{2\pm}(u)$.

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

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

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

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

[26] The influence function for $\widehat{\tau}(u)$ is provided in Lemma 5. If desired, one can alternatively estimate the influence function and then estimate the variance of $\widehat{\tau}(u)$ by the sample variance of the estimated influence function. Similarly we can use the influence function for $\hat{\pi}^*$ provided in Lemma 6 to estimate the variance of $\hat{\pi}^*$.

Plugging in all 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 $\hat{\tau}^{bc}(u)$. By Theorem 5, $\mathsf{V}^{bc}_{\tau,n}(u) \equiv \frac{\mathsf{V}_\tau(u)}{nh^2} + \frac{\mathsf{V}_{\mathsf{B}_\tau}(u)+\rho^{-5}\mathsf{C}_\tau(u;\rho)}{nh^2\rho^{-6}}$. Estimation of $\mathsf{V}_\tau(u)$ is discussed above. For the Uniform kernel, $\mathsf{V}_{\mathsf{B}_\tau}(u) = 9.765625\mathsf{V}_\tau(u)$ by equation (21), 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 (22). Plugging in $\widehat{\mathsf{V}}_\tau(u)$ for a chosen ρ, one can obtain $\widehat{\mathsf{V}}^{bc}_{\tau,n}(u)$.

Consider further the standard error of $\hat{\pi}^*$. By Theorem 4, $\mathsf{V}_\pi \equiv \mathsf{V}^m_\pi + \mathsf{V}^q_\pi$, where $\mathsf{V}^m_\pi \equiv \frac{C_{\mathsf{V}}\left(\mathbb{E}\left[\mathbb{V}[Y|T,R]|R=r_0^+\right]+\mathbb{E}\left[\mathbb{V}[Y|T,R]|R=r_0^-\right]\right)}{f_R(r_0)\left(\int_{\mathcal{U}}|\Delta q(u)|du\right)^2}$ and $\mathsf{V}^q_\pi \equiv \frac{C_{\mathsf{V}}}{f_R(r_0)}\int_{\mathcal{U}}\int_{\mathcal{U}}(\min\{u,v\}-vu)\left(\frac{\Lambda^+(u)\Lambda^+(v)}{f^+_{T|R}(u)\,f^+_{T|R}(v)} + \frac{\Lambda^-(u)\Lambda^-(v)}{f^-_{T|R}(u)\,f^-_{T|R}(v)}\right)dvdu$ with $\Lambda^\pm(u) \equiv \frac{\left(m'^{\pm}_t(u)-\pi^*\right)w^*(u)}{\Delta q(u)}$. Estimation of $\Delta q(u)$, $f_R(r_0)$, and $f^\pm_{T|R}(u)$ is described at the beginning of this section. $w^*(u)$ is estimated in Step 4 estimation in the main text. The only unknowns involved in V_π are $\mathbb{E}\left[\mathbb{V}\left[Y|T,\,R=r_0^\pm\right]|R=r_0^\pm\right]$ and $m'^{\pm}_t(\mathrm{u})$, which appear in $\Lambda^\pm(\mathrm{u})$, $\mathrm{u} = u, v$. Consider first estimating $\mathbb{E}\left[\mathbb{V}\left[Y|T,\,R=r_0^+\right]|R=r_0^+\right]$. By the law of iterated expectations, $\mathbb{E}\left[\mathbb{V}\left[Y|T,\,R=r_0^+\right]|R=r_0^+\right] = \mathbb{E}\left[\mathbb{E}\left[\left(Y-\mathbb{E}Y|T,\,R=r_0^+\right)^2|T,\,R=r_0^+\right]|R=r_0^+\right] = \mathbb{E}\left[\left(Y-\mathbb{E}\left[Y|T,\,R=r_0^+\right]\right)^2|R=r_0^+\right]$. One can first estimate $\mathbb{E}\left[Y|T=T_i,\,R=r_0^+\right]$ by a local linear regression, and then estimate $\mathbb{E}\left[\mathbb{V}\left[Y|T,\,R=r_0^+\right]|R=r_0^+\right]$ by $\widehat{\theta}_0^+$ from the following local linear regression

$$\begin{aligned}&\left(\widehat{\theta}_0^+(u),\widehat{\theta}_1^+(u)\right)\\ =\;&\arg\min_{\theta_0^+,\theta_1^+}\sum_{\{i\,R_i\geq r_0\}}K\left(\frac{R_i-r_0}{b\sigma_R}\right)\left(\left(Y_i-\hat{\mathbb{E}}\left[Y|T=T_i,\,R=r_0^+\right]\right)^2-\theta_0^+-\theta_1^+(R_i-r_0)\right)^2.\end{aligned}$$

$m'^{\pm}_t(u)$ can be estimated by $\widehat{b}^\pm_2(u)$ from Step 2 local linear regression described in the main text. Plugging in the estimates of $\Delta q(u)$, $m'^{\pm}_t(u)$, and $w^*(u)$, one get estimates of $\Lambda^\pm(\mathrm{u})$.

Further plugging in the estimates of $\Delta q(u)$, $f_R(r_0)$, $f^\pm_{T|R}(u)$, $\Lambda^\pm(u)$, $\mathbb{E}\left[\mathbb{V}\left[Y|T,\,R=r_0^\pm\right]|R=r_0^\pm\right]$, and the constant C_{V}, and replacing integration by summation, one can obain $\widehat{\mathsf{V}}_\pi = \widehat{\mathsf{V}}^m_\pi + \widehat{\mathsf{V}}^q_\pi$.

Consider lastly the standard error of the bias-corrected estimator $\hat{\pi}^{bc}$. By Theorem 6, $\mathsf{V}^{bc}_{\pi,n} \equiv \frac{\mathsf{V}_\pi}{nh} + \frac{\mathsf{V}_{\mathsf{B}_\pi}+\rho^{-3}\mathsf{C}_\pi}{nh\rho^{-5}}$. Estimation of V_π is provided above. For the Uniform kernel, $\mathsf{V}_{\mathsf{B}_\pi} = 1.641\mathsf{V}^m_\pi$ by equation (25) and $\mathsf{C}_\pi = (3.125\rho - 2.5\rho^3)\mathsf{V}^m_\pi$ when $\rho \leq 1$, and $\mathsf{C}_\pi = (2.5 - 1.875/\rho)\mathsf{V}^m_\pi$ when $\rho > 1$ by equation (26). Estimation of V^m_π is discussed above. Plugging in the estimates of V_π and V^m_π and the constant C_π, one obtain $\widehat{\mathsf{V}}^{bc}_{\pi,n}$.

C.3 Optimal bandwidth estimation

Given consistent estimates of $\mathsf{B}_\tau(u)$, $\mathsf{V}_\tau(u)$, B_π, and V_π in the previous section, by the plug-in rule, one can obtain the estimated bandwidth $\widehat{h}^*_\tau(u) = \left(\widehat{\mathsf{V}}_\tau(u)/\left(2\widehat{\mathsf{B}}^2_\tau(u)\right)\right)^{1/6}n^{-1/6}$ and $\hat{h}^*_\pi = \left(\widehat{\mathsf{V}}_\pi/\left(4\widehat{\mathsf{B}}^2_\pi\right)\right)^{1/5}n^{-1/5}$.

D Supplementary empirical analysis

Table A3 Impacts of log(capital) on bank outcomes (undersmoothing)

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	0.843	(0.249)***	-0.157	(0.247)	0.136	(0.120)
	0.12	0.777	(0.230)***	-0.223	(0.216)	0.133	(0.117)
	0.14	0.734	(0.237)***	-0.266	(0.225)	0.125	(0.136)
	0.16	0.687	(0.282)**	-0.313	(0.260)	0.132	(0.151)
	0.18	0.677	(0.276)**	-0.323	(0.249)	0.107	(0.147)
	0.20	0.681	(0.257)***	-0.319	(0.229)	0.103	(0.139)
	0.22	0.665	(0.265)**	-0.335	(0.235)	0.102	(0.145)
	0.24	0.639	(0.249)**	-0.361	(0.221)	0.088	(0.139)
WQ-LATE		0.700	(0.650)	-0.300	(0.608)	0.116	(0.354)

Note: The first panel presents estimated Q-LATEs at equally spaced quantiles; The last row presents the estimated WQ-LATEs; The bandwidths are set to be $h = 4\sigma_R n^{-0.23} = 1039.5$ for R and $h_T = 4\sigma_T n^{-0.23} = 0.3905$, which satisfies the undersmoothing conditions for the Q-LATE or WQ-LATE estimator in Theorems 3 and 4; The trimming thresholds are determined by using a preliminary bandwidth The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R$ or 779.6; Analytical standard errors are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

Table A4 Impacts of log (capital) on bank outcomes (bias-corrected estimates)

Q-LATE	Quantile	Log(assets)		Log(leverage)		Suspension	
	0.10	0.949	(0.295)***	-0.051	(0.282)	0.005	(0.132)
	0.12	0.915	(0.261)***	-0.085	(0.247)	-0.018	(0.123)
	0.14	0.899	(0.276)***	-0.101	(0.254)	-0.017	(0.128)
	0.16	0.862	(0.356)**	-0.138	(0.313)	-0.033	(0.153)
	0.18	0.858	(0.360)**	-0.142	(0.306)	-0.064	(0.150)
	0.20	0.871	(0.362)**	-0.129	(0.304)	-0.070	(0.151)
	0.22	0.819	(0.351)**	-0.181	(0.295)	-0.077	(0.153)
	0.24	0.865	(0.340)**	-0.135	(0.282)	-0.091	(0.143)
	0.26	0.883	(0.335)***	-0.117	(0.279)	-0.089	(0.141)
WQ-LATE		0.873	(0.718)	-0.127	(0.655)	-0.051	(0.344)

Note: The first panel presents the bias-corrected estimates of Q-LATEs at equally spaced quantiles; The last row presents the bias-corrected estimates of WQ-LATEs; The standardized AMSE optimal bandwidth for the WQ-LATE estimator is $h_\pi^* = 0.91$ (the standardized AMSE optimal bandwidth for the Q-LATE estimator h_τ^* ranges from 0.72 to 1.1); The bandwidths in the estimation are then set to be $h_R = h_\pi^* \sigma_R = 1108.0$ and $h_T = h_\pi^* \sigma_T = 0.4173$; The bandwidths used to estimate the biases are 2 times of the main bandwidths; The trimming thresholds are determined by using a preliminary bandwidth for R equal to $3/4h_R = 831.0$; The bandwidths used to estimate the biases are 2 times of the main bandwidths; Analytical standard errors are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

References

[1] Angrist, J. D. and M. Rokkanen (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," Journal of the American Statistical Association, 110(512), 1331-1344.

[2] Arai, Y., Y. Hsu, T. Kitagawa, I. Mourifie, and Y. Wan (2017): "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Design," Working paper.

[3] Bertanha, M. (2016): "Regression Discontinuity Design with Many Thresholds," Working Paper.

[4] Blundell, R., and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in Advances in Economics and Econometrics, Vol. II, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press, 312-357.

[5] Brinch, C. N., M. Mogstad, and M. Wiswall, (2017): "Beyond LATE with a Discrete Instrument," Journal of Political Economy, 125(4).

[6] Bugni F. and I.A. Canary (2018): "Testing Continuity of a Density via g-order Statistics in the Regression Discontinuity Design," working paper.

[7] Caetano, C. and J. C. Escanciano, (2017): "Identifying Multiple Marginal Effects with a Single Instrument," Working paper.

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

[9] Canay, I. A. and V. Kamat (2018): "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," The Review of Economic Studies, forthcoming.

[10] Cattaneo, M. D., B. R. Frandsen, and R. Titiunik (2015): "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," Journal of Causal Inference, 3(1), 1-24.

[11] Cattaneo, M. D., Jansson, M. and Ma, X. (2017): "Simple Local Polynomial Density Estimators," Tech. rep., Working Paper.

[12] Cattaneo, M. D., M. Jansson, and X. Ma (2018): "Manipulation Testing Based on Density Discontinuity," The Stata Journal, 18 (1), 234-261.

[13] Carneiro, P., J. J. Heckman, and E. Vytlacil, (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," Econometrica 78, 377-394.

[14] Chernozhukov, V., Fernández-Val, I., Galichon, A. (2010): "Quantile and Probability Curves without Crossing," Econometrica 78(3), 1093-1125.

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

[16] Chernozhukov, V. and Hansen, C. (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," Journal of Econometrics, 132, 491-525.

[17] Chernozhukov, V., G. Imebns, and W. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," Journal of Econometrics, 139, 4-14.

[18] Chesher, A. (2003): "Identification in Nonseparable Models," Econometrica, 71, 1405-1441.

[19] Chiang, D. H., Y. Hsu, and Y. Sasaki (2018): "Robust Uniform Inference for Quantile Treatment Effects in Regression Discontinuity Designs," Working paper.

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

[21] Dong, Y. and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," Review of Economics and Statistics, 97(5), 1081-1092.

[22] Dong, Y. (2016): "Alternative Assumptions to Identify LATE in Regression Discontinuity Designs," Working paper.

[23] Dong, Y. (2017): "Regression Discontinuity Designs with Sample Selection," Journal of Business and Economic Statistics.

[24] Dong, Y. and S., Shen (2018): "Testing for Rank Invariance or Similarity in Program Evaluation," The Review of Economics and Statistics, 100 (1), 78-85.

[25] Fang, Z. and A. Santos (2015): "Inference on Directionally Differentiable Functions," Working Paper, arXiv preprint arXiv:1404.3763.

[26] Feir, D., T. Lemieux, and V. Marmer (2016): "Weak Identification in Fuzzy Regression Discontinuity Designs," Journal of Business & Economic Statistics, 2 (34), 185-196.

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

[28] Florens, J. P., J. J., Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," Econometrica, 76, 1191-1206.

[29] Gerard, F., M. Rokkanen, and C. Rothe (2018): "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," Working paper.

[30] 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.

[31] Hall, P. (1992): The Bootstrap and Edgeworth Expansion. Springer, New York.

[32] Heckman, J. J. and E. J. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," Econometrica, 73(3), 669-738.

[33] Heckman, J. J. and E. J. Vytlacil (2007): "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," Handbook of Econometrics 6, in: J.J. Heckman and E.E. Leamer (ed.), 4779-4874.

[34] Horowitz, J. (2001): "The Bootstrap," Handbook of Econometrics V, in: J. J. Heckman and E. Leamer (ed.), 3159-3228.

[35] Horowitz, J. and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," Econometrica, 75, 1191-1208.

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

[37] 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.

[38] Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," Review of Economic Studies,76, 1071-1102.

[39] Ma, L. and R. Koenker (2006): "Quantile Regression Methods for Recursive Structural Equation Models," Journal of Econometrics, 134, 471-506.

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

[41] Newey, W. K. and J. L. Powell (2003): "Nonparametric Instrumental Variables Estimation," Econometrica, 71, 1565-1578.

[42] Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," Econometrica, 67, 565-603.

[43] Otsu, T., K.-L. Xu, and Y. Matsushita (2013): Estimation and inference of discontinuity in density, " Journal of Business & Economic Statistics, 31 507-524.

[44] Otsu, T., K.-L. Xu, and Y. Matsushita (2015): "Empirical Likelihood for Regression Discontinuity Design," Journal of Econometrics, 186, 94-112.

[45] Pinkse, J. (2000): "Nonparametric Two-Step Regression Functions When Regressors and Error Are Dependent," Canadian Journal of Statistics, 28, 289-300.

[46] Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66, 688-701.

[47] Thistlethwaite, D. L. and D. T. Campbell (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," The Journal of Educational Psychology, 51(6), 309-317.

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

[49] Qu, Z. and J. Yoon (2015b): "Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs," Working Paper.

[50] van der Vaart, A. (2000): Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[51] Vytlacil, E. (2002): “Independence, Monotonicity, and Latent Index Models: An Equivalence Result, ” Econometrica, 70(1), 331-341.