RIETI Discussion Paper Series 19-E-058

Regression Discontinuity Designs with a Continuous Treatment

DONG, Yingying
University of California Irvine

LEE, Ying-Ying
University of California Irvine

GOU, Michael
PricewaterhouseCoopers

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
https://www.rieti.go.jp/en/

RIETI Discussion Paper Series 19-E-058

August 2019

Regression Discontinuity Designs with a Continuous Treatment

Yingying Dong

Ying-Ying Lee

Michael Gou*

Abstract

Many empirical applications of regression discontinuity (RD) designs involve a continuous treatment. This paper establishes identification and bias-corrected robust inference for such RD designs. Causal identification is achieved by utilizing changes in the distribution of the continuous treatment at the RD threshold (including the usual mean change as a special case). Applying the proposed approach, we estimate the impacts of capital holdings on bank failure in the pre-Great Depression era. Our RD design takes advantage of the minimum capital requirements which change discontinuously with town size. We find that increased capital has no impacts on the long-run failure rates of banks.

Keywords: Regression discontinuity (RD) design, Continuous treatment, Control variable, Robust inference, Distributional change, Rank invariance, Rank similarity, Capital regulation, Bank failure

JEL classification: C21, C26, E58

This study is conducted as a part of the Project "Economic Analysis of the Development of the Nursing Care Industry in China and Japan" undertaken at the Research Institute of Economy, Trade and Industry (RIETI).

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

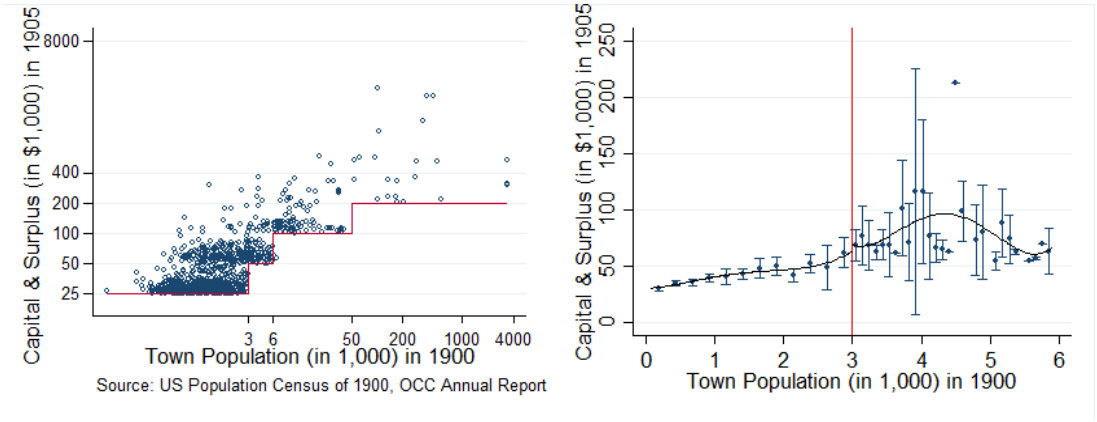

Figure 1: Scatter plot (left) and RD mean plot around the first threshold (right) of bank capital against town population

1 Introduction

Are banks less likely to fail when they hold more capital? To provide a credible estimate of the causal effect of capital holdings on bank failure, one needs some quasi-experimental variation in bank capital. As seen in Figure 1 (left), one potential source of variation is the relationship between minimum capital requirements and town size in the early 20th century of the United States – as town size crosses certain thresholds, minimum capital requirements (marked by the solid line) jump up and the bottom of the capital distribution shifts up correspondingly. Given this relationship, one may be tempted to apply the standard RD design to estimate the impacts of capital holdings, with town size as the running variable, capital holdings as the treatment variable, and bank failure as the outcome.

There are two issues with this approach. First, the standard RD design assumes a binary treatment, while the treatment variable here, capital holdings, is continuous. Hahn, Todd, and van der Klaauw (2001) show that under proper conditions, the RD local Wald ratio with a binary treatment identifies an average treatment effect for compliers at the RD threshold. Even when the same RD local Wald ratio is valid for a continuous treatment, the interpretation would be more complicated.[1] We discuss

[1]Unlike the standard RD design, which can be classified into sharp and fuzzy designs, there is generally no such distinction with a continuous treatment. The RD local Wald ratio would not reduce to a single difference even when everyone complies with the policy rule and changes treatment when crossing the RD threshold.

this point in greater detail later. Second and more importantly, the discontinuous relationship between minimum capital holdings and town size generates only a weak "first-stage" discontinuity in the relationship between mean capital holdings and town size. Figure 1 (right) plots the mean capital against town size along with the 95% confidence intervals. No significant changes are found in the mean capital at the first policy threshold, where most of the banks are present. Applying the standard RD design would be difficult.

Our empirical example is not alone. Many empirical applications of RD designs involve continuous treatments (see, for recent examples, Oreopoulos, 2006, Card, Chetty, and Weber, 2007, Schmieder, von Wachter, and Bender, 2012, Pop-Eleches and Urquiola, 2013, and Clark and Royer, 2013, Isen, Rossin-Slater, and Walker, 2017, and Corbi, Papaioannou, and Surico, 2018, Agarwal, Chomsisengphet, Mahoney, and Stroebel, 2018, Dell and Querubin, 2018). Empirical researchers typically apply the standard RD estimand for a binary treatment to applications with continuous treatments. Causal identification relies solely on the mean shift of the treatment.

In practice, public policies or welfare programs do not necessarily target the average units. Instead they may target some parts (e.g., top or bottom) or features of the treatment distribution. Examples include minimum school leaving age, minimum wage, maximum welfare benefits, government transfers that are capped at certain levels, or pollution ceiling set by the environmental protection agency. When policies shift these minimum or maximum requirements, focusing on the mean treatment change may miss the true sources of identification.

The obvious question is then how one might proceed when confronting RD designs with a continuous treatment. In this paper we answer that question. We establish causal identification and robust inference for the class of RD designs with a continous treatment. We show that identification can be achieved by utilizing any changes in the distribution of the treatment variable at the RD threshold. These include not only the usual mean change, but also changes at other quantiles (e.g., lower quantiles, as in the case of bank capital regulation). By focusing on where the true changes are in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects.

We first identify quantile specific local average treatment effects (Q-LATEs). These Q-LATEs provide information on treatment effect heterogeneity at different treatment

levels. We further identify a local weighted average treatment effect averaging over the treatment distribution (WQ-LATE). Importantly, the WQ-LATE estimand 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. In addition, we provide bias-corrected robust inference for Q-LATE and WQ-LATE, as well as their asymptotic mean squared error (AMSE) optimal bandwidths.

In the final part of the paper we quantify the impacts of capital holdings on bank failure, particularly among those banks targeted by the capital regulation. We show that while capital requirements induce small banks to hold more capital, these banks adjust their assets to lead to only a "scale-up" effect. On average a 1% increase in capital leads to almost a 1% increase in assets among those banks at lower quantiles of the capital distribution. Their leverages are not significantly lowered and their long run (up to 24 years) rates of suspension stay unchanged.

Our paper complements the existing studies of the classical RD design with a binary treatment.[2] Our paper further complements several important strands of literature on causal model identificaiton, which typically 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, Frölich and Melly 2013), and the marginal treatment effect (MTE) literature (see, e.g., Heckman and Vytlacil 2005, 2007, Carneiro, Heckman and Vytlacil, 2010). Important work discussing causal identification with a continuous 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

[2]For theoretical discussion of the standard RD design, see, e.g., Hahn, Todd, and van der Klaauw (2001), Porter (2003), Lee (2009), Imbens and Kalyanaraman (2012), Frandsen, Frölich, 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 (2019), Arai et al. (2017), Chiang, Hsu, and Sasaki (2018), Bugni and Canay (2018), Canay and Kamat (2018), Cattaneo, Jansson, and Ma (2018), and Gerard, Rokkanen, and Rothe (2018).

Torgovitsky, 2015). In contrast, we allow for multidimensional unobservables in both the first-stage and outcome equations and exemplify identification with a binary 'IV'.[3]

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 treatment response heterogeneity in covariates. In contrast, we consider treatment response heterogeneity at different points of the treatment distribution. In addition, the RD QTE model of Frandsen, Frölich, 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 defines the causal parameters of interest, and provides our main identification results. Section 3 provides a robust estimand that incorporates the standard RD estimand as a special case. Section 4 proposes convenient tests for our identifying assumptions and discusses including covariates. Section 5 describes estimation. Section 6 provides bias-corrected robust inference and the AMES optimal bandwidths. Section 7 presents the empirical analysis. Short concluding remarks are provided in Section 8. All proofs, inference based on undersmoothing, estimation details of the biases, variances, and AMSE optimal bandwidths, as well as additional empirical analyses are gathered in the Appendix.

2 Identification

We discuss causal identification in RD designs with a continuous treatment, following the 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).

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 the treatment. Let $R \in \mathcal{R} \subset \mathbb{R}$ be the continuous running variable that partly determines the treatment. Consider an outcome equation $Y = G(T, R, \varepsilon)$,

[3] Torgovitsky (2015) 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).

where $\varepsilon \in \mathcal{E} \subset \mathbb{R}^{d_\varepsilon}$ is allowed to be of arbitrary dimension.[4]

Assume that at a known threshold value of the running variable, r_0, treatment changes discontinuously. Define $Z \equiv \mathbf{1}(R \geq r_0)$, where $\mathbf{1}(\cdot)$ is an indicator function equal to 1 if the expression in the parentheses is true and 0 otherwise. The continuous treatment variable can be written as $T = T_1 Z + T_0 (1 - Z)$, where T_z, $z = 0, 1$, is the potential treatment when Z is set at a hypothetical value z. Assume T_z has a reduced-form equation $T_z = q_z(R, U_z)$ with an unobserved disturbance U_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 focuses on $R \in \mathcal{R}$, where $\mathcal{R}$ is an arbitrarily small compact interval around r_0.

Assumption 1 (Quantile representation)**.** *For $z = 0, 1$ and any $r \in \mathcal{R}$, the conditional distribution of T_z given $R = r$ is continuous with a strictly increasing CDF $F_{T_z|R}(T_z, r)$, and $q_z(r, u)$ is strictly monotonic in u.*

Assumption 1 imposes monotonicity on unobserved heterogeneity in the first-stage. Given Assumption 1, one can conveniently take $q_z(r, u)$ as the conditional u quantile of T_z given $R = r$. $U_z = F_{T_z|R}(T_z, R) \sim Unif(0, 1)$ is then the conditional rank of T_z given R.[5]

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

Assumption 2 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

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

[5]Suppose more generally $T_z = H_z(R, V_z)$, $z = 0, 1$, where V_z is a continuous random variable that is not necessarily uniformly distributed and $H_z(R, v)$ is strictly monotonic in v. One can simply normalize $T_z = H_z(R, V_z)$ to be its quantile representation by a strictly monotonic transformation of V_z, i.e., set $U_z = F_{V_z|R}(V_z, R)$.

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 2 in practice requires the 'no manipulation' condition that units cannot sort to be just above or below the RD threshold (McCrary, 2008).

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

Assumption 3.2 assumes local rank similarity, a weaker condition than Assumption 3.1. Without conditioning on ε, U_0 and U_1 given $R = r_0$ both follow a uniform distribution over the unit interval, i.e., $U_0|\,(R = r_0) \sim U_1|\,(R = r_0)$ by construction. Local rank similarity 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 3 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 (along with Assumptions 1 and 2) have readily testable implications. See Dong and Shen (2018) and Frandsen and Lefgren (2018) for discussion on the testable implications of rank invariance or rank similarity in treatment models. In Section 4 we discuss a convenient test for these assumptions in our setting.

The following lemma shows that conditioning on $U \equiv U_1 Z + U_0(1 - Z)$, any changes in the outcome at the RD threshold are causally related to changes in the treatment.

Lemma 1. *Let Assumptions 1-3 hold.*

1. $T \perp \varepsilon|\,(U, R)$.

2. For any integrable function of Y, $\Gamma(Y)$, and any $u \in (0,1)$,

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

Lemma 1.1 states that U is a control variable conditional on R. The defining feature of any 'control variable' is that conditional on this variable, treatment is exogenous to the outcome of interest. Note that here the 'IV' $Z \equiv \mathbf{1}(R \geq r_0)$ is binary and is a deterministic function of a possibly endogenous covariate R. When $U = u$ and $R = r \in \mathcal{R} \backslash r_0$, T is deterministic, i.e., $T = q_1(r,u)$ for $r > r_0$, and $T = q_0(r,u)$ for $r < r_0$; whereas when $U = u$ and $R \to r_0$, T can randomly take on two potential values $q_0(r,u)$ and $q_1(r,u)$. Therefore, causal identification with this control variable U is still local to the RD cutoff, which is a generic feature of the RD design. Lemma 1.1 can follow analogously from Theorem 1 of Imbens and Newey (2009).[6]

Lemma 1.2 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 $q_1(r_0,u)$. Lemma 1.2 gives the average effect of the 'IV' Z on the outcome $\Gamma(Y)$ for individuals at the u quantile of the treatment distribution.

Based on Lemma 1, 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

$$\begin{aligned}\tau(u) &\equiv \mathbb{E}\left[\frac{G(T_1, r_0, e) - G(T_0, r_0, e)}{T_1 - T_0} | U = u, R = r_0\right] \\ &= \int \frac{G\left(q_1(r_0,u), r_0, e\right) - G\left(q_0(r_0,u), r_0, e\right)}{q_1(r_0,u) - q_0(r_0,u)} dF_{\varepsilon|U,R}(e,u,r_0) \\ &= \frac{\mathbb{E}\left[Y | U_1 = u, R = r_0\right] - \mathbb{E}\left[Y | U_0 = u, R = r_0\right]}{q_1(r_0,u) - q_0(r_0,u)}, \end{aligned} \tag{1}$$

where $\frac{G(T_1,r_0,e)-G(T_0,r_0,e)}{T_1-T_0}$ is the (standardized) individual causal effect, so Q-LATE captures an average causal effect for individuals at the u quantile of the treatment at

[6]Note, however, that Imbens and Newey (2009) focus on a continuous IV, which when equipped with a large support assumption that essentially requires the IV varies a lot after possibly conditioning on exogenous covariates, yields more general identification results.

the RD threshold. Q-LATE $\tau(u)$ can be further written as the ratio of the reduced-form effect of Z on Y to that of Z on T at the u quantile of the treatment T.[7]

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

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

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

When 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 T. In particular, following Lemma 5 of Angrist, Graddy, and Imbens (2000), one can write

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

for $\varkappa(u) \equiv \big(q_1(r_0, u) - q_0(r_0, u)\big)^{-1}$ under standard regularity conditions and interchange of the order of integration and differentiation. That is, Q-LATE $\tau(u)$ is a weighted average derivative, averaging over the change in T at a given quantile u at r_0. It follows that WQ-LATE $\pi(w)$ is also a weighted average derivative, averaging over both changes in T at a given quantile u and over $\mathcal{U}$ at the RD threshold.

To identify Q-LATE and WQ-LATE, we impose the following first-stage assumption.

Assumption 4 (First-stage)**.** $q_1(r_0, u) \neq q_0(r_0, u)$ *for at least some* $u \in (0, 1)$.

Assumption 4 requires that the distribution functions $F_{T_1|R}(t, r_0)$ and $F_{T_0|R}(t, r_0)$ are not the same. This is in contrast to the standard RD first-stage assumption requiring a mean change, i.e., $\mathbb{E}[T_1|R = r_0] \neq \mathbb{E}[T_0|R = r_0]$. The weak IV literature considers the case when $\mathbb{E}[T|Z = 1] - \mathbb{E}[T|Z = 0]$ is close to zero. There, inference on the Wald ratio estimator is known to be difficult. We instead seek to provide alternative ways to identify and estimate causal treatment effects. Intuitively, if treatment changes

[7] Consider the special case where Y is given by a linear correlated random coefficients model of the form $Y_i = a_i + b_i T_i$. Q-LATE $\tau(u) = \mathbb{E}[b_i|U = u, R = r_0]$, which captures the average partial effect for units ranked at the u quantile of the treatment distribution at the RD threshold.

concentrate in a few quantiles, instead of looking at the average change, focusing on where the true changes are in the treatment distribution may strengthen identification.[8]

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 2, $q(r,U)$ is right and left continuous in r at $r = r_0$ for all $u \in (0,1)$. 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 1–4, for any $u \in \mathcal{U}$, Q-LATE $\tau(u)$ is identified and is given by*

$$\tau(u) = \frac{m^+(u) - m^-(u)}{q^+(u) - q^-(u)}. \tag{2}$$

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

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. $w(u)$ is required to be non-negative; otherwise, $\pi(w)$ can be a weighted difference of the average treatment effects among those who change treatment levels at the RD threshold. The next section shows that the standard RD estimand can be expressed as a WQ-LATE, using a particular weighting function. In the special case when treatment effect is locally constant, the weighting function does not matter. With any valid weighting function, one can identify the same homogenous treatment effect. Replacing Y by any integrable function $\Gamma(Y)$ in the above, one can readily identify Q-LATE and WQ-LATE on $\Gamma(Y)$.

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 po-

[8] When changes at some treatment quantiles are small, one can modify our trimming threshold for U to be above some constant some constant $c > 0$ instead of 0.

tential outcome corresponding to the treatment value $t \in \mathcal{T}$ be $Y_t \equiv G(t, R, \varepsilon)$. Under Assumptions 1-3, $F_{Y_{t_1}|U,R}(y, u, r_0) = \lim_{r \to r_0^+} \mathbb{E}\left[\mathbf{1}(Y \leq y) | T = q^+(u), R = r\right]$ 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 $Q_{Y_{t_1}|U,R}(v, u, r_0) - Q_{Y_{t_0}|U,R}(v, u, r_0)$ for any $v \in (0, 1)$ and $u \in \mathcal{U}$, where $Q_{Y_{t_z}|U,R}(v, u, r_0) \equiv F^{-1}_{Y_{t_z}|U,R}(v, u, r_0)$.

3 Robust estimand

In this section, we discuss the standard RD estimand and show that it can be expressed as a WQ-LATE, using a particular weighting function. We then seek to provide a 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, and rewrite it as follows

$$\begin{aligned}
\pi^{RD} &\equiv \frac{\lim_{r \to r_0^+} \mathbb{E}[Y|R = r] - \lim_{r \to r_0^-} \mathbb{E}[Y|R = r]}{\lim_{r \to r_0^+} \mathbb{E}[T|R = r] - \lim_{r \to r_0^-} \mathbb{E}[T|R = r]} \\
&= \frac{\int_0^1 \{\mathbb{E}[Y|U_1 = u, R = r_0] - E[Y|U_0 = u, R = r_0]\} du}{\int_0^1 (q_1(r_0, u) - q_0(r_0, u)) du} \qquad (3) \\
&= \int_{\mathcal{U}} \tau(u) \frac{\Delta q(u)}{\int_{\mathcal{U}} \Delta q(u) du} du, \qquad (4)
\end{aligned}$$

where $\Delta q(u) \equiv q_1(r_0, u) - q_0(r_0, u)$, (3) follows from the smoothness conditions in Assumption 2, and (4) follows from Assumption 3 and the definition of Q-LATE $\tau(u)$. Equation (4) shows that under our assumptions, the standard RD estimand identifies a weighted average of Q-LATEs, using as weights $w^{RD}(u) \equiv \Delta q(u) / \int_{\mathcal{U}} \Delta q(u) du$.

The above requires $\Delta q(u) \geq 0$ or $\Delta q(u) \leq 0$ for all $u \in \mathcal{U}$ to ensure $w^{RD}(u) \geq 0$ over $\mathcal{U}$. Otherwise, when $\Delta q(u)$ can switch signs over $\mathcal{U}$, the local Wald ratio π^{RD}

would be undefined if the denominator $\int_{\mathcal{U}} \Delta q\,(u)\,du = 0$. Further if $\int_{\mathcal{U}} \Delta q\,(u)\,du \neq 0$, π^{RD} would be a weighted difference of average treatment effects for units with positive treatment changes and units with negative treatment changes. The following assumption is sufficient for $w^{RD}(u) \geq 0$.

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

Assumption 3b requires that treatment T is weakly increasing or weakly decreasing almost surely in Z. Under Assumption 3b, $\Delta q\,(u) \geq 0$ or $\Delta q\,(u) \leq 0$ for all $u \in (0, 1)$.

Unlike Assumption 3, which imposes rank restrictions, Assumption 3b imposes a sign restriction on the treatment changes at the RD threshold. Angrist, Graddy, and Imbens (2000, Assumption 4) have made a similar assumption in identifying a general simultaneous equations system with binary IVs.

When Assumption 3 local rank invariance or rank similarity does not hold, Q-LATE involved in equation (4) does not have a causal interpretation. However, the RD estimand may still identify a causal parameter under Assumption 3b monotonicity. This is parallel to the standard RD designs with a binary treatment, where the local Wald ratio identifies a LATE for compliers. We formally state this result in the following Lemma 2.

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

The proof of Lemma 2 shows that Assumptions 1, 2, 3b, and 4, the standard RD estimand with a continuous treatment identifies a weighted average of individual treatment effects among those individuals who change their treatment intensity at the RD threshold.[9] When further $G(T, R, \varepsilon)$ is continuously differentiable in its first argument, the identified effect 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 2 in the Appendix.

[9] Card, Lee, Pei, & Weber (2015, Section A.2) shows a simpler expression when the treatment T is a deterministic function of the running variable.

3.2 Robust WQ-LATE estimand

The discussion so far suggests that the standard RD estimand in general requires monotonicity in order to be a causal estimand. Monotonicity does not always hold in practice. Both monotonicity and local rank assumptions impose restrictions on the first-stage heterogeneity, but neither assumption implies the other. Monotonicity imposes a sign restriction on $T_1 - T_0$ at $R = r_0$, while the local rank invariance or rank similarity imposes a rank restriction on the joint distribution of T_1 and T_0 at $R = r_0$. It is then useful to have an estimand that is valid under either assumption.

Theorem 2 below provides a robust WQ-LATE estimand that is valid under either monotonicity or local rank invariance or rank similarity.

Theorem 2 (Robust Estimand)**.** *Let Assumptions 1, 2 and 4 hold. Then under either Assumption 3 or 3b,*

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

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

When monotonicity holds, $\pi^* = \pi^{RD}$; otherwise, when monotonicity does not hold, but our rank assumption holds, $\pi^* = \pi(w^*) \equiv \int_{\mathcal{U}} \tau(u) w^*(u) du$ for $w^*(u) \equiv \frac{|\Delta q(u)|}{\int_{\mathcal{U}} |\Delta q(u)| du}$.[10] Either way, π^* identifies a weighted average treatment effect among those individuals who change their treatment intensity at the RD threshold. The two alternative assumptions specify different ways in which individuals can respond to crossing the RD threshold.[11] Our robust estimand π^* then provides a robust way to aggregate these individual treatment effects.

[10] Theorems 1 and 2 in fact only require Assumptions 3 and 3b to hold for U_z over the sub-support $\mathcal{U} \subseteq (0, 1)$, for $z = 0, 1$.

[11] Under monotonicity $\Pr(T_1 \geq T_0 | R = r_0) = 1$, π^* identifies a weighted average of individual causal effects $\frac{G(q_1(r_0,u_1),r_0,\varepsilon) - G(q_0(r_0,u_0),r_0,\varepsilon)}{q_1(r_0,u_1) - q_0(r_0,u_1)}$ among those having $q_1(u_1) - q_0(u_0) > 0$ (or among those having $q_1(u_1) - q_0(u_0) < 0$ under $\Pr(T_1 \leq T_0 | R = r_0) = 1$). Under the rank restriction, π^* identifies a weighted average of $\frac{G(q_1(r_0,u),r_0,\varepsilon) - G(q_0(r_0,u),r_0,\varepsilon)}{q_1(r_0,u) - q_0(r_0,u)}$ among those having $\varepsilon | (U_1 = u, R = r_0) \sim \varepsilon | (U_0 = u, R = r_0)$ for any $u \in \mathcal{U}$.

4 Discussion

We impose Assumptions 1 - 4 for identification. The quantile and first-stage conditions are assumptions imposed on estimable functions and observables, respectively. They can be directly verified in the data. In the following we briefly discuss the testable implications of the local smoothness and rank restrictions and propose convenient tests for them. It is worth emphasizing that what we propose is a joint test for the implications of both assumptions, similar to Kitagawa (2015) and Arai et al. (2018).

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, $U_0 | (\varepsilon, R = r_0) \sim U_1 | (\varepsilon, R = r_0)$. 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 $u \in (0, 1)$. Further by Assumption 2, $F_{\varepsilon|U_z,R}(e, u, r)$ and hence $F_{X|U_z,R}(x, u, r)$, $z = 0, 1$, is continuous at $r = r_0$. It follows that the right and left limits of $F_{X|U,R}(x, u, r) = Z F_{X|U_1,R}(x, u, r) + (1 - Z) F_{X|U_0,R}(x, u, r)$ exist at $r = r_0$. Therefore, to test the implications of the local smoothness and rank restrictions, one can test the following null hypothesis

$$H_0: \lim_{r \to r_0^+} \mathbb{E}[\mathbf{1}(X \le x) | U = u, R = r] - \lim_{r \to r_0^-} \mathbb{E}[\mathbf{1}(X \le x) | U = u, R = r] = 0, \tag{6}$$

$$\forall x \in \mathcal{X}, u \in \mathcal{U}.$$

The left-hand side of equation (6) corresponds to the numerator of equation (2) with Y being replaced by $\mathbf{1}(X \le x)$. Testing for our identifying assumptions then amounts to testing that the 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.

Complementary to our proposed tests, conventional RD validity tests can be used to provide suggestive evidence for local smoothness. In particular, one can test smoothness of the conditional means of covariates or the density of the running variable. See, e.g., McCrary (2008), Otsu, Xu, and Matsushita (2013), Bugni and Canay (2018), Canay and Kamat (2018), and Cattaneo, Jansson, and Ma (2018). In addition, Arai et

al. (2018) provide a joint test for the identifying assumptions of the standard fuzzy RD design with a binary treatment.

The above leverages covariates for testing purposes. Our discussion of identification does not directly condition on covariates. In practice, rank similarity may be more plausible when conditioning on relevant covariates (see, e.g., Chernozhokov and Hansen, 2005). One may relax the rank restrictions to assume that local rank similarity holds after conditioning on additional covariates other than the running variable. Let Assumptions 1, 2, and 3 hold conditioning on covariates. All our results would then hold conditional on covariates. To obtain the unconditional (on covariates) Q-LATEs, one may average over the covariates distribution at a given treatment quantile.

5 Estimation

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

For simplicity, we use the same kernel function $K(\cdot)$ for all estimation. Let the bandwidths for T and R be h_T and h_R, respectively. 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. Denote as $\widehat{\theta}$ the estimate of any parameter θ. Given a sample of n $i.i.d.$ observations $\{(Y_i, T_i, R_i)\}_{i=1}^n$ from (Y, T, R), we estimate Q-LATE 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 $q^+(u)$ by $\widehat{q}^+(u) \equiv \widehat{a}_0$ from the local linear quantile regression

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

where $\rho_u(\alpha) = \alpha(u - \mathbf{1}(\alpha < 0))$ is the standard check function.[12] $q^-(u)$ can be estimated similarly using observations below r_0.

[12] 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 (2015). Both papers show that the monotonized estimators share the same first-order limiting distribution with the initial local linear estimator.

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 $\epsilon_n^2 \left(\sup_{u\in\mathcal{U}} \left||\Delta\hat{q}(u)| - |\Delta q(u)|\right|\right)^{-1} = o_p(1)$.[13] For all $u \in \widetilde{\mathcal{U}}$, estimate $m^+(u)$ by $\widehat{m}^+(u) \equiv \widehat{b}_0$ from the local linear regression

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

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

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

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

Our identification theory requires trimming out treatment quantiles where there are no changes at the RD threshold, i.e., $\Delta q(u) = 0$, whereas in practice we do not know the true $\Delta q(u)$. To avoid any pre-testing 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 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 $\epsilon_n \equiv \max_{u\in\mathsf{U}^{(l)}} se\left(\Delta\tilde{q}(u)\right) \times 1.96$, where $\Delta\tilde{q}(u)$ is a preliminary Step 1 estimator of the treatment quantile change, using the bandwidth $\tilde{h}\sigma_R$ such that $\tilde{h}/h \to 0$ and $n\tilde{h}^2/h \to \infty$. We discuss the choice of h next in Section 6.[14] 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.

[13] 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.

[14] 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 $a < b < (a+1)/2$. The associated standard errors satisfy $se(\Delta\tilde{q}(u)) = O_p\left(\left(n\tilde{h}\right)^{-1/2}\right) > se(\Delta\hat{q}(u)) = O_p\left(\left(nh\right)^{-1/2}\right)$.

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 $\mathcal{U}$.

The above describes estimation of Q-LATEs or WQ-LATE. 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.

6 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 $\Delta\hat{q}(u)$. We overcome these complications by extending the results of Kong, Linton, and Xia (2010) and Qu and Yoon (2015). Qu and Yoon (2015) provide uniform convergence results for local linear quantile regressions, while Kong, Linton, and Xia (2010) establish strong uniform convergence results for local polynomial estimators.

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

We follow the popular approach of Calonico, Cattaneo, and Titiunik (2014) and develop robust inference for our bias-corrected estimators. The robust inference takes into account the added variability due to bias correction in deriving large sample dis-

tributions. We also present the asymptotically mean squared error (AMSE) optimal bandwidths for both the Q-LATE and WQ-LATE estimators by minimizing the AMSE. Imbens and Kalyanaraman (2012) propose the AMSE optimal bandwidth for the standard RD estimator. The robust confidence intervals for the bias-corrected estimators deliver valid inference when the AMSE optimal bandwidths are used.

We impose the following assumptions for asymptotics.

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

2. For any $t \in \mathcal{T}_z$, $z = 0, 1$, and $r \in \mathcal{R}$, $\mathbb{E}\left[G(T_z, R, \varepsilon)|T_z = t, R = r\right]$ has bounded fourth order derivatives; the conditional variance $\mathbb{V}\left[G(T_z, R, \varepsilon)|T_z = t, R = r\right]$ 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 $\int_{-\infty}^{\infty} \upsilon^2 K(\upsilon) d\upsilon > 0$.

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

We present the preliminary asymptotic distributions of the main estimators $\hat{\tau}(u)$ and $\hat{\pi}^*$ in Appendix B, followed by the inference theory for the undersmoothing approach. The next section presents the inference theory for the bias-corrected approach.

6.1 Bias-corrected robust inference

Denote the leading bias for $\hat{\tau}(u)$ as $h^2\mathsf{B}_\tau(u)$ and the bias 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 (B.1) of Lemma 4 and equation (B.3) of Lemma 5 in Appendix B. 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 following 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. Denote as b the bandwidth used in the bias estimation.

Bias correction reduces biases, but also introduces variability. When the added variability of estimating the bias is not accounted for, the empirical coverage of the resulting confidence interval 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 3 (Asymptotic distribution of $\hat{\tau}^{bc}(u)$)**.** *Let Assumptions 1-5 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* $nh^3\max\{1, h^6/b^6\} \to \infty$, *then for any* $u \in \mathcal{U}$,

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

The exact forms of $\mathsf{V}_\tau(u)$, $\mathsf{V}_{\mathsf{B}_\tau}(u)$ *and* $\mathsf{C}_\tau(u;\rho)$ *are given in equations (B.2), (B.12), and (B.13), respectively, in Appendix B.*

The variance $\mathsf{V}^{bc}_{\tau,n}(u)$ 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 3 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 3 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 $\mathsf{V}_\tau(u)$. For example, for the Uniform kernel and $\rho = 1$, $\mathsf{V}^{bc}_{\tau,n}(u) = 13.89\mathsf{V}_\tau(u)/(nh^2)$.

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

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

The exact forms of V_π , $\mathsf{V}_{\mathsf{B}_\pi}$, *and* C_π *are given in equations (B.4), (B.16), and (B.17), 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 3, Theorem 4 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 $\mathsf{V}^{bc}_{\pi,n} \equiv \mathsf{V}_\pi/(nh)$. 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

[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.

estimator $\widehat{\mathsf{B}}_\tau$ is first-order while the actual estimator $\hat{\pi}^*$ is of smaller order, and hence $\mathsf{V}^{bc}_{\pi,n} \equiv \mathsf{V}_{\mathsf{B}_\pi}/(nb\rho^{-4})$.

Given our results, one can estimate the robust variances by the plug-in estimators. Details for estimating the biases and variances are provided in Appendix C.

As a convenient alternative, one may apply the standard bootstrap based on drawing n observations with replacement to obtain standard errors and confidence intervals. The bootstrap is known to be valid for the local linear mean and quantile estimators in $\Delta\hat{m}(u)$ and $\Delta\widehat{q}(u)$ for any u (see, e.g., Horowitz, 2001). Since $\Delta\hat{m}(u)$ and $\hat{\tau}(u)$ are differentiable functions of the local linear mean and quantile estimators, the bootstrap is valid for $\hat{\tau}(u)$ by the standard delta method. To show the bootstrap validity for $\widehat{\pi}^*$, it suffices to show the uniform stochastic expansion for $\widehat{\pi}^*$ based on the bootstrap sample.[16] This additional technicality is out of the scope of the current paper. Bootstrap validity for the bias-corrected estimators $\hat{\tau}^{bc}(u)$ and $\hat{\pi}^{bc}$ follows the same arguments.

6.2 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 5 and Theorem 6 below. Further details for estimating these AMSE optimal bandwidths are provided in Appendix C.

Theorem 5 (AMSE optimal bandwidth for $\hat{\tau}(u)$)**.** *Let Assumptions 1-5 hold. If $h = h_n \to 0$ and $nh^2 \to \infty$, then the mean squared error of $\hat{\tau}(u)$ is $\mathbb{E}\left[\left(\hat{\tau}(u) - \tau(u)\right)^2\right] = h^4\mathsf{B}_\tau(u)^2 + \left(nh^2\right)^{-1}\mathsf{V}_\tau(u) + o\left(h^4 + \left(nh^2\right)^{-1}\right)$; further if $\mathsf{B}_\tau(u) \neq 0$, the bandwidth that minimizes the AMSE is $h^*_\tau = \left(\mathsf{V}_\tau(u)/\left(2\mathsf{B}^2_\tau(u)\right)\right)^{1/6} n^{-1/6}$.*

[16] 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 3 hold for bootstrap estimators $\Delta\widehat{q}^*(u)$ and $\Delta\hat{m}^*(u)$ for any u. Bootstrap validity then follows from the fact that $\tau(u)$ is a differentiable function of $\Delta m(u)$ and $\Delta q(u)$. $\hat{\pi}^*$ 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 Theorem 3.1 in Fang and Santos (2019), the standard bootstrap is valid for inference on $\sum_{u\in\mathcal{U}^{(l)}}\pi(u)\frac{|\Delta q(u)|}{\sum_{u\in\mathcal{U}^{(l)}}|\Delta q(u)|}$, where $\mathcal{U}^{(l)} \equiv \{u \in \mathsf{U}^{(l)} \ \ |\Delta q(u) > \epsilon_n|\}$ for a fixed number of grid points l. As $l \to \infty$, uniformity over u is sufficient for valid inference on π^*.

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 3. Therefore, one can apply the above AMSE optimal bandwidth and then conduct the bias-corrected robust inference provided in Theorem 3.

Theorem 6 (AMSE optimal bandwidth for $\hat{\pi}^*$)**.** *Let Assumptions 1-5 hold. If* $h = h_n \to 0$ *and* $nh \to \infty$*, then the mean squared error of* $\hat{\pi}^*$ *is* $\mathbb{E}\left[\left(\hat{\pi}^* - \pi^*\right)^2\right] = h^4 \mathsf{B}_\pi^2 + (nh)^{-1} \mathsf{V}_\pi + o\left(h^4 + (nh)^{-1}\right)$*; further if* $\mathsf{B}_\pi \neq 0$*, the bandwidth that minimizes the AMSE is* $h_\pi^* = \left(\mathsf{V}_\pi / \left(4\mathsf{B}_\pi^2\right)\right)^{1/5} n^{-1/5}$.

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 4. These AMSE optimal bandwidths trade off squared biases with variances, so when the biases are small, the AMSE optimal bandwidths can be large.

7 Empirical analysis

The United States banking system in the early 20th century was characterized as a fragile system consisting of thousands of unit banks. Minimum capital requirements were set in place to prevent bank failures; however, bank runs and banking panics were prevalent in the pre-Great Depression era. Are banks less likely to fail when they hold more capital? It is an important question both in this historical context and in light of the current debate on the macroprudencial vs microprudencial approach to financial regulation. The macroprudencial approach promotes higher capital requirements, especially in economic upturns (Hanson, Kashyap, and Stein, 2011). It aims to create incentives for troubled banks to raise new capital rather than shrink assets to restore their damaged capital ratios (the percentage of a bank's capital to its risk-weighted assets), since many institutions shrinking assets simultaneously is likely to be more damaging to the economy.

It is challenging to evaluate the causal impacts of capital holdings on bank failure, as higher capital requirements can be responses to bank runs or banking panics instead of the other way around. The regulation regime in the early 20th century United States provides a unique opportunity for one to nonparametrically identify the true causal impacts of capital holdings on bank responses and outcomes. As shown previously,

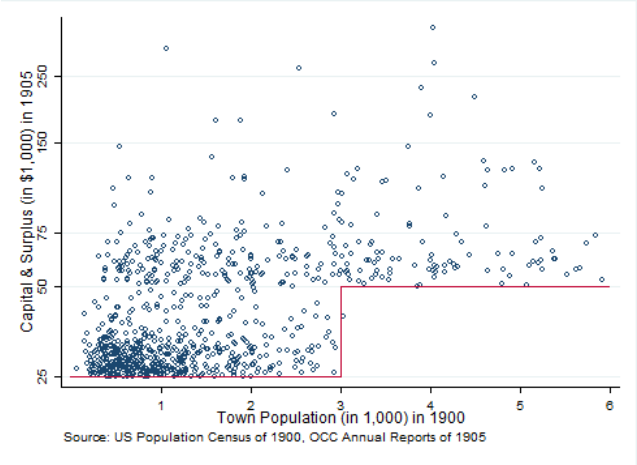

Figure 2: Minimum capital requirements around the town population 3,000 in 1905

the minimum capital requirements were assigned based on the population size of the town a bank operated in. The requirements changed abruptly at various population thresholds.

Figure 2 presents a close-up of banks operating in towns with a population around 3,000, the first regulatory threshold. Over 80% of the towns in our sample have a population near the 3,000 threshold. 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. We therefore focus on the first regulatory threshold and explore the exogenous changes in the capital distribution at this threshold for identification. As is clear from Figure 2, the bottom of the capital distribution shifts up at the 3,000 population threshold.

We estimate the impacts of capital requirements on bank capital (i.e., the first stage impact of Z on T), and further the causal relationships between the induced higher 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. We quantify the possible heterogenous effects (or lack of those) of increased capital at various capital levels.

7.1 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. 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). The minimum capital requirements changed in 1990. Before 1900, the first regulatory threshold did not exist. National banks were required to have a minimum capital of $50,000 regardless of whether they operated in a town above or below the 3,000 population threshold. National banks established before 1900 might be subject to either the old or new regulatory regime, depending on when they were rechartered. We do not have the recharter information, so we focus on national banks that were established after 1900 for clean identification.[17]

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.[18] 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 in 1900, as reported by the 1900 census. Our 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

[17] This is unlike Guo (2016), who analyzed a larger sample of banks established both before and after 1990.

[18] 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

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 leverages. However, these simple correlations may not reflect the true causal relationships. As we can see, towns with more than 3,000 people are associated with older banks, a lower percentage of farm land in their counties, and higher manufacturing output per capita. These results highlight the importance to seek for local identification. Causal relationships would be confounded if comparing banks far away from the threshold.

7.2 Estimation results

Table 2 reports the estimated changes in log capital from 0.10 to 0.90 quantiles and the estimated mean change. Figure 3 visualizes the estimated quantile curves of log capital right above or below the policy threshold (left) and the estimated quantile changes (right) along with their 95% point-wise confidence bands. Consistent with the visual evidence in Figure 2, results in Table 2 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.[19] Given that there is no mean change in log capital at the policy threshold, we cannot apply the standard RD design.

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 7 and 8; Standard errors are in parentheses; *** Significant at the 1% level; ** Significant at the 5% level.

One may be concerned that the data are censored at the minimum capital, implying mass points at $25,000 and $50,000 around the 3,000 population threshold. Assumption 1 would then be invalid. Our data do not suggest a censored distribution for bank capital. Less than 1% of the banks below and less than 2% of the banks above hold the minimum capital in our local estimation sample around the 3,000 threshold. In addition, Figure 3 (left panel) shows that there are no flat regions at the lower ends the estimated quantile curves of log capital right above or right below this threshold.

Table 3 presents the estimated Q-LATEs. These estimates use the AMSE optimal bandwidth given in Theorem 6.[20] We report bootstrapped standard errors that are clustered at the town level, since capital regulation varies at the town level. Alternative

[19] These first-stage estimates are based on a bandwidth implied by the undersmoothing conditions in Theorems 7 and 8 in Appendix B, so we don't have to be concerned with bias correction. Further analysis using other bandwidths suggests that the estimated quantile changes are robust to different bandwidth choices.

[20] 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.

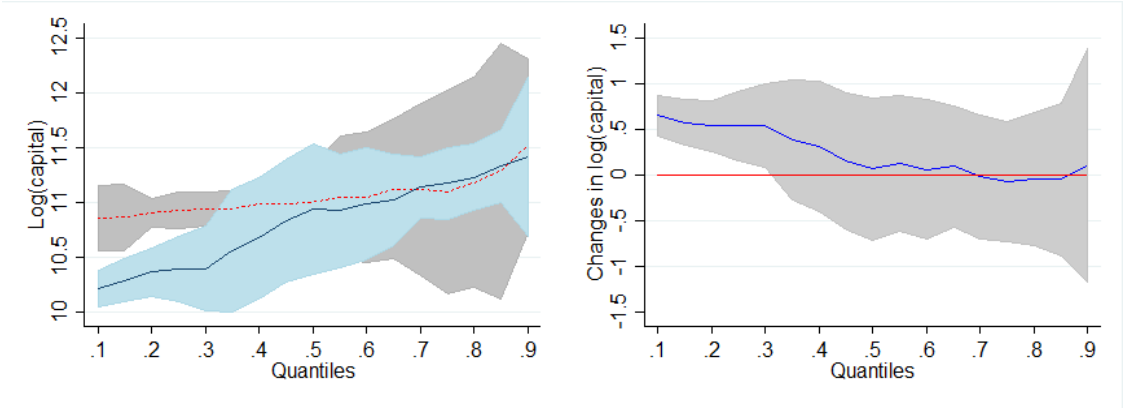

Figure 3: Estimated quantile curves above and below the population threshold 3,000 (left) and the estimated changes at different quantiles (right).

Table 3 Effects of log(capital) on bank outcomes

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 $3/4 h_R = 831.0$; Bootstrapped standard errors are clustered at the town level and are in the parentheses; ***Significant at the 1% level, **Significant at the 5% level.

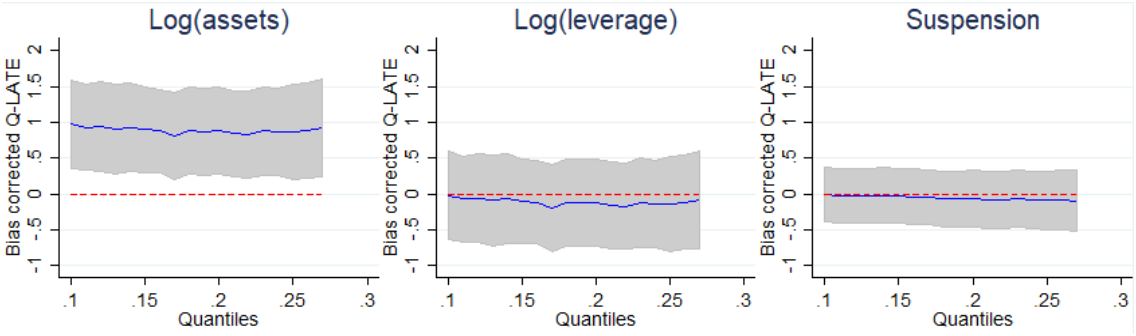

Figure 4: Estimated Q-LATEs at different quantiles

results based on undersmoothing and results with analytical standard errors (without clustering) are presented in Appendix D. For brevity, Table 3 presents the estimated Q-LATEs at selected quantiles. Figure 4 illustrates the estimated Q-LATEs at a finer grid of quantiles along with the 95% confidence intervals.

The estimated Q-LATEs for log assets 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 requirements. 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 small and insignificant.

Figure 5 plots the estimated WQ-LATEs against different bandwidth choices. The estimated WQ-LATEs are robust to a wide range of bandwidths.

7.3 Validity checks

We have estimated the impacts of increased bank capital among banks with low levels of capital. Validity of these estimates requires our local smoothness and rank restrictions to hold. In the following, we evaluate validity of these assumptions.

We first perform the usual standard RD tests to provide suggestive evidence for

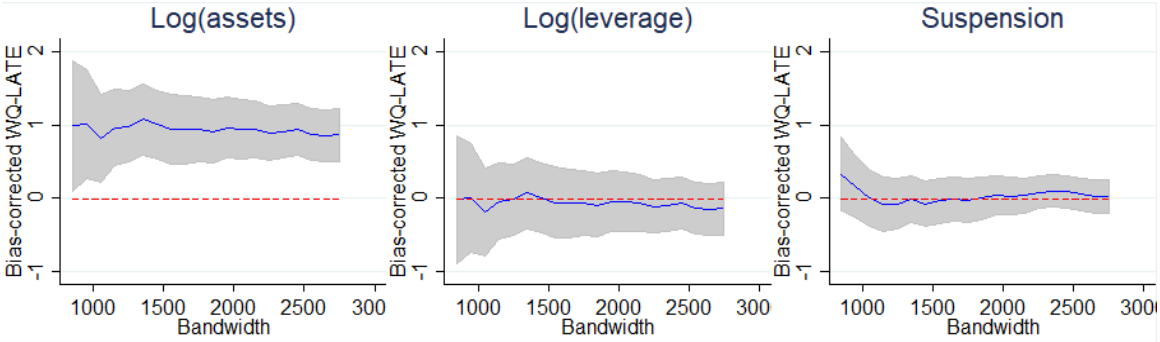

Figure 5: Estimated WQ-LATEs by different bandwidths

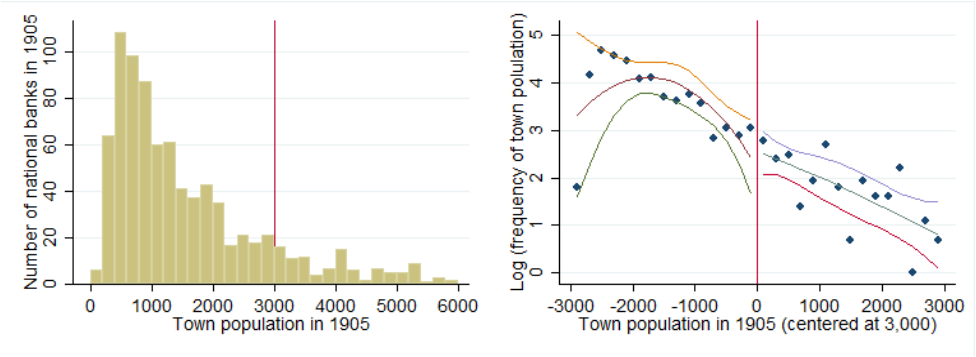

Figure 6: Histogram and the empirical density of town population

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.

Following the standard practice, we test smoothness of the density of town population near the policy 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 6 and 7 provide visual evidence of smoothness. The left graph in Figure 6 presents the histogram of the town population, while the right graph presents 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.

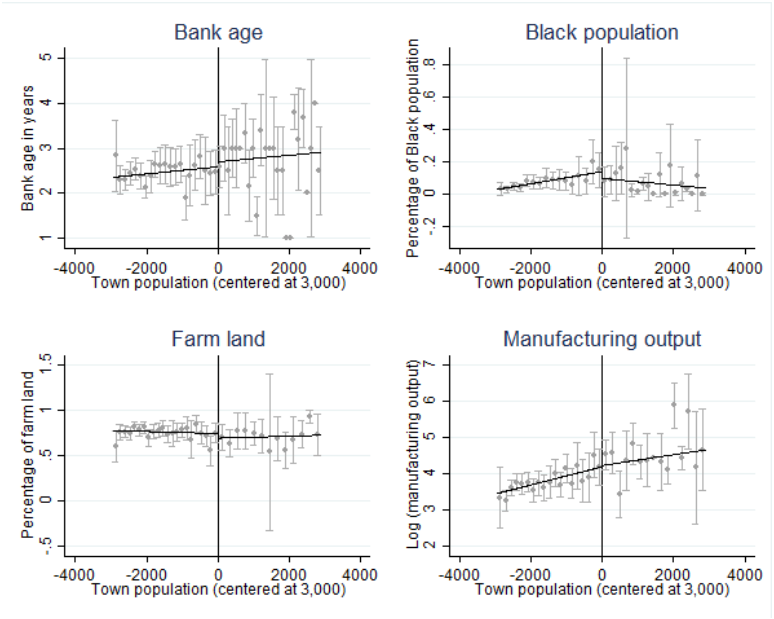

Figure 7: Conditional means of covariates conditional on town population

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.

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. These results suggest that smoothness conditions are plausible in our empirical setting.

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.

We next perform our proposed joint test. For simplicity, instead of testing the entire distribution of covariates, we test the low order (raw) moments of covariates. That is, we replace the outcome variable 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 WQ-LATEs. We use the same bandwidth and specification as those used to produce the main estimates in Table 3. Results of these falsification tests are presented in Table 6. Figures 8 and 9 further visualize the results. None of these estimates are significant. Overall we cannot reject validity of the local smoothness and rank restrictions.

7.4 Policy implications

Our empirical analysis shows that while higher capital requirements induce banks at the bottom of the capital distribution to hold more capital, these banks adjust their assets proportionately. That is, banks simply scale up without a ratio regulation. Their leverages and long-run risk of failure remain almost unchanged. This analysis sheds light on the U.S. banking crisis in the early twentieth century, when bank runs and bank panics occurred often – 29 banking panics occurred from 1865 to 1933.

Note that while earlier regulations focus on the dollar amount of capital, modern regulations focus on capital ratios. Our results support such a regime shift. Interest-

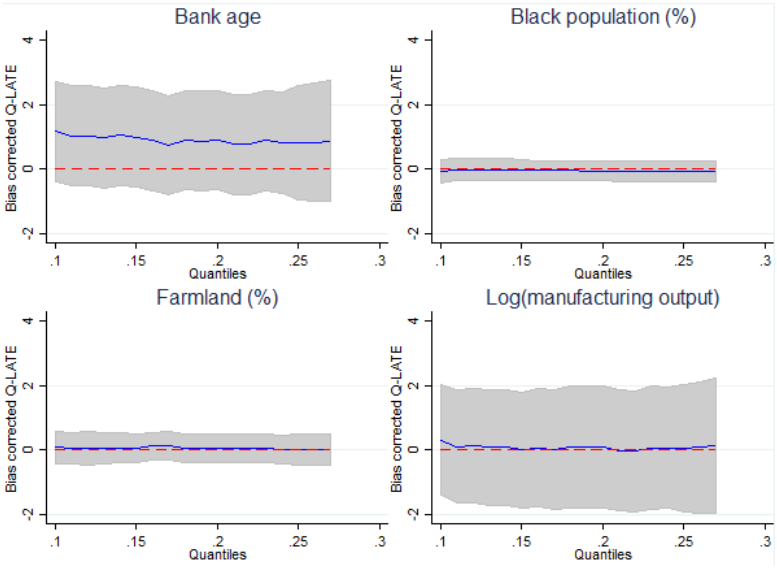

Figure 8: Estimated Q-LATEs on covariates (first moments)

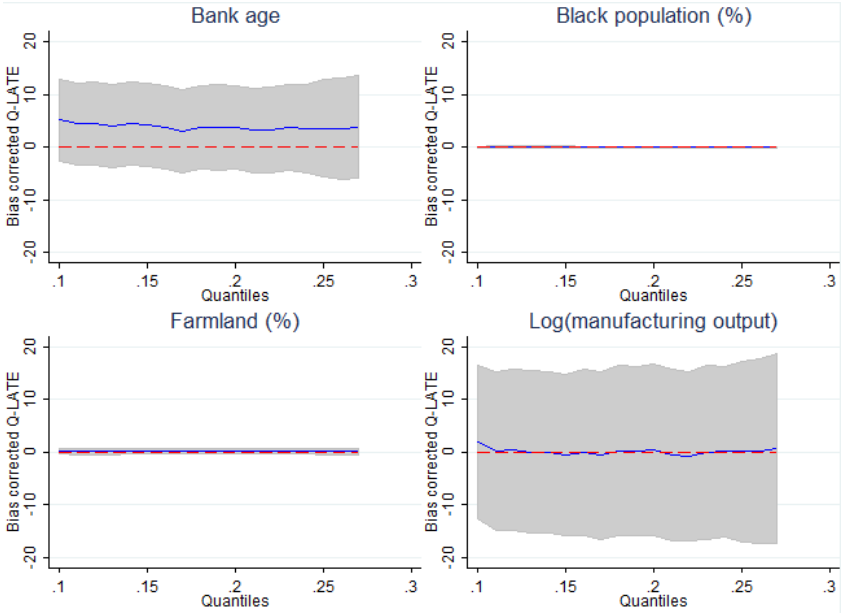

Figure 9: Estimated Q-LATEs on covariates (second moments)

ingly, existing studies suggest that under a ratio regulation, troubled banks in a financial crisis tend to shrink assets rather than raise new capital to restore their damaged capital ratios, even when the latter is more desirable from a social perspective. Based on this and what we have learned from our empirical exercise, a better practice seems to be supplementing the capital ratio regulation with higher capital requirements. This is precisely what the macroprudential approach promotes. For example, in discussing the macroprudential approach, Hanson, Kashyap, and Stein (2011) note "...it may be especially helpful in thinking about the phase-in of higher capital requirements under Basel III."

8 Conclusion

An empirically important class of RD designs involve continuous treatments. This is the first paper to provide identification and inference theory for the RD design with a continuous treatment. We utilize for identification any treatment distributional changes (including the usual mean change as a special case) at the RD threshold.

Our model applies to a large class of policies that target parts or features of the treatment distribution, such as changing the mean, changing the variance or shifting one or both tails of the distribution. Treatment changes are generally responses to relevant policies, and such policies may target some parts (e.g., top or bottom) or features of the treatment distribution. By focusing on where the true changes are in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects.

We identify both quantile specific treatment effects (Q-LATEs) and a weighted average treatment effect (WQ-LATE) at the RD threshold. We also provide bias-corrected robust inference along with the AMSE optimal bandwidths for the identified treatment effects. Our approach complements the standard RD design and the related weak identification approach, since we can identify treatment effect heterogeneity at different treatment intensities. Compared with the standard RD local Wald ratio, the proposed WQ-LATE estimator has the advantage of being robust to possible failure of the monotonicity assumption. It incorporates the standard RD local Wald ratio as a special case; it is valid under either the local rank restriction or the monotonicity assumption.

In our empirical scenario, the minimum capital regulation shifts up the bottom of the capital distribution, but leads to no mean changes. Estimating the causal impacts of capital holdings would be difficult by just applying the standard RD design. However, taking advantage of lower quantile changes in the capital distribution allow for precisely estimating the causal impacts of increased bank capital.

We show that while the higher capital requirements induce small banks to hold more capital, these banks adjust their assets proportionately to lead to only a "scale-up" effect. A 1% increase in capital leads to a close to 1% increase in assets among all banks at the lower quantiles of the capital distribution. As a result, the long-run (up to 24 years, from 1905 to 1929) risk of suspension for those banks stays the same. These results help us better understand the frequent bank runs and banking panics prior to the Great Depression. These results are also useful in considering the macroprudential approach to financial regulation, which promotes higher capital requirements under the ratio regulation regime.

References

[1] Agarwal, S., S. Chomsisengphet, N. Mahoney, and J. Stroebel (2018): "Do Banks Pass through Credit Expansions to Consumers Who want to Borrow?" The Quarterly Journal of Economics, 133(1), 129-190.

[2] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91(434), 444-455.

[3] Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Journal of Econometrics, 113(2), 231-263.

[4] Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," Econometrica, 70(1), 91-117.

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

[6] 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.

[7] Arai, Y., Y. Hsu, T. Kitagawa, I. Mourifie, and Y. Wan (2018): "Testing Identifying Assumptions in Fuzzy Regression Discontinuity Design," Working paper.

[8] Berger, A. and C. Bouwman (2013): "How Does Capital Affect Bank Performance during Financial Crises," Journal of Financial Economics, 109, 146-176.

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

[10] 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.

[11] Brinch, C. N., M. Mogstad, and M. Wiswall, (2017): "Beyond LATE with a Discrete Instrument," Journal of Political Economy, 125(4), 985-1039.

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

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

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

[15] Canay, I. A. and V. Kamat (2018): "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," The Review of Economic Studies, 85(3), 1577-1608.

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

[17] Card, D., R. Chetty, and A. Weber (2007): "The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job?" American Economic Review, 97(2), 113-118.

[18] 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.

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

[20] Clark, D., and H. Royer (2013): "The Effect of Education on Adult Mortality and Health: Evidence from Britain," American Economic Review, 103(6), 2087-2120.

[21] 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.

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

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

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

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

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

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

[28] Corbi, R., E. Papaioannou, and P. Surico (2018): "Regional Transfer Multipliers," The Review of Economic Studies, forthcoming.

[29] Dell, M. and P. Querubin (2018): "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies," The Quarterly Journal of Economics, 133(2), 701-764.

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

[31] Dong, Y. (2019): "Regression Discontinuity Designs with Sample Selection," Journal of Business and Economic Statistics, 37(1), 171-186.

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

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

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

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

[36] 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.

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

[38] Frandsen B. and L. Lefgren (2018): "Testing Rank Similarity," The Review of Economics and Statistics, 100, 86-91.

[39] Frölich, M. and B. Melly (2013): “Unconditional Quantile Treatment Effects under Endogeneity,” Journal of Business & Economic Statistics, 31(3), 346-357.

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

[41] Guo, M. (2016): “Did Capital Requirements Promote Bank Stability in the Early 20th Century United States?" Working paper.

[42] 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.

[43] Hanson, S. G., A. K Kashyap, and J. C. Stein (2011): “A Macroprudential Approach to Financial Regulation," Journal of Economic Perspectives, 25(1), 3-28.

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

[45] 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.

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

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

[48] Imbens, G. and J. Angrist (1994): “Identification and Estimation of Local Average Treatment Effects,” Econometrica, 62(2), 467-475.

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

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

[51] Isen, A., M. Rossin-Slater, and W. R. Walker (2017): "Every Breath You Take?Every Dollar You?ll Make: The Long-Term Consequences of the Clean Air Act of 1970," Journal of Political Economy, 125(3), 848-902.

[52] Kitakawa (2015): "A Test for Instrument Validity," Econometrica, 83(5), 2043-2063.

[53] 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.

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

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

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

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

[58] Oreopoulos, P. (2006): "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter," American Economic Review, 96(1), 152-175.

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

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

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

[62] Pop-Eleches, C. and M. Urquiola (2013): "Going to a Better School: Effects and Behavioral Responses," American Economic Review, 103(4), 1289-1324.

[63] Porter, J. (2003): "Estimation in the Regression Discontinuity Model, " Working paper.

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

[65] Qu, Z. and J. Yoon (2018): "Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs," Journal of Business & Economic Statistics, forthcoming.

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

[67] Schmieder, J.F., T. von Wachter and S. Bender (2012): "The Effects of Extended Unemployment Insurance over the Business Cycle: Evidence from Regression Discontinuity Estimates over 20 Years," Quarterly Journal of Economics, 127(2), 701-752.

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

For Online Publication: Supplemental Appendix for "Regression Discontinuity Designs with a Continuous Treatment"

Yingying Dong, Ying-Ying Lee, Michael Gou

This Appendix is organized as follows. Section A provides proofs for the lemmas, theorem, and corollary presented in Section 2 Identification. Sections B.1 and B.2 provide some preliminary lemmas along with their proofs to facilitate deriving the asymptotic properties for the proposed estimators. Sections B.3 and B.4 then present proofs for the theorems presented in Section 6 Inference. Section C describes how to estimate the biases, variances, and AMSE optimal bandwidths presented in Sections 6.2 and 6.3. Section D provides additional results for the empirical analysis.

A Proofs for Section 2 Identification

Proof of Lemma 1.1 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 $\eta(T)$, $\eta(T)$ is constant conditional on $U = u$ and $R = r$. In particular,

$$\begin{aligned}
&\mathbb{E}[\eta(T) | U = u, R = r] \\
&= \mathbb{E}[\eta(T_1) | U_1 = u, R = r] \mathbf{1}(r \geq r_0) + \mathbb{E}[\eta(T_0) | U_0 = u, R = r] \mathbf{1}(r < r_0) \\
&= \eta(q_1(r, u) \mathbf{1}(r \geq r_0) + q_0(r, u) \mathbf{1}(r < r_0)).
\end{aligned}$$

For any bounded functions $\gamma(\varepsilon)$ and $\eta(T)$,

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

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

By Assumption 3 local rank invariance or 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 $u \in (0, 1)$. Then

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

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

Conditional on $U = u$, T can take on two potential values at $r = r_0$, $T = t_z \equiv q_z(r_0,u)$, $z = 0,1$. The above shows $f_{\varepsilon|T,U,R}(e,t_1,u,r_0) = f_{\varepsilon|T,U,R}(e,t_0,u,r_0) = f_{\varepsilon|U,R}(e,u,r_0)$ for any $u \in (0,1)$ and $e \in Supp(\varepsilon|U = u, R = r_0)$; therefore $T \perp \varepsilon|(U, R = r_0)$.

Proof of Lemma 1.2 For simplicity, the following assumes that Γ is an identity mapping, i.e., $\Gamma(Y) = Y$. The derivation can be readily extended to any integrable functional Γ.

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

where the first equality follows from Assumption 1; the second equality follows from the fact $Y = G(T,R,\varepsilon)$; the third equality follows from the smoothness conditions in Assumption 2, and the last equality follows from the fact that Assumption 3 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, $T = q(r,u) = q_0(r,u)(1-Z) + q_1(r,u)Z$. Further by smoothness of $q_z(r,u)$, $z = 0,1$ in Assumption 2, the right and left limits of $q(r,u)$ $r = r_0$ exist, i.e., $q_1(r_0,u) = \lim_{r\to r_0^+} q(r,u)$ and $q_1(r_0,u) = \lim_{r\to r_0^+} q(r,u)$. Equation (2) holds following Lemma 1. $\pi(\omega) \equiv \int_{\mathcal{U}} \tau(u)\, \omega(u)\, du$ is identified since since $\tau(u)$ is identified, the weighting function $\omega(u)$ is assumed to be known or estimable, and the set $\mathcal{U} \equiv \{u \in (0,1)\colon |q_1(r_0,u) - q_0(r_0,u)| > 0\}$ is identified since $q_z(r,u)$, $z = 0,1$ is identified.

Proof of Lemma 2 For notational convenience, the following derivation uses $q_z(U_z)$ to denote $q_z(r_0,U_z)$, $z = 0,1$. Assumption 3b monotonicity states $\Pr(q_1(U_1) \geq q_0(U_0)) =$

1 or $\Pr\left(q_1(U_1) \leq q_0(U_0)\right)$
$= 1$. Without loss of generality, we assume the former is true. Given the smoothness conditons in Assumption 2, we have

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

where the outer integration in the last equality is over $\mathcal{I}^2 \equiv \left\{(u_0, u_1) \in (0,1) \times (0,1)\text{: } q_1(u_1) - q_0(u_0) > 0\right\}$ and $\widetilde{w}^{RD}\left(u_0, u_1\right) \equiv \frac{q_1(u_1) - q_0(u_0)}{\iint_{\mathcal{I}^2}\left(q_1(u_1) - q_0(u_0)\right) F_{U_0,U_1|R=r_0}\left(du_0, du_1\right)}$. When monotonicity holds, $\widetilde{w}^{RD}\left(u_0, u_1\right) > 0$ and $\iint_{\mathcal{I}^2} \widetilde{w}^{RD}(u_0, u_1) F_{U_0,U_1|R=r_0}(du_0, du_1) = 1$. That is, under Assumptions 2, 3b, and 4, 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 $q_1(u_1) - q_0\left(u_0\right) > 0$ for any $(u_0, u_1) \in \mathcal{I}^2$.

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

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

where $\overline{w}^{RD}(u_0,u_1) \equiv \frac{\Pr(q_0(u_0)\leq t\leq q_1(u_1)|R=r_0)}{\int\int\int \Pr(q_0(u_0)\leq t\leq q_1(u_1))du_0du_1dt}$, the second to the last equality follows from the law of iterated expectations and interchanging the order of integration under standard regularity conditions.

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

Alternatively, when Assumption 3b monotonicity holds, under Assumptions 1, 2 and 4, $\pi^* = \pi^{RD}$. π^{RD} identifies a causal parameter by Lemma 2.

B Proofs for Section 6 Inference

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

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

The following constants are defined by the kernel function. $\kappa_j \equiv \int_0^\infty \nu^j K(\nu)d\nu$, $\lambda_j \equiv \int_0^\infty \nu^j K^2(\nu)d\nu$, $C_{\mathsf{V}} \equiv 4\big(\kappa_2^2\lambda_0 - 2\kappa_1\kappa_2\lambda_1 + \kappa_1^2\lambda_2\big)\big(\kappa_2 - 2\kappa_1^2\big)^{-2}$, $C_{\mathsf{B}} \equiv \big(\kappa_2^2 - \kappa_1\kappa_3\big)\big(\kappa_2 - 2\kappa_1^2\big)^{-1}$, and $C_{\mathsf{C}} \equiv \int_0^\infty K(\nu/\rho)K(\nu)d\nu\big(\rho\kappa_2\int_0^\infty K(\nu/\rho)K(\nu)d\nu - \kappa_1\int_0^\infty \nu K(\nu/\rho)K(\nu)d\nu\big)$.[21]

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

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

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

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

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

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

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

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

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

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 4. *Let Assumptions 1-5 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*

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

$\mathsf{B}_1^\pm(u)$ *and* $\mathsf{B}_2(u)$ *are given in Lemma 3, the influence function* $IF_{\tau i}(u) \equiv \left(Z_i(\phi_{2i}^+(u) + \Phi_{1i}^+(u)(m_t'^+(u) - \tau(u))) - (1 - Z_i)(\phi_{2i}^-(u) + \Phi_{1i}^-(u)(m_t'^-(u) - \tau(u))) \right) (\Delta q(u))^{-1}$, *and* $\Phi_{1i}^\pm(u)$, $\phi_{2i}^\pm(u)$, *and Rem are given in Lemma 3.*

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

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

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

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

$$\mathsf{B}_\pi \equiv \int_\mathcal{U} \mathsf{B}_\tau(u) w^*(u)\, du + \int_\mathcal{U} \left(\mathsf{B}_1^+(u) - \mathsf{B}_1^-(u) \right) \left(\tau(u) - \pi^* \right) \frac{w^*(u)}{\Delta q(u)} du, \tag{B.3}$$

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

$$\Phi_{21i}^\pm \equiv \left(Y_i - m(T_i, r_0^\pm) - m_r'(T_i, r_0^\pm)(R_i - r_0) \right) \frac{w^*\left(F_{T|R}(T_i, r_0^\pm)\right)}{\Delta q\left(F_{T|R}(T_i, r_0^\pm)\right)}$$
$$\times \frac{2\left(\kappa_2 - \kappa_1(R_i - r_0)/(h\sigma_R)\right)}{f_R(r_0)\left(\kappa_2 - 2\kappa_1^2\right)} \frac{1}{h\sigma_R} K\left(\frac{R_i - r_0}{h\sigma_R} \right) \mathbf{1}\left(F_{T|R}(T_i, r_0^\pm) \in \mathcal{U} \right),$$

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

(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*

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

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

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

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

$$V_\pi^q \equiv \frac{C_V}{\sigma_R f_R(r_0)} \int_{\mathcal{U}} \int_{\mathcal{U}} (\min\{u,v\} - vu) \left(\frac{\Lambda^+(u)\Lambda^+(v)}{f^+_{T|R}(u)\, f^+_{T|R}(v)} + \frac{\Lambda^-(u)\Lambda^-(v)}{f^-_{T|R}(u)\, f^-_{T|R}(v)} \right) dv\, du. \tag{B.6}$$

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

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

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

B.2 Asymptotic distributions under undersmoothing

Theorem 7 below presents the asymptotic distribution of $\hat{\tau}(u)$ under 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 7 (Asymptotic distribution of $\hat{\tau}(u)$)**.** *Let Assumptions 1-5 hold. If* $h = h_n \to 0$, $nh^3 \to \infty$, *and* $nh^6 \to 0$, *then for* $u \in \mathcal{U}$

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

The exact form of $V_\tau(u)$ *is given by equation (B.2) of Lemma 4 in Appendix B.*

The bandwidth conditions in Theorem 7 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 7 implies $\sqrt{nh^2}\left(\hat{\tau}(u) - \tau(u)\right) \to_d \mathcal{N}(0, V_\tau(u))$, where $V_\tau(u)$ is the asymptotic variance of $\sqrt{nh^2}\hat{\tau}(u)$. The $100(1-\alpha)\%$ confidence interval for $\tau(u)$ is then given by $\left[\hat{\tau}(u) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{V_\tau(u)/(nh^2)}\right]$, where $\Phi^{-1}_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. One can estimate $V_\tau(u)$ by the usual plug-in estimator, i.e., replacing the unknown parameters involved with their consistent estimates.

Theorem 8 below similarly presents the asymptotic distribution of $\hat{\pi}^*$ using 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 8 (Asymptotic distribution of $\hat{\pi}^*$)**.** *Let Assumptions 1-5 hold. If* $h = h_n \to 0$, $nh^4 \to \infty$, $nh^5 \to 0$, *and the number of grid points* $l \to \infty$, *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 (B.4) of Lemma 5 in Appendix B.*

The bandwidth conditions in Theorem 8 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 8, $\sqrt{nh}\left(\hat{\pi}^* - \pi^*\right) \to_d \mathcal{N}(0, \mathsf{V}_\pi)$, where V_π is the asymptotic variance of $\sqrt{nh}\hat{\pi}^*$.

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

B.3 Proofs for Section B.1

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

Proof of Lemma 3

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

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

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

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

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

(M) Proof for $\Delta\hat{m}(u)$ Kong, Linton, and Xia (2010) provide a uniform Bahadur representation for the local polynomial regression that is uniform over the interior

support of the regressors. In the following, we extend their results to the case when one of the regressors R is evaluated at the boundary point r_0.

Decompose $\hat{m}^+(u)-m^+(u) = \hat{m}^+(u)-\tilde{m}^+(u)+\tilde{m}^+(u)-m^+(u)$, where $\tilde{m}^+(u) = \hat{\mathbb{E}}\big[Y|T = q^+(u),\, R = r_0\big]$ is the infeasible estimator using the true $q^+(u)$. By Corollary 1 of Kong, Linton, and Xia (2010), the following asymptotic linear representation holds:[22] $\tilde{m}^+(u)-m^+(u)-\mathbb{B}\left[\tilde{m}^+(u)\right] = n^{-1}\sum_{i=1}^n Z_i\phi_{2i}^+(u)+O_p\left(\left(\log n/(nh^2)\right)^{3/4}\right)$ uniformly over $u \in \mathcal{U}$, where the bias

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

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

Note $(1,0,0)S_1^{-1} = \big(2\kappa_2, -2\kappa_1, 0\big)/\big(\kappa_2 - 2\kappa_1^2\big)$ and $(1,0,0)S_1^{-1}Q_1 = 2(C_{\mathsf{B}}, 0, \kappa_2)$. Then $\mathsf{B}_2(u) \equiv \mathbb{B}\big[\tilde{m}^+(u)\big]-\mathbb{B}\big[\tilde{m}^-(u)\big] = C_{\mathsf{B}}\sigma_R^2\big(m_r''^+(u)-m_r''^-(u)\big)+\kappa_2\sigma_T^2\big(m_t''^+(u)-m_t''^-(u)\big)$.

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

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

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

Proof of Lemma 4

(I) From the proof of Lemma 3, $\|\Delta\hat{q}-\Delta q\|_\infty = O_p\left((\log n/(nh))^{1/2}+h^2\right)$, $\|\Delta\hat{m}-$

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

$\Delta m\|_\infty = O_p\left(\left(\log n/\left(nh^2\right)\right)^{1/2} + h^2\right)$, and uniformly over $u \in \mathcal{U}$,

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

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

(D) Consider the asymptotic variance $\mathsf{V}_\tau(u)$. 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 $\lim_{r\to r_0^+}\mathbb{E}\left[Y - \left(m^+(u) + m_r'^+(u)(R - r_0) + m_t'^+(u)(T - q^+(u))\right)\middle|T = q^+(u), R = r\right] = 0$, 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

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

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

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

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 mo-

ment. The Lyapounov condition holds, $\left(\sum_{i=1}^n \mathbb{V}\left[IF_{\tau i}(u)\right]\right)^{-3/2} \sum_{i=1}^n \mathbb{E}\left[|IF_{\tau i}(u)|^3\right] = O\left((nh^{-2})^{-3/2}\right) \sum_{i=1}^n \mathbb{E}\left[|IF_{\tau i}(u)|^3\right] = O\left(\left(nh^2\right)^{-1/2}\right) = o(1)$.

Proof of Lemma 5

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

$$\begin{aligned}\tilde{\pi} - \pi^* &= \int_{\mathcal{U}} \hat{\tau}(u)\hat{w}^*(u)du - \int_{\mathcal{U}} \tau(u)w^*(u)\,du \\ &= \int_{\mathcal{U}} \left(\hat{\tau}(u)-\tau(u)\right) w^*(u)\,du + \int_{\mathcal{U}} \tau(u)\left(\tilde{w}^*(u)-w^*(u)\right)du \\ &\quad + \int_{\mathcal{U}} \left(\hat{\tau}(u)-\tau(u)\right)\left(\tilde{w}^*(u)-w^*(u)\right)du,\end{aligned} \tag{B.7}$$

where the last term is $O_p\left(\left((\log n/(nh^2))^{1/2}+h^2\right)\left((\log n/(nh))^{1/2}+h^2\right)\right)$ by Lemma 4.

First consider the estimation error in the estimated weighting function $\tilde{w}^*(u)$ in equation (B.7). 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 (2019), $\hat{A}(u) - A(u) = |\Delta\hat{q}(u)| - |\Delta q(u)| = n^{-1}\sum_{i=1}^n \phi_{1i}(u)\left(\mathbf{1}\left(\Delta q(u) > 0\right) - \mathbf{1}\left(\Delta q(u) < 0\right)\right) + O_p(h^3) + o_p\left((nh)^{-1/2}\right) = O_p\left((nh)^{-1/2} + h^2\right)$, since $\mathbf{1}\left(\Delta q(u) = 0\right) = 0$ for $u \in \mathcal{U}$. It follows that $\hat{B} - B = \int_{\mathcal{U}}\left(\hat{A}(u) - A(u)\right)du + o(1) = n^{-1}\sum_{i=1}^n \int_{\mathcal{U}} \phi_{1i}(u)\left(\mathbf{1}(\Delta q(u) >\right.$

$0) - \mathbf{1}(\Delta q(u) < 0)\big)du + o_p\big((nh)^{-1/2}\big) = O_p\big((nh)^{-1/2} + h^2\big)$. Then

$$
\begin{aligned}
&\int_{\mathcal{U}} \tau(u)\left(\tilde{w}^*(u) - w^*(u)\right)du \\
&= \int_{\mathcal{U}} \frac{\tau(u)}{B}\left(\hat{A}(u) - A(u)\right)du - \frac{\pi^*}{B}\left(\hat{B} - B\right) \\
&\quad + O_p\left(\int_{\mathcal{U}} |\tau(u)|du \left\|\hat{A} - A\right\|_\infty \left|\hat{B} - B\right|\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{U}} \left(\frac{\tau(u)}{B} - \frac{\pi^*}{B}\right)\phi_{1i}(u)\left(\mathbf{1}\left(\Delta q(u) > 0\right) - \mathbf{1}\left(\Delta q(u) < 0\right)\right)du \\
&\quad + \ O_p\left(\log n/(nh) + h^4\right) + o_p\left((nh)^{-1/2}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{U}} \left(\tau(u) - \pi^*\right)\phi_{1i}(u)\frac{w^*(u)}{\Delta q(u)}du + O_p\left(\log n/(nh) + h^4\right) + o_p\left((nh)^{-1/2}\right)
\end{aligned}
\tag{B.8}
$$

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

Next consider the first term in (B.7). Let $\mathsf{m}^+(v) \equiv \lim_{r\to r_0^+}\mathbb{E}\big[Y\big|T = v, R = r\big]$, $\mathsf{m}_r^+(v) \equiv \lim_{r\to r_0^+}\partial\mathbb{E}\big[Y\big|T = v, R = r\big]/\partial r$, and $\mathsf{m}_t^+(v) \equiv \lim_{r\to r_0^+}\partial\mathbb{E}\big[Y\big|T = t, R = r\big]/\partial t|_{T=v}$. By change of variable $v = q^+(u)$, $dv = du\partial q^+(u)/\partial u = du f_R(r_0)/f_{TR}^+(u)$. Then $\phi_{2i}^+(u)$ defined in Lemma 3 becomes

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

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

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

The last equality follows by letting $\mathrm{U}_{zi} \equiv F_{T_z|R}(T_{zi}, r_0) \sim Unif(0, 1)$ for $z \in \{0, 1\}$. Thus $T_{1i} = q^+(\mathrm{U}_{1i})$ and $\mathsf{m}^+(T_{1i}) = m^+(\mathrm{U}_{1i})$. The same argument applies to $\mathcal{U}^-$, where $\Delta q(u) < 0$ for $u \in \mathcal{U}^-$. Then together with the influence function derived in Lemma 4, the first term in equation (B.7) is given by

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

Together with (B.8), we obtain the asymptotic linear representation for $\hat{\pi}^*$.

(D) The asymptotic variance V_π is derived using the influence function in Lemma 5(I),

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

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

$\int_{\mathcal{U}} Z_i \Phi_{1i}^{+}(u)\Lambda^{+}(u)du\big] = O(h)$. Then for V_{π}^{q},

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

For V_{π}^{m}, $\lim_{n\to\infty} h\mathbb{V}\left[\mathbf{1}\left(\mathrm{U}_i \in \mathcal{U}\right) Z_i \Phi_{21i}^{+}\right]$ is

$$
\begin{aligned}
&\lim_{n\to\infty} h\mathbb{E}\Bigg[Z\mathbb{E}\Bigg[\left(Y - \left(m^{+}(\mathrm{U}) + m_r'^{+}(\mathrm{U})\left(R - r_0\right)\right)\right)^2 \Bigg| \mathrm{U}, R\Bigg]\left(\frac{w^{*}(\mathrm{U})}{\Delta q(\mathrm{U})}\right)^2 \mathbf{1}\left(\mathrm{U} \in \mathcal{U}\right) \\
&\quad \times \frac{4\left(\kappa_2 - \kappa_1\frac{R-r_0}{h_R}\right)^2}{f_R^2(r_0)(\kappa_2 - 2\kappa_1^2)^2}\frac{1}{h^2\sigma_R^2}K^2\left(\frac{R-r_0}{h\sigma_R}\right)\Bigg] \\
&= \frac{C_{\mathsf{V}}}{\sigma_R f_R(r_0)}\frac{\lim_{r\to r_0^{+}} \mathbb{E}\left[\mathbb{V}\left[Y\middle|\mathrm{U}, R\right]\mathbf{1}\left(\mathrm{U}_1 \in \mathcal{U}\right)\middle|R = r\right]}{\left(\int_{\mathcal{U}} |\Delta q(u)|du\right)^2}
\end{aligned}
\tag{B.9}
$$

by $(w^{*}(\mathrm{U}_i)/\Delta q(\mathrm{U}_i))^2 = \left(\int_{\mathcal{U}} |\Delta q(u)|du\right)^{-2}$. Note $\lim_{r\to r_0^{+}} \mathbb{V}\big[Y\big|\mathrm{U}_1 = F_{T_1|R}(T_1, r_0), R = r\big] = \lim_{r\to r_0^{+}} \mathbb{V}\big[Y\big|T_1 = q^{+}(\mathrm{U}_1), R = r\big] = \sigma^{2+}(\mathrm{U}_1)$ and $\mathrm{U}_1 \sim Unif(0,1)$. Thus in (B.9), $\lim_{r\to r_0^{+}} \mathbb{E}\big[\mathbb{V}\big[Y\big|\mathrm{U}, R\big]\mathbf{1}\big(\mathrm{U}_1 \in \mathcal{U}\big)\big|R = r\big] = \int_{\mathcal{U}} \sigma^{2+}(u)du$.

To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. By the bandwidth conditions, the Lyapounov condition $\left(\sum_{i=1}^{n} \mathbb{V}\big[IF_{\pi i}\big]\right)^{-3/2} \times \sum_{i=1}^{n} \mathbb{E}\big[|IF_{\pi i}|^3\big] = O\big((nh^{-1})^{-3/2}\big)\sum_{i=1}^{n} \mathbb{E}\big[|IF_{\pi i}|^3\big] = O\big((nh)^{-1/2}\big) = o(1)$ 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 $\lim_{l\to\infty} \mathsf{U}^{(l)} = (0,1)$, $\lim_{l\to\infty} \widetilde{\mathcal{U}} = \widehat{\mathcal{U}} \equiv \{u \in (0,1) \ \ |\Delta\hat{q}(u)| > \epsilon_n\}$ for any n. It follows that $\lim_{l\to\infty} l^{-1}\sum_{u_j\in\widetilde{\mathcal{U}}} |\Delta\hat{q}(u_j)| = \int_{\widehat{\mathcal{U}}} |\Delta\hat{q}(u)|du$ 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 $\mathcal{U}$. By Lemma 6, $\int_{\widehat{\mathcal{U}}} |\Delta\hat{q}(u)|du - \int_{\mathcal{U}} |\Delta\hat{q}(u)|du = \int_0^1 |\Delta\hat{q}(u)|(\hat{\chi}(u) - \chi(u))du = o_p((nh)^{-1/2})$. The smoothness condition in Assumption 5.2 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 6. Therefore as $l \to \infty$ and $n \to \infty$, $\hat{\pi}^{*} - \tilde{\pi} = o_p((nh)^{-1/2})$.

Proof of Lemma 6 Rewrite

$$\begin{aligned}
&\hat{\chi}(u) - \chi(u) \\
&= \mathbf{1}\left(|\Delta\hat{q}(u)| > \epsilon_n, |\Delta q(u)| \leq 0\right) - \mathbf{1}\left(|\Delta\hat{q}(u)| \leq \epsilon_n, |\Delta q(u)| > 0\right) \\
&= \mathbf{1}\left(|\Delta\hat{q}(u)| > \epsilon_n, |\Delta q(u)| \leq 0\right) - \mathbf{1}\left(|\Delta\hat{q}(u)| \leq \epsilon_n < 2\epsilon_n < |\Delta q(u)|\right) \quad \text{(B.10)} \\
&\quad - \mathbf{1}\left(|\Delta\hat{q}(u)| \leq \epsilon_n, 0 < |\Delta q(u)| \leq 2\epsilon_n\right). \quad \text{(B.11)}
\end{aligned}$$

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 (B.10) $\mathbf{1}\big(|\Delta\hat{q}(u)| > \epsilon_n, |\Delta q(u)| \leq 0\big) \leq \mathbf{1}\big(\left||\Delta\hat{q}(u)| - |\Delta q(u)|\right| > \epsilon_n\big) = 0$ with probability approaching one (w.p.a.1) for any $u \in \mathcal{U}$. Thus $\big(\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)| - |\Delta q(u)|\right|\big)^{-1} \times \int_0^1 |\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)| > \epsilon_n, |\Delta q(u)| \leq 0\big)du = 0$ w.p.a.1. It then implies that $\int_0^1 |\Delta\hat{q}(u)|\mathbf{1}\big(|\Delta\hat{q}(u)| > \epsilon_n, |\Delta q(u)| \leq 0\big)du = o_p\big(\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)| - |\Delta q(u)|\right|\big)$. The same argument applies to the second term in (B.10) and implies that $\int_0^1 |\Delta\hat{q}(u)|$ $\mathbf{1}\big(|\Delta\hat{q}(u)| \leq \epsilon_n < 2\epsilon_n < |\Delta q(u)|\big)du = o_p\big(\sup_{u\in\mathcal{U}}\left||\Delta\hat{q}(u)| - |\Delta q(u)|\right|\big)$.

For the term in (B.11), note that $\int_0^1 \mathbf{1}\big(0 < |\Delta q(u)| \leq 2\epsilon_n\big)du = F(2\epsilon_n)$ denotes the CDF of $|\Delta q(U)|$ with $U \sim Unif(0,1)$. By the smoothness Assumption 5.1, we can apply a Taylor series expansion $F(2\epsilon_n) = F'(0)2\epsilon_n + o(\epsilon_n) = O(\epsilon_n)$. Therefore

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

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

B.4 Proofs of Theorem 7, Theorem 3, and Theorem 5 for $\tau(u)$

Proof of Theorem 7 Lemma 4 implies Theorem 7 by letting the bias be of smaller order, i.e., $\sqrt{nh^2}h^2\mathsf{B}_\tau(u) = o(1)$.

Proof of Theorem 3 The following derives the terms $\mathsf{V}_{\mathsf{B}_\tau}(u)$ and $\mathsf{C}_\tau(u;\rho)$ in the asymptotic variance of $\sqrt{nh^2}\hat{\tau}^{bc}(u)$, which are due to bias-correction. They are defined

as follows.

$$\mathsf{V}_{\mathsf{B}_\tau}(u) \equiv \mathsf{V}_\tau(u)\frac{4\lambda_0}{C_{\mathsf{V}}}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}e_1e_1^\top S_2^{-1}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right) \text{ and} \tag{B.12}$$

$$\mathsf{C}_\tau(u;\rho) \equiv -\mathsf{V}_\tau(u)\frac{8\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top S_2^{-1}e_1}{\lambda_0 C_{\mathsf{V}}(\kappa_2-2\kappa_1^2)}C_{\mathsf{C}}. \tag{B.13}$$

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 threshold: $\widehat{\mathsf{B}}_\tau^+ - \mathsf{B}_\tau^+ = \left\{\widehat{\mathsf{B}}_2^+ - \mathsf{B}_2^+ - \mathsf{B}_1^+\left(\hat{\tau}-\tau\right) + \left(\hat{\mathsf{B}}_1^+ - \mathsf{B}_1^+\right)(m_t'^+ - \tau)\right\}/\Delta q + Rem_\tau$. 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 (B.14) below and the convergence rates of the derivatives in $\widehat{\mathsf{B}}_2^+$: $\|\hat{m}_r''^+ - m_r''^+\|_\infty = O_p\left(\left(\log n/(nb^6)\right)^{1/2}+b\right)$, $\|\hat{m}_t''^+ - m_t''^+\|_\infty = O_p\left(\left(\log n/(nb^6)\right)^{1/2}+b\right)$, and $\|\hat{m}_t'^+ - m_t'^+\|_\infty = O_p\left(\left(\log n/(nb^4)\right)^{1/2}+b^2\right)$. Lemma 3 in Qu and Yoon (2018) suggests $\|\hat{q}_r''^+ - q_r''^+\|_\infty = O_p\left(\left(\log n/(nb^5)\right)^{1/2}+b\right)$. Thus it can be shown that the term associated with $\hat{q}_r''^+$ in $\widehat{\mathsf{B}}_1^+$ and the remainder terms Rem_τ are of smaller order.

$$\begin{aligned}
&\widehat{\mathsf{B}}_\tau^+ - \mathsf{B}_\tau^+ \\
&= \frac{1}{\Delta q}\left\{b^{-2}\left(C_{\mathsf{B}}e_4+\kappa_2 e_6\right)^\top \beta_{n2}^{*+} + \mathbb{B}\left[\hat{\mathsf{B}}_2^+\right]\right\} + O_p\left(\frac{1}{b^2}\left(\frac{\log n}{nb^2}\right)^{3/4}\right) \\
&\quad - \frac{\mathsf{B}_1^+}{\Delta q}\left(\hat{\tau}-\tau\right) + \frac{C_{\mathsf{B}}\sigma_R^2}{\Delta q}\left(\hat{q}_r''^+ - q_r''^+\right)\left(m_t'^+ - \tau\right) + Rem_\tau \\
&= O_p\left(\left(\log n/(nb^6)\right)^{1/2} + b + \left(\log n/(nh^2)\right)^{1/2} + h^2\right),
\end{aligned} \tag{B.14}$$

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 $\beta_2(\underline{x}) \equiv \left(m^+, m_r'^+\sigma_R, m_t'^+\sigma_T, m_r''^+\sigma_R^2, \lim_{r\to r_0^+}\frac{\partial^2 m(t,r)}{\partial r\partial t}\big|_{t=q^+(u)}\sigma_R\sigma_T, m_t''^+\sigma_T^2\right)^\top$. β_{n2}^{*-}

is defined as β_{n2}^{*+} by replacing Z_i with $1 - Z_i$ and $+$ with $-$.

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

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

where the influence function

$$\begin{aligned}
IF_{\tau^{bc}i} \equiv \frac{1}{\Delta q}\Bigg\{ & Z_i\left(\phi_{2i}^+ + \Phi_{1i}^+\left(m_t'^+ - \tau\right)\right) - \frac{h^2}{b^2}\left(C_\mathsf{B}e_4 + \kappa_2 e_6\right)^\top \beta_{n2}^{*+} \\
& - (1 - Z_i)\left(\phi_{2i}^- + \Phi_{1i}^-\left(m_t'^- - \tau\right)\right) + \frac{h^2}{b^2}\left(C_\mathsf{B}e_4 + \kappa_2 e_6\right)^\top \beta_{n2}^{*-}\Bigg\}. \qquad \text{(B.15)}
\end{aligned}$$

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

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

where the second moment term

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

Therefore

$$
\mathbb{V}\left[\beta_{n2}^{*+}\right]=\frac{2\lambda_0^2\sigma^{2+}}{nb^2\sigma_T\sigma_R f_{TR}^+}W_2S_2^{-1}e_1e_1^\top S_2^{-1}W_2+O\left((nb)^{-1}\right).
$$

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 $(C_{\mathsf{B}}e_4+\kappa_2e_6)^\top W_2=2\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top$, we obtain $\mathsf{V}_{\mathsf{B}_\tau}(u)$ defined in (B.12) by showing that the sample above the threshold contributes

$$
\frac{\sigma^{2+}}{f_{TR}^+}\frac{1}{\sigma_T\sigma_R(\Delta q)^2}8\lambda_0^2\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^\top S_2^{-1}e_1e_1^\top S_2^{-1}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right).
$$

For the covariance term,

$$
\begin{aligned}
&\mathbb{C}\left[Z_i\phi_{2i}^+,\beta_{n2}^{*+}\right] \\
=&\frac{1}{n}\frac{2W_2S_2^{-1}B_n^{-1}}{f_{TR}^{+2}\left(\kappa_2-2\kappa_1^2\right)}\mathbb{E}\Big[K_h\left(T-q^+,R-r_0\right)K_b\left(T-q^+,R-r_0\right) \\
&\times\left(Y-\left(m^+ +m_r'^+\left(R-r_0\right)+m_t'^+\left(T-q^+\right)\right)\right)\left(Y-\mu\left(\underline{X}-\underline{x}\right)^\top W_2^{-1}\beta_2(\underline{x})\right) \\
&\times\left(\kappa_2-\kappa_1(R-r_0)/h\right)\mu\left(\underline{X}-\underline{x}\right)Z\Big],
\end{aligned}
$$

where the expectation term is

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

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

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

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 (B.13)

by showing that the sample above the threshold contributes

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

Therefore $\mathsf{V}_{\tau}^{bc}=O\big((nh^2)^{-1}+h^4(nb^6)^{-1}\big)$ and $\mathbb{B}\big[\hat{\tau}^{bc}\big]=-h^2\mathbb{B}\big[\hat{\mathsf{B}}_{\tau}\big]+O(h^3)=O\big(h^3+h^2b\big)$ is of smaller order by the conditions $n\min\{h^6,b^6\}\max\{h^2,b^2\}\to 0$. We have the asymptotic linear representation in (B.15), $\hat{\tau}^{bc}-\tau=n^{-1}\sum_{i=1}^{n}IF_{\tau^{bc}i}+o_p\big((nh^2)^{-1/2}+h^2(nb^6)^{-1/2}\big)$. To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. When $h/b\to\rho\in 0,\infty)$, (B.15) implies $\sqrt{nh^2}\big(\hat{\tau}^{bc}-\tau-\mathbb{B}\big[\hat{\tau}^{bc}\big]\big)=\sqrt{nh^2}n^{-1}\sum_{i=1}^{n}IF_{\tau^{bc}i}+o_p(1)$. The Lyapounov condition holds, $\big(\sum_{i=1}^{n}\mathbb{V}\big[IF_{\tau^{bc}i}\big]\big)^{-3/2}\sum_{i=1}^{n}\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big((nh^{-2})^{-3/2}\big)\sum_{i=1}^{n}\mathbb{E}\big[|IF_{\tau^{bc}i}|^3\big]=O\big(n^{-1/2}h^3\big(h^{-4}+\rho^6b^{-4}\big)\big)=O\big(\big(nh^2\big)^{-1/2}\big)=o(1)$. Then $\sqrt{nh^2}\big(\hat{\tau}^{bc}(u;h,b)-\tau(u)\big)\to_d\mathcal{N}\big(0,\mathsf{V}_{\tau}^{bc}(u)\big)$.

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

Proof of Theorem 5 Theorem 5 follows by minimizing the AMSE implied by Lemma 4. The asymptotic distribution becomes $n^{1/3}\big(\hat{\tau}(u)-\tau(u)\big)\to_d\mathcal{N}\big(c_u^2\mathsf{B}_{\tau}(u),c_u^{-2}\mathsf{V}_{\tau}(u)\big)$, where $c_u\equiv\big(\mathsf{V}_{\tau}(u)/\big(2\mathsf{B}_{\tau}^2(u)\big)\big)^{1/6}$.

B.5 Proofs of Theorem 8, Theorem 4, and Theorem 6 for π^*

Proof of Theorem 8 Lemma 5 implies Theorem 8 by letting the bias be of smaller order, i.e., $\sqrt{nh}h^2\mathsf{B}_{\pi}=o(1)$.

Proof of Theorem 4 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.

$$\mathsf{V}_{\mathsf{B}_{\pi}}\equiv\mathsf{V}_{\pi}^{m}C_{\mathsf{V}}^{-1}4\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}S_2^{-1}\Lambda_2S_2^{-1}\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)\text{ and}\tag{B.16}$$

$$\mathsf{C}_{\pi}\equiv-\mathsf{V}_{\pi}^{m}\frac{8\left(C_{\mathsf{B}}e_4+\kappa_2e_6\right)^{\top}S_2^{-1}}{C_{\mathsf{V}}\left(\kappa_2-2\kappa_1^2\right)}\int_0^{\infty}K(\upsilon)K(\upsilon/\rho)\mathbf{v}_2(\kappa_2-\kappa_1\upsilon/\rho)d\upsilon,\tag{B.17}$$

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

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

First derive the 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

$$\begin{aligned} IF_{\pi^{bc}i} \equiv Z_i\Bigg\{ &\int_{\mathcal{U}} \Phi_{1i}^+(u)\Lambda^+(u)du - \rho^2\left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^\top \Phi_{22i}^+(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right) \\ &+ \Phi_{21i}^+(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}1}\right)\Bigg\} - (1 - Z_i)\Bigg\{\Phi_{21i}^-(h)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right) \\ &+ \int_{\mathcal{U}} \Phi_{1i}^-(u)\Lambda^-(u)du - \rho^2\left(C_{\mathsf{B}}e_4 + \kappa_2 e_6\right)^\top \Phi_{22i}^-(b)\mathbf{1}\left(T_i \in \mathcal{T}_{\mathcal{U}0}\right)\Bigg\} \end{aligned} \tag{B.18}$$

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

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

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

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

The leading term in Rem_π is $O_p(\|\hat{\mathsf{B}}_\tau - \mathsf{B}_\tau\|_\infty \|\Delta\hat{q} - \Delta q\|_\infty) = O_p\big(\big(\big(\log n/(nb^6)\big)^{1/2} + b + \big(\log n/(nh)\big)^{1/2} + h^2\big)\big(\big(\log n/(nh)\big)^{1/2} + h^2\big)\big)$. 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 4 and Lemma 5,

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

By the same argument in the proof of Lemma 5, the third term associated with $IF_{\tau i}(u)$ is $O_p\big(h^2\big((nh)^{-1/2} + h^2\big)\big)$, 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 (B.14). One can show that

$$
\int_{\mathcal{U}} \frac{w^*(u)}{\Delta q(u)} \left\{ b^{-2} \left(C_{\mathsf{B}} e_4 + \kappa_2 e_6\right)^\top \beta_{n2}^{*+}(u) + \mathbb{B}\big[\hat{\mathsf{B}}_2^+\big] - \mathsf{B}_1^+(u) \left(\hat{\tau}(u) - \tau(u)\right) \right\} du \tag{B.19}
$$

$$
= O_p\left((nb^5)^{-1/2} + b + (nh)^{-1/2} + h^2\right).
$$

To see why, the second term associated with $\mathbb{B}\big[\hat{\mathsf{B}}_2^+\big]$ 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 5 with the additional weight $\mathsf{B}_1^+(\mathrm{U}_i)/\Delta q(\mathrm{U}_i)$. For the first term in (B.19), we use the same arguments as those in deriving (B.9) in the proof of Lemma 8. By change of variable $v = q^+(u)$

and $s = (\upsilon - T_i)/b_T$, we have

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

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

$$
\begin{aligned}
&\mathbb{E}\left[\Phi_{22i}^{+2}(b) \mathbf{1} \left(\mathrm{U}_i \in \mathcal{U} \right) Z_i \right] \\
&= W_2 S_2^{-1} \mathbb{E}\Bigg[\left(Y - \left(m^{+}(\mathrm{U}) + m_r'^{+}(\mathrm{U}) (R - r_0) + \frac{1}{2} m_r''^{+}(\mathrm{U}) (R - r_0)^2 \right) \right)^2 \\
&\quad \times \left(1, \frac{R - r_0}{b}, 0, \left(\frac{R - r_0}{b} \right)^2, 0, 0 \right)^\top \left(1, \frac{R - r_0}{b}, 0, \left(\frac{R - r_0}{b} \right)^2, 0, 0 \right) \\
&\quad \times \left(\frac{w^*(\mathrm{U})}{\Delta q(\mathrm{U})} \right)^2 K_b^2 (R - r_0) \mathbf{1} \left(\mathrm{U} \in \mathcal{U} \right) Z \Bigg] \frac{S_2^{-1} W_2}{f_R^2(r_0)} \\
&= W_2 S_2^{-1} \int_0^{\infty} \int_{\mathcal{T}} \mathbf{1} \left(F_{T_1|R}(T|r_0) \in \mathcal{U} \right) \mathbf{v}^\top \mathbf{v} K^2(\upsilon) \, \mathbb{E}\Bigg[\Bigg(Y - \Bigg(m^{+}(\mathrm{U}) + m_r'^{+}(\mathrm{U})(b\upsilon) \\
&\quad + \frac{1}{2} m_r''^{+}(\mathrm{U})(b\upsilon)^2 \Bigg) \Bigg)^2 \Big| \mathrm{U} = F_{T_1|R}(T|r_0), \, R = r_0 + b\upsilon \Bigg] f_{TR}(T, r_0 + b\upsilon) dT d\upsilon \\
&\quad \times \frac{S_2^{-1} W_2}{b \sigma_R B^2 f_R^2(r_0)} \\
&= \frac{\mathbb{E}\left[\mathbb{V}\left[Y \middle| \mathrm{U}, R \right] \mathbf{1} \left(\mathrm{U} \in \mathcal{U} \right) \middle| R = r_0^{+} \right]}{b \sigma_R B^2 f_R(r_0)} W_2 S_2^{-1} \Lambda_2 S_2^{-1} W_2 + o(b^{-1}) = O(b^{-1}).
\end{aligned}
$$

Thus the first term in (B.19) is $O_p((nb^5)^{-1/2})$. Then $\rho^2 \left(C_{\mathsf{B}} e_4 + \kappa_2 e_6 \right)^\top \Phi_{22i}^{+}(b)$ contributes to the asymptotic variance of $\hat{\pi}^{bc}$ by a term of order $\rho^4 (nb)^{-1} = (nh\rho^{-5})^{-1}$. We obtain $\mathsf{V}_{\mathsf{B}_\pi}$ defined in (B.16) by showing that the sample above the cutoff con-

tributes

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

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

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

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

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

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 (B.17) by showing that the sample above the cutoff contributes

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

Therefore $\mathbb{V}\big[\hat{\pi}^{bc}\big]=O\big((nh)^{-1}+(nb^5h^{-4})^{-1}\big)$ and $\mathbb{B}\big[\hat{\pi}^{bc}\big]=O\big(h^2(h+b)\big)$ that is smaller-order by the bandwidth conditions $n\min\{h^5,b^5\}\max\{h^2,b^2\}\to 0$. To show asymptotic normality, we apply Lyapounov CLT with third absolute moment. When $h/b\to\rho\in(0,\infty)$, (B.18) implies $\sqrt{nh}\big(\hat{\pi}^{bc}-\pi^*-\mathbb{B}\big[\hat{\pi}^{bc}\big]\big)=\sqrt{nh}n^{-1}\sum_{i=1}^n IF_{\pi^{bc}i}+o_p(1)$. The Lyapounov condition $\big(\sum_{i=1}^n\mathbb{V}\big[IF_{\pi^{bc}i}\big]\big)^{-3/2}\times\sum_{i=1}^n\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big]=O\big((nh^{-1})^{-3/2}\big)\sum_{i=1}^n\mathbb{E}\big[|IF_{\pi^{bc}i}|^3\big]=O\big(n^{-1/2}h^{3/2}h^{-2}\big)=O\big((nh)^{-1/2}\big)=o(1)$ holds. Then $\sqrt{nh}\big(\hat{\pi}^{bc}(h,b)-\pi^*\big)\to_d\mathcal{N}\big(0,\mathsf{V}_\pi^{bc}\big)$.

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

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

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

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

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^*. We focus on estimating the unknown parameters defined above the RD cutoff. Corresponding parameters defined below the cutoff can be estimated analogously.

C.1 Biases estimation

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

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 $2\widehat{\alpha}_2$ from the following local quadratic quantile regression with a chosen bandwidth b,

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

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

from the following local quadratic regression

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

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

For the standard error of $\widehat{\tau}(u)$, Theorem 7 gives $\mathsf{V}_\tau(u) \equiv \frac{2\lambda_0 C_{\mathsf{V}}}{(\Delta q(u))^2 f_R(r_0)\sigma_R\sigma_T}\left(\frac{\sigma^{2+}(u)}{f_{T|R}^+(u)} + \frac{\sigma^{2-}(u)}{f_{T|R}^-(u)}\right)$.[23] 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 $g = 0.7344 n^{\pm -1/6}$ for $\hat{f}_{T|R}^{\pm}(u)$ and $g = 1.843 n^{-1/5}$ for $\hat{f}_R(r_0)$.

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

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

[23] The influence function for $\widehat{\tau}(u)$ is provided in Lemma 4. 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 5 to estimate the variance of $\hat{\pi}^*$.

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

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 3, $\mathsf{V}^{bc}_{\tau,n}(u) \equiv \frac{\mathsf{V}_\tau(u)}{nh^2} + \frac{\mathsf{V}_{\mathsf{B}_\tau}(u)+\rho^{-5}\mathsf{C}_\tau(u;\rho)}{nh^2\rho^{-6}}$. 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 (B.12), and $\mathsf{C}_\tau(u;\rho) = 3.125\rho^3\mathsf{V}_\tau(u)$ when $\rho \leq 1$, and $\mathsf{C}_\tau(u;\rho) = 37.5(\rho/3 - 1/4)\mathsf{V}_\tau(u)$ when $\rho > 1$ by equation (B.13). Plugging in $\widehat{\mathsf{V}}_\tau(u)$ for a chosen ρ, one can obtain $\widehat{\mathsf{V}}^{bc}_{\tau,n}(u)$.

Consider the standard error of $\hat{\pi}^*$. By Theorem 8, $\mathsf{V}_\pi \equiv \mathsf{V}^m_\pi + \mathsf{V}^q_\pi$, where $\mathsf{V}^m_\pi \equiv \frac{C_{\mathsf{V}}\int_{\mathcal{U}}(\sigma^{2+}(u)+\sigma^{2-}(u))du}{\sigma_R f_R(r_0)(\int_{\mathcal{U}}|\Delta q(u)|du)^2}$ and $\mathsf{V}^q_\pi \equiv \frac{C_{\mathsf{V}}}{\sigma_R f_R(r_0)}\int_{\mathcal{U}}\int_{\mathcal{U}}\left(\min\{u,\upsilon\} - \upsilon u\right)\left(\frac{\Lambda^+(u)\Lambda^+(\upsilon)}{f^+_{T|R}(u)\,f^+_{T|R}(\upsilon)} + \frac{\Lambda^-(u)\Lambda^-(\upsilon)}{f^-_{T|R}(u)\,f^-_{T|R}(\upsilon)}\right)d\upsilon du$ with $\Lambda^\pm(u) \equiv \left(m'^{\pm}_t(u) - \pi^*\right)\frac{w^*(u)}{\Delta q(u)}$. Estimation of $\Delta q(u)$, $f_R(r_0)$, $f^\pm_{T|R}(u)$, and $\sigma^{2\pm}(u)$ is described at the beginning of this section. $w^*(u)$ is estimated in Step 4 estimation in the main text.

The only unknown involved in V_π is $m'^{\pm}_t(\mathrm{u})$, which appears in $\Lambda^\pm(\mathrm{u})$, $\mathrm{u} = u, \upsilon$. $m'^{\pm}_t(u)$ 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 $\Delta q(u)$, $f_R(r_0)$, $f^\pm_{T|R}(u)$, $\Lambda^\pm(u)$, $\sigma^{2\pm}(u)$, 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 4, $\mathsf{V}^{bc}_{\pi,n} \equiv \frac{\mathsf{V}_\pi}{nh} + \frac{\mathsf{V}_{\mathsf{B}_\pi}+\rho^{-3}\mathsf{C}_\pi}{nh\rho^{-5}}$. Estimation of V_π is provided above. For the Uniform kernel, $\mathsf{V}_{\mathsf{B}_\pi} = 1.641\mathsf{V}^m_\pi$ by equation (B.16) 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 (B.17). 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 bandwidths 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 estimate the AMSE optimal bandwidths by $\widehat{h}^*_\tau(u) = \left(\widehat{\mathsf{V}}_\tau(u)/\left(2\widehat{\mathsf{B}}^2_\tau(u)\right)\right)^{1/6} n^{-1/6}$ 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 D.1 presents the estimated impacts of log capital on the outcomes of interest, using bandwidths consistent with undersmoothing and so no bias correction is made. These estimates show similar patterns as those bias-corrected robust inference results reported in the main text. 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 precision compared with estimates by undersmoothing.

Figure D.1 plots the estimated Q-LATEs based on undersmoothing at different quantiles. Figure D.2 shows how the estimated WQ-LATEs based on undersmoothing changes with bandwidths. Tables D.2 and D.3 present estimates using analytical standard errors.

Table D.1 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 7 and 8; 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.

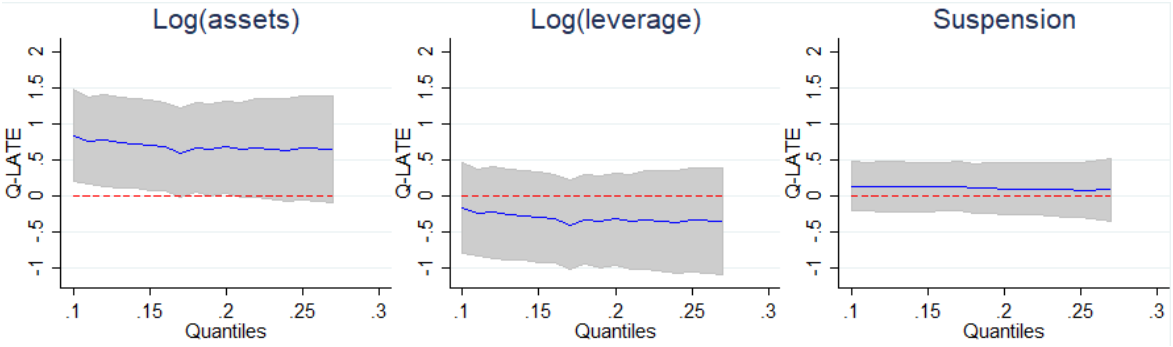

Figure D.1: Estimated Q-LATEs at different quantiles (undersmoothing)

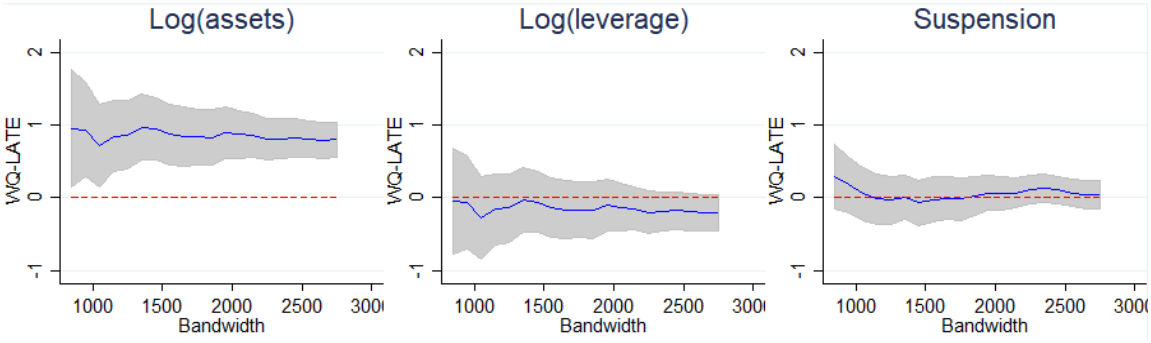

Figure D.2: Estimated WQ-LATEs by different bandwidths (undersmoothing)

Table D.2 Impacts of log (capital) on bank outcomes (analytical SEs)

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.

Table D.3 Impacts of log(capital) on bank outcomes (undersmoothing with analytical SEs)

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 $h_R = 4\sigma_R n^{-0.23} = 1039.5$ 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 7 and 8; 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.