





Regression Discontinuity Designs With a Continuous Treatment

Yingying Dong, Ying-Ying Lee & Michael Gou


To cite this article: Yingying Dong, Ying-Ying Lee & Michael Gou (2021): Regression Discontinuity Designs With a Continuous Treatment, Journal of the American Statistical Association, DOI: [10.1080/01621459.2021.1923509](https://doi.org/10.1080/01621459.2021.1923509)

To link to this article: <https://doi.org/10.1080/01621459.2021.1923509>

 View supplementary material [↗](#)

 Published online: 21 Jun 2021.

 Submit your article to this journal [↗](#)

 Article views: 732

 View related articles [↗](#)

 View Crossmark data [↗](#)



Regression Discontinuity Designs With a Continuous Treatment

Yingying Dong, Ying-Ying Lee, and Michael Gou

Department of Economics, University of California, Irvine, CA

ABSTRACT

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

KEYWORDS

Continuous treatment;
Treatment quantiles; Rank
invariance; Rank similarity;
Double-robust identification

1. Introduction

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

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

Hahn, Todd, and van der Klaauw (2001) showed that under proper conditions, the RD local Wald ratio with a binary treatment identifies a local average treatment effect (LATE) for

compliers at the RD threshold. In RD designs with a continuous treatment, empirical researchers typically apply this RD local Wald ratio to estimate the causal effect of the continuous treatment. Causal identification and inference rely solely on the mean shift of the continuous treatment variable. A few issues arise with this practice. The first issue is interpretation—we show in Section 3 that the LATE interpretation no longer holds with a continuous treatment. Intuitively, there are an infinite number of potential outcomes, and compliers are not immediately defined.

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

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

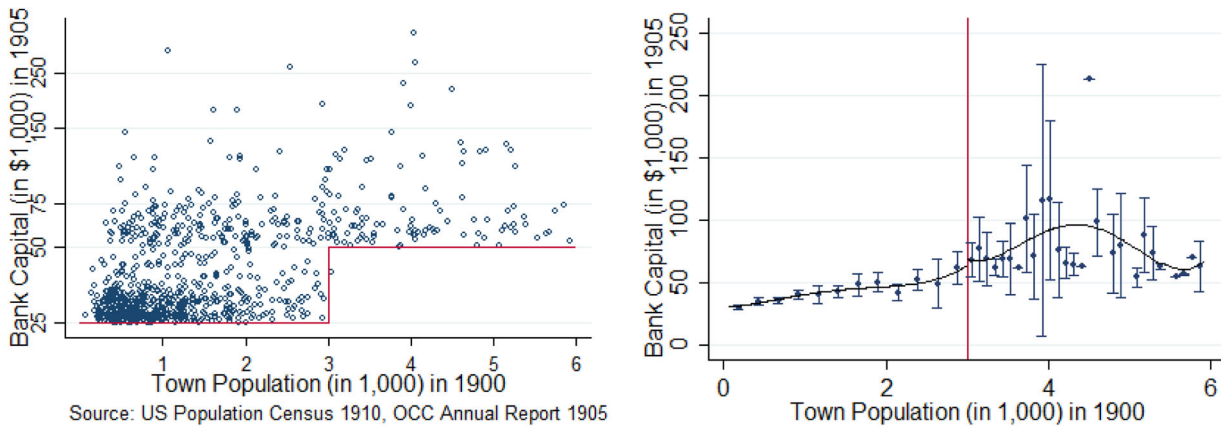


Figure 1. Scatterplot (left) and the RD mean plot (right) of bank capital against town population.

ceiling set by the environmental protection agency. Focusing on the mean treatment may miss the true sources of identification, that is, where the true changes are in the treatment distribution.

In this article, we show that causal identification can be achieved by using any changes in the distribution of the continuous treatment variable at the RD threshold. These include not only the usual mean change, but also changes at various points of the treatment distribution. By focusing on where the true exogenous changes are in the treatment distribution, we provide what are likely to be the most policy relevant treatment effects.

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

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

Our article adds to the growing literature of RD designs, which focuses on binary treatments. See Imbens and Lemieux (2008) for an early review of the RD literature. For more recent reviews, see Cattaneo, Idrobo, and Titiunik (2020, 2021) and Cattaneo, Titiunik, and Vazquez-Bare (2020). Note that our model is different than the RD quantile treatment effect (RD QTE) model discussed by Frandsen, Frölich, and Melly (2012). The RD QTE model still requires a binary treatment along with a continuous outcome. In contrast, our model requires a continuous treatment with either a discrete or continuous outcome. RD QTE captures treatment effect heterogeneity at different points of the outcome distribution, while our Q-LATE parameter captures treatment effect heterogeneity at different points of the treatment distribution. Caetano, Caetano, and Escanciano (2020) discussed identification and estimation of RD designs with a multivalued treatment variable. A continuous treatment has been considered in the literature of regression kink (RK) designs (Card et al. 2015). In RK designs, identification relies on treatment assignment as a kinked function of the running variable.

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

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

2. Identification

In this section, we discuss nonparametric identification of RD designs with a continuous treatment. To fix the idea, ignore the running variable for now. Consider a continuous treatment T and a binary “IV” Z . For an observation i , let $T_i = a_i + b_i Z_i$, where a_i and b_i are random coefficients. Typically, one would estimate a constant coefficient regression in the first stage of the linear IV model, where the constant coefficient of the binary Z captures the exogenous change in the mean treatment. Here, we show that under proper conditions, the random coefficient b_i captures exogenous changes in the distribution of the treatment, which can be used for identification.

2.1. Basic Setup

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

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

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

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

Assumption 1 (Quantile representation). $q_z(r, u)$, $z = 0, 1$, is strictly monotonic in u for any $r \in \mathcal{R}$, where \mathcal{R} is an arbitrarily small closed interval around r_0 . The conditional distribution of T_z given $R = r$ is continuous with a strictly increasing CDF $F_{T_z|R}(t, r)$.

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

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

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

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

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

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

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

Assumption 3.2 assumes rank similarity, a weaker condition than Assumption 3.1. Without conditioning on ε , U_0 and U_1 given $R = r_0$ both follow a uniform distribution over the unit interval, that is, $U_0 | (R = r_0) \sim U_1 | (R = r_0)$ by construction. Local rank similarity permits random “slippages” from the common rank level in the treatment distribution just above or just below the RD cutoff. Rank similarity has been proposed to identify quantile treatment effects (QTEs) in IV models (Chernozhukov and Hansen 2005). Unlike the IV QTE model, we impose the similarity assumption on the ranks of potential treatments, instead of the ranks of potential outcomes. In our empirical analysis, Assumption 3.2 requires that the probability for a bank to stay at the certain rank of the capital distribution stays the same regardless of whether it is in a town with a population just above or just below 3000.

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

The above identifying assumptions have potentially testable implications. Under either Assumption 3.1 or 3.2,

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

2.2. Identification Results

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

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

1. $\lim_{r \rightarrow r_0^-} f_{\varepsilon|TR}(e, q_0(r, u), r) = \lim_{r \rightarrow r_0^+} f_{\varepsilon|TR}(e, q_1(r, u), r) = \lim_{r \rightarrow r_0} f_{\varepsilon|UR}(e, u, r)$ for $e \in \mathcal{E}$.
2. $\lim_{r \rightarrow r_0^+} \mathbb{E}[Y|U = u, R = r] - \lim_{r \rightarrow r_0^-} \mathbb{E}[Y|U = u, R = r] = \int (G(t_1(u), r_0, e) - G(t_0(u), r_0, e)) F_{\varepsilon|UR}(de, u, r_0)$.

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

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

Here, the ‘‘IV’’ $Z \equiv \mathbf{1}(R \geq r_0)$ is binary and is a deterministic function of a possibly endogenous covariate R . Given $U = u$ and $R = r$, T is deterministic, that is, $T = q_1(r, u)$ for $r \geq r_0$, and $T = q_0(r, u)$ for $r < r_0$. Causal identification with this control variable U is therefore local to the RD cutoff, which is a generic feature of the RD design. In contrast, Imbens and Newey (2009) focused on a continuous IV and aim to identify different causal objects than ours.

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

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

It follows that $\lim_{r \rightarrow r_0^+} \mathbb{E}[Y|U = u, R = r] - \lim_{r \rightarrow r_0^-} \mathbb{E}[Y|U = u, R = r] = \mathbb{E}[Y_{t_1(u)} - Y_{t_0(u)} | U = u, R = r_0]$. Based

on Lemma 1.2, we can define the causal parameters of interest, Q-LATE and WQ-LATE.

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

$$\tau(u) \equiv \int \frac{G(t_1(u), r_0, e) - G(t_0(u), r_0, e)}{t_1(u) - t_0(u)} F_{\varepsilon|UR}(de, u, r_0) \quad (1)$$

$$= \mathbb{E} \left[\frac{Y_{t_1(u)} - Y_{t_0(u)}}{t_1(u) - t_0(u)} \middle| U = u, R = r_0 \right], \quad (2)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

To aggregate Q-LATE, one simple weighting function is equal weighting, that is, $w(u) = 1 / \int_{\mathcal{U}} 1 du$. One may choose other properly defined weighting functions. $w(u)$ is required to be nonnegative; otherwise, when $w(u)$ is allowed to be negative, some weights will be greater than 1 and $\pi(w)$ will be some weighted difference of the average treatment effects among those who change treatment levels at the RD threshold. The next section shows that the standard RD estimand can be expressed as a WQ-LATE, using a particular weighting function. In the special case where the treatment effect is locally constant, the weighting function does not matter. With any valid weighting functions, one can identify the same homogeneous treatment effect.

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

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

3. Double-Robust Identification

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

3.1. Standard RD Estimand

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

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

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

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

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

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

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

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

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

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

3.2. Double-Robust Identification

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

Theorem 2 (Double-robust identification). Let [Assumptions 1, 2,](#) and [4](#) hold. Then under either [Assumption 3](#) or [3b](#),

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

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

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

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

4. Estimation

The proposed estimands for Q-LATE and WQ-LATE involve conditional means and quantiles at a boundary point. Following

the standard practice of the RD literature, we estimate Q-LATE and WQ-LATE by local linear mean and quantile regressions.

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

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

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

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

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

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

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

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

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

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

In practice, one can choose $\epsilon_n = \max_{u \in \mathcal{U}^{(l)}} \text{se}(\Delta \tilde{q}(u)) \times 1.96$, where $\Delta \tilde{q}(u)$ is a preliminary Step 1 estimator of the treatment quantile change, using the bandwidth \tilde{h}_R such that $\tilde{h}_R/h_R \rightarrow 0$ and $n\tilde{h}_R^2/h_R \rightarrow \infty$. We discuss the choice of h_R in [Section 5](#). The associated standard errors satisfy $\text{se}(\Delta \tilde{q}(u)) = O_p((n\tilde{h}_R)^{-1/2}) > \text{se}(\Delta \hat{q}(u)) = O_p((nh_R)^{-1/2})$. By this procedure, insignificant estimates (at the 5% significance level) of $\Delta \hat{q}(u)$ along with some significant but small estimates

will be trimmed out. Since $\sup_{u \in \mathcal{U}} \left| |\Delta \hat{q}(u)| - |\Delta q(u)| \right| = O_p((nh_R)^{-1/2})$, the conditions for ϵ_n given in Lemma 6 are satisfied. Consider specifically the bandwidth sequences $h_R = cn^{-a}$ and $\tilde{h}_R = cn^{-b}$ for some constants $0 < a, b < 1$ and $c > 0$. The required conditions for ϵ_n are satisfied when choosing b such that $a < b < (a + 1)/2$.

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

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

5. Inference

The proposed estimators have several distinct features, which make analyzing their asymptotic properties challenging. First, the local polynomial estimator in Step 2 involves a continuous treatment variable T , in addition to the running variable R . Evaluating T over its interior support and evaluating R at the boundary point r_0 complicates the analysis. Second, we need to account for the sampling variation of $\hat{q}^\pm(u)$ from Step 1, which appears in both the numerator and denominator of $\hat{\tau}(u)$, as well as in the weighting function $\hat{w}^*(u)$ for $\hat{\pi}^*$. Third, our estimation involves a trimming procedure that is based on the estimated quantile change $\Delta \hat{q}(u)$. We overcome these complications by extending the results of Kong, Linton, and Xia (2010) and Qu and Yoon (2015). Qu and Yoon (2015) provided uniform convergence results for local linear quantile regressions, while Kong, Linton, and Xia (2010) established uniform convergence results for local polynomial estimators.

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

We develop robust inference for our bias-corrected estimators, similar to the robust bias-corrected inference of Calonico, Cattaneo, and Titiunik (2014) in the context of the standard

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

We impose the following assumptions for asymptotics.

- Assumption 5 (Asymptotics).**
1. For any $t \in \mathcal{T}_z, z = 0, 1, r \in \mathcal{R}$, and $u \in \mathcal{U}$, $f_{TzR}(t, r)$ 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 $r \in \mathcal{R}$, $\mathbb{E}[G(T_z, R, \varepsilon) | T_z = t, R = r]$ has bounded fourth order derivatives; the conditional variance $\mathbb{V}[G(T_z, R, \varepsilon) | T_z = t, R = r]$ is continuous and bounded away from zero; the conditional density $f_{TzR|Y}(t, r, y)$ is bounded for any $y \in \mathcal{Y}$. $\mathbb{E}[|Y - \mathbb{E}[Y | T_z, R]|^3] < \infty$ for $z = 0, 1$.
 3. The kernel function K is bounded, positive, compactly supported, symmetric, having finite first-order derivative, and satisfying $\int_{-\infty}^{\infty} v^2 K(v) dv > 0$.

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

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

5.1. Inference on Q-LATE

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

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

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

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

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. We therefore present the asymptotic distributions of the bias-corrected estimators $\hat{\tau}^{bc}(u)$, taking into account the sampling variation induced by bias correction.

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

$$\frac{\hat{\tau}^{bc}(u) - \tau(u)}{\sqrt{V_{\tau,n}^{bc}(u)}} \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{where}$$

$$V_{\tau,n}^{bc}(u) \equiv \left(\frac{V_{\tau}(u)}{nh^2} + \frac{V_{B_{\tau}}(u)}{nb^6 h^{-4}} + \frac{C_{\tau}(u; \rho)}{nhb} \right) \frac{1}{c_R c_T}.$$

The exact forms of $V_{\tau}(u)$, $V_{B_{\tau}}(u)$ and $C_{\tau}(u; \rho)$ are given in Equations (B.1), (B.2), and (B.3) in Appendix B, respectively.

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

Without loss of generality, we assume that the bandwidths $h_R = c_R h$ and $h_T = c_T h$ are of the same order. We show in Lemma 4 in Appendix B that h_R and h_T have the same first-order impact on $\hat{\tau}(u)$. This is because the local linear estimator of $\mathbb{E}[Y|T, R]$ in Step 2 dominates the first-order asymptotically linear representation, and the quantile regression of T on R in Step 1 is of smaller order. In addition, we derive the optimal bandwidths that minimize the AMSE of $\hat{\tau}(u)$ in **Theorem 4**. The resulting AMSE optimal bandwidths are of the same order $n^{-1/6}$.

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

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

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

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

5.2. Inference on WQ-LATE

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

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

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

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

$$\frac{\hat{\pi}^{bc} - \pi^*}{\sqrt{V_{\pi,n}^{bc}}} \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{where}$$

$$V_{\pi,n}^{bc} \equiv \left(\frac{V_{\pi}}{nh} + \frac{V_{B_{\pi}}}{nb^5 h^{-4}} + \frac{C_{\pi}(\rho)}{nb^2 h^{-1}} \right) \frac{1}{c_R}. \quad (6)$$

The exact forms of V_{π} , $V_{B_{\pi}}$, and $C_{\pi}(\rho)$ are given in Equations (B.4), (B.6), and (B.7) in Appendix B, respectively.

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

$V_{\pi,n}^{bc}$ consists of three terms: V_{π} is from the variance of the actual estimator $\hat{\pi}^*$, $V_{B_{\pi}}$ is from the variance of the bias estimator $h_R^2 \widehat{B}_{R\pi} + h_T^2 \widehat{B}_{T\pi}$, and $C_{\pi}(\rho)$ is from the covariance between $\hat{\pi}^*$ and $h_R^2 \widehat{B}_{R\pi} + h_T^2 \widehat{B}_{T\pi}$. Similar to **Theorem 3**, **Theorem 5**

incorporates three limiting cases depending on ρ . When $h/b \rightarrow \rho = 0$, $\hat{\pi}^*$ is first-order and the bias estimator is of smaller order. Then $V_{\pi,n}^{bc} \equiv V_{\pi}/(nhc_R)$. When $h/b \rightarrow \rho \in (0, \infty)$, both $\hat{\pi}^*$ and the bias estimator contribute to the asymptotic variance. When $h/b \rightarrow \infty$, the bias estimator is first-order and $\hat{\pi}^*$ is of smaller order. Then $V_{\pi,n}^{bc} \equiv V_{B_{\pi}}/(nb^5h^{-4}c_R)$.

Note that Q-LATE $\tau(u)$ is a function of T and R , while WQ-LATE π^* is a weighted average of $\tau(u)$ averaging over T and hence is only a function of R . The asymptotic theory for $\hat{\pi}^*$ in Lemma 5 of Appendix B shows that the leading variance is of order $1/\sqrt{nh_R}$. In theory, one can choose a small bandwidth for T , in particular $h_T = c_{Tn}h$ for $c_{Tn} \rightarrow 0$, such that the leading bias associated with h_T , $h_T^2B_{T\pi}$, becomes first order ignorable compared with the leading bias associated with h_R , $h_R^2B_{R\pi}$. The leading bias of $\hat{\pi}^*$ can then be simplified to $h_R^2B_{R\pi}$. It follows that the first-order asymptotic property of $\hat{\pi}^*$ will not depend on h_T . These are features of the general marginal integration or partial mean of the nonparametrically estimated conditional mean function (see, e.g., Newey 1994). Nevertheless, $h_T^2B_{T\pi}$ might not be ignorable in finite samples. The finite-sample performance of the bias-corrected estimator could be compromised, if the bias term associated with h_T was ignored. Our bias-corrected estimator $\hat{\pi}^{bc}$ and the associated robust inference therefore take into account $h_T^2\widehat{B}_{T\pi}$.

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

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

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

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

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

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

6. Empirical Analysis

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

As shown in Figure 1, the requirement depends on town sizes and changes abruptly at the town population threshold 3000. The required minimum capital is \$25,000 for a bank located in a town with a population less than 3000, and jumps to \$50,000 for a bank located in a town with a population at or above 3000. There are two other population thresholds, 6000 and 50,000, at which the minimum capital requirement changes. Our empirical analysis focuses on the population threshold 3000, since about 88% of banks in our sample are located in towns with a population below 6000.

Let the continuous treatment T be bank capital, and the running variable R be town population. Furthermore, let Z indicate whether a bank is located in a town with 3000 or more people. We consider three outcomes of interest (Y): total assets, leverage, and an indicator of whether a bank suspended its operation in the following 24 years. Leverage is defined as the ratio of a bank's total assets to capital, which is a measure of the amount of risk a bank engages in. Logged values are used for bank capital, assets and leverage, as these variables have rather skewed distributions. We estimate the impacts of the minimum capital requirement on the distribution of bank capital (i.e., the first stage impact of Z on T), and further the impacts of higher capital on the three outcomes of interest (i.e., the impacts of T on Y). We also quantify any possible treatment effect heterogeneity at various levels of bank capital.

Our data come from three sources: the annual reports of the office of the comptroller of the currency (OCC), Rand

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

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

6.1. Estimation Results

Figure 2 visualizes the estimated quantile curves of log capital above or below the policy threshold (left) and the estimated quantile changes (right) along with their 95% point-wise confidence bands. These estimates are generated using $\hat{h}_{R\pi}^* = 1462.76$. For simplicity, all estimates in the empirical analysis use uniform kernels, unless otherwise stated. Consistent with the visual evidence in Figure 1, Figure 2 suggests that significant

changes only occur at roughly the bottom 30 percentiles of the distribution of log capital. The estimated changes are also larger at lower quantiles. In contrast, the estimated mean change in log capital using $\hat{h}_{R\pi}^* = 1462.76$ is 0.107 with a standard error of 0.148. The estimated mean change by the default CCT rdrobust package (using $\hat{h}_R = 803.58$ and a triangular kernel) is 0.141 with a standard error of 0.171. The lack of a significant mean change in bank capital suggests that the standard fuzzy RD design does not apply.

Figure 3 illustrates the bias-corrected estimates of Q-LATEs at different quantiles along with their 95% confidence intervals. The main bandwidths used for estimation are $\hat{h}_{R\pi}^* = 1426.76$ and $\hat{h}_{T\pi}^{rot} = 0.441$. The bandwidth for bias estimation is set to be $4.5n^{-1/8} = 2308.67$, corresponding to $\rho = 0.618$ (See Appendix C.3 for details). A preliminary bandwidth $3/4\hat{h}_{R\pi}^* = 1097.07$ is used to determine the trimming thresholds. Alternative results based on undersmoothing or bootstrapped standard errors (with or without being clustered at the town level) are presented in Appendix D.2. Clustering seems to have little impacts based on the bootstrapped standard errors. Our analytical standard errors therefore do not take into account possible clustering at the town level.

As shown in Figure 3, the estimated Q-LATEs for log assets are around 1 at various low quantiles of log capital. All estimates are significant at the 1% level. The corresponding WQ-LATE is estimated to be 1.034, which is also significant at the 1% level, so on average, a 1% increase in capital leads to roughly a 1%

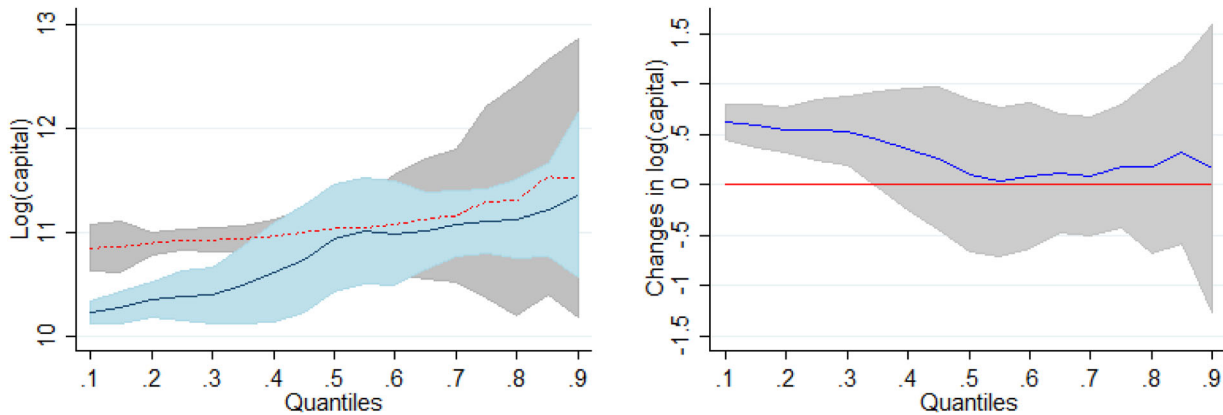


Figure 2. Estimated quantile curves of bank capital above and below the population threshold 3,000 (left) and quantile changes (right).

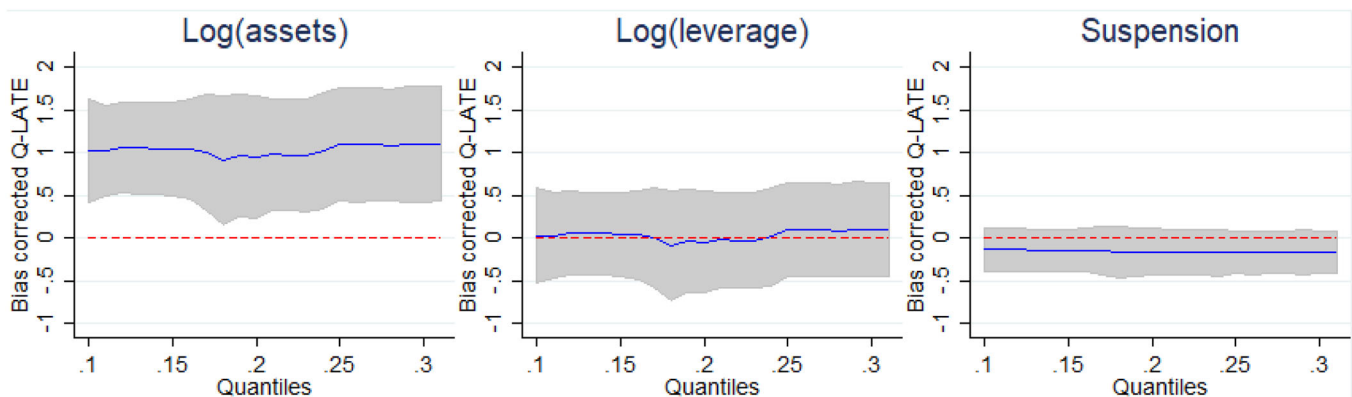


Figure 3. Bias-corrected estimates of Q-LATEs at different quantiles.

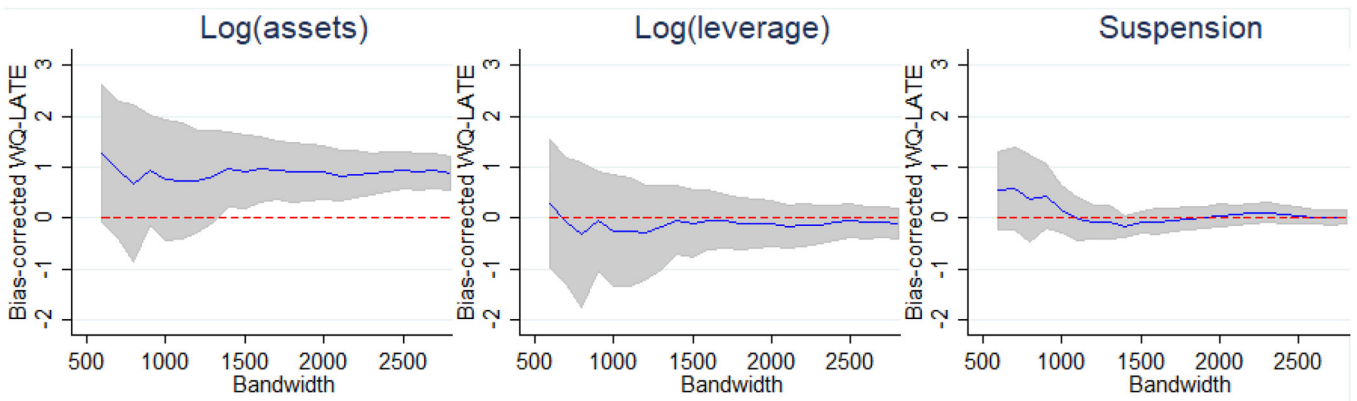


Figure 4. Bias-corrected estimates of WQ-LATEs by different bandwidths.

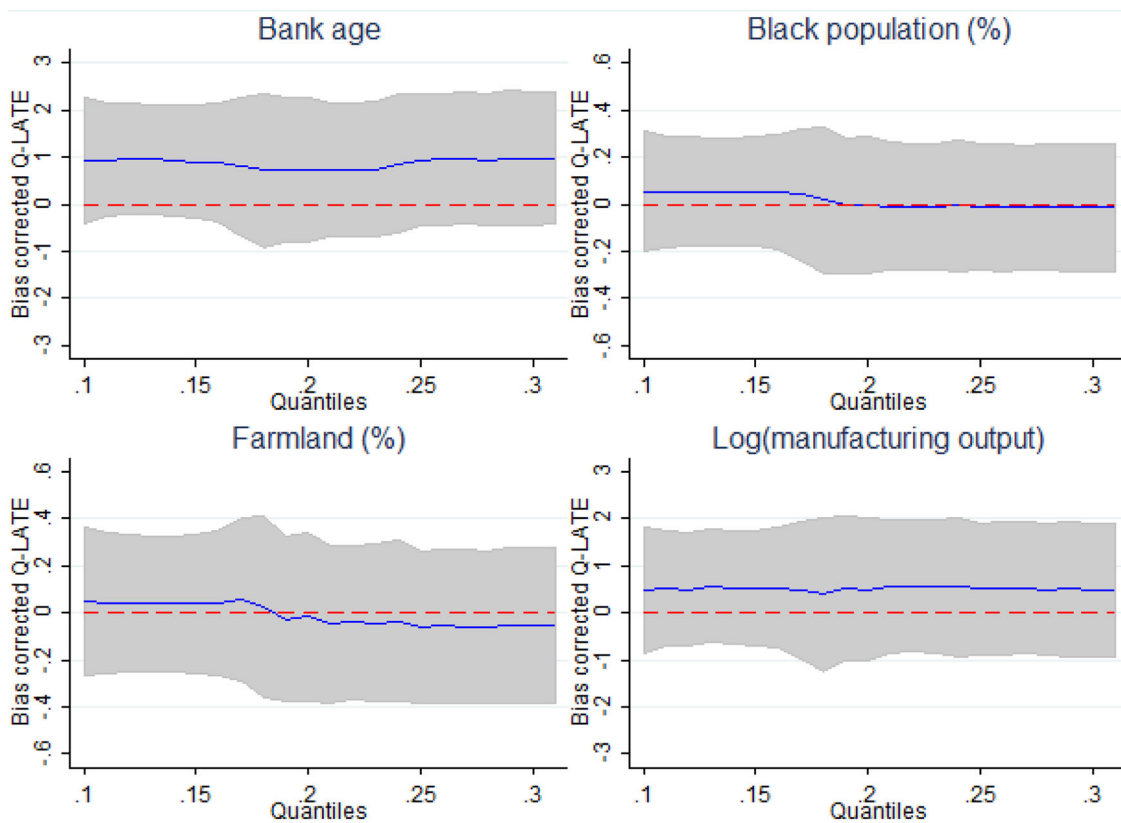


Figure 5. Bias-corrected estimates of Q-LATEs on covariates (first moments).

increase in assets among those banks at lower quantiles of the capital distribution. The estimated impacts on log leverage and those on the long-run risk of suspending operation are small and insignificant.

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

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

Overall, our empirical analysis suggests that while the minimum capital requirement induces small banks (i.e., banks at the bottom 30% of the capital distribution) to hold more capital, these banks adjust their assets proportionately. That

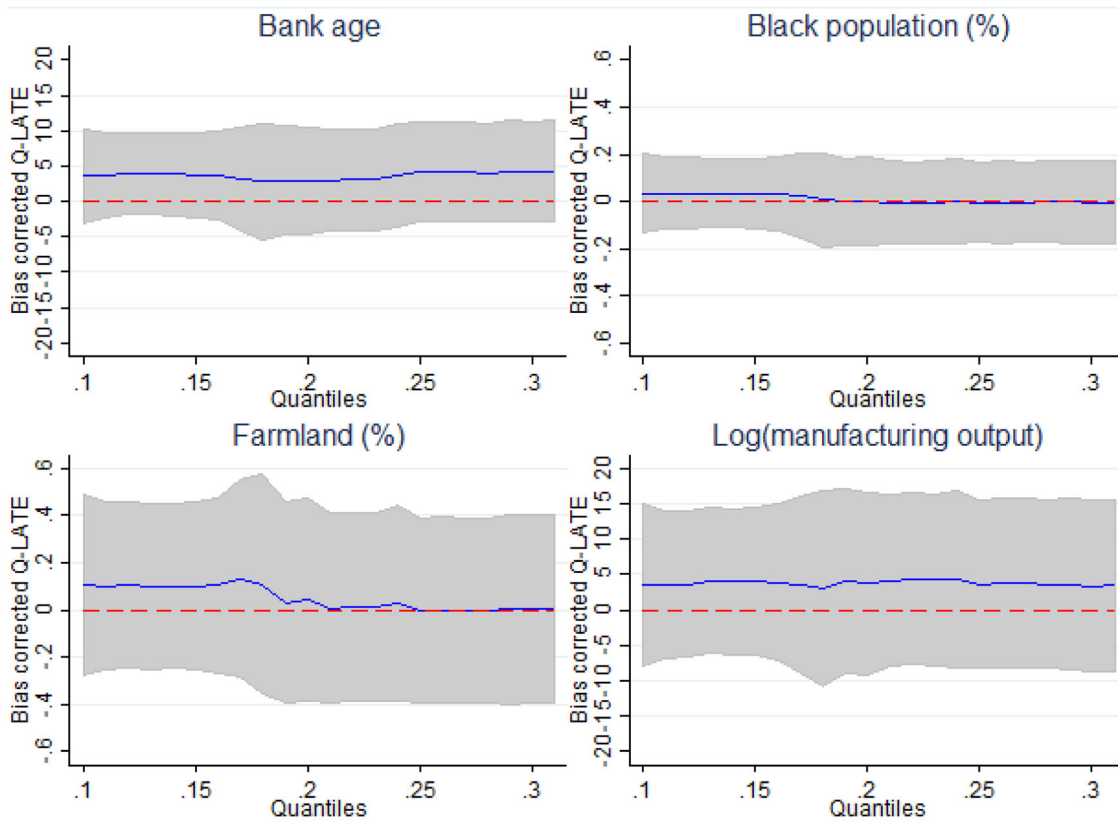


Figure 6. Bias-corrected estimates of Q-LATEs on covariates (second moments).

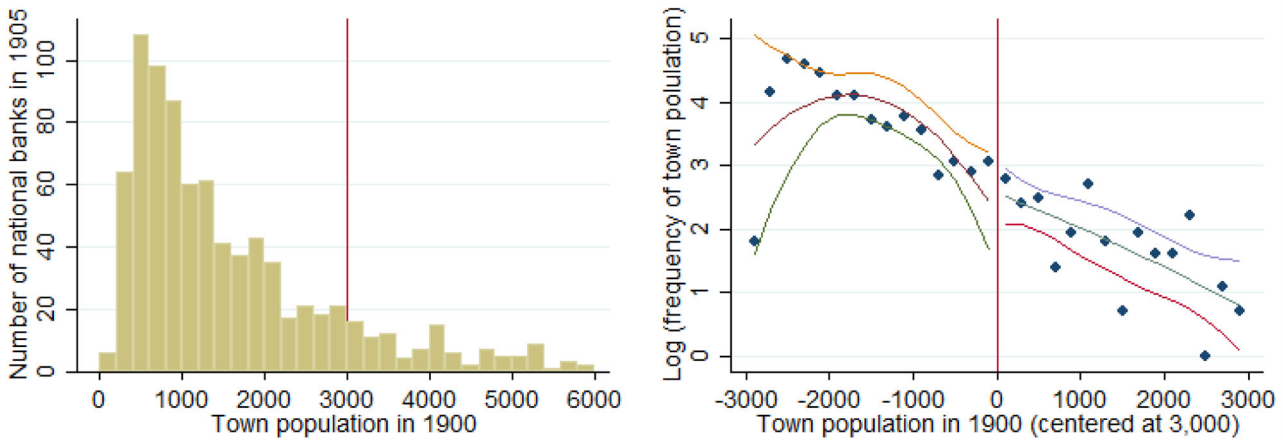


Figure 7. Histogram and the empirical density of town population.

is, banks simply scale up without a ratio regulation. As a result, their leverages and long-run risk of failure remain almost unchanged. These results help us better understand the frequent bank runs and banking panics prior to the Great Depression.

6.2. Validity Checks

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

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

In addition to our joint tests, we also perform the standard RD validity checks, including the density test and covariates smoothness test. Details of these tests and formal testing results

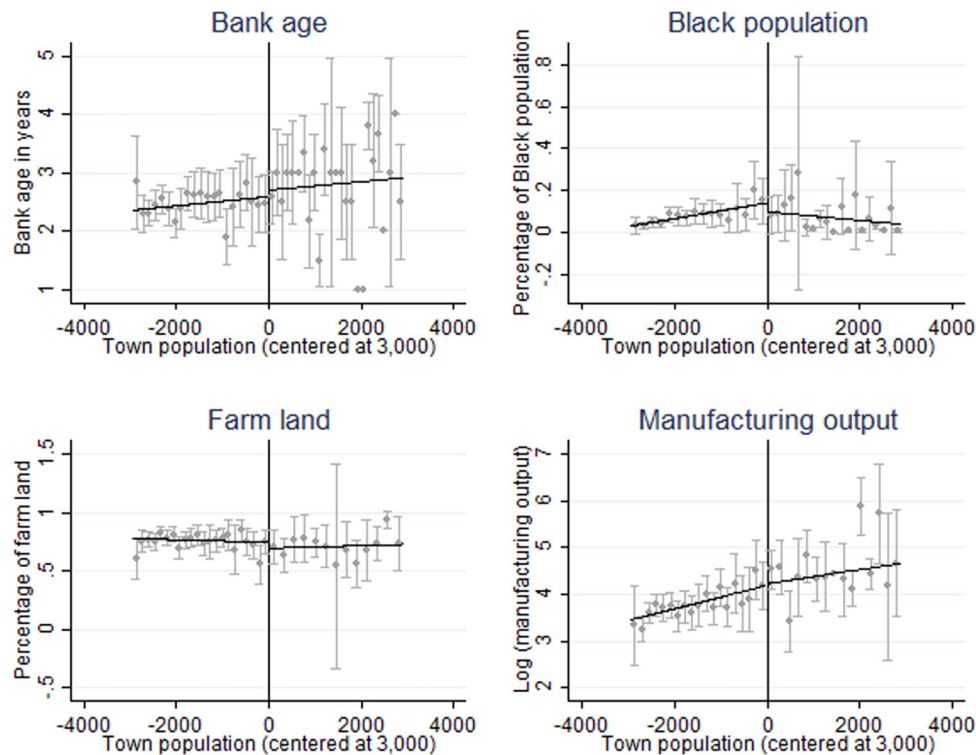
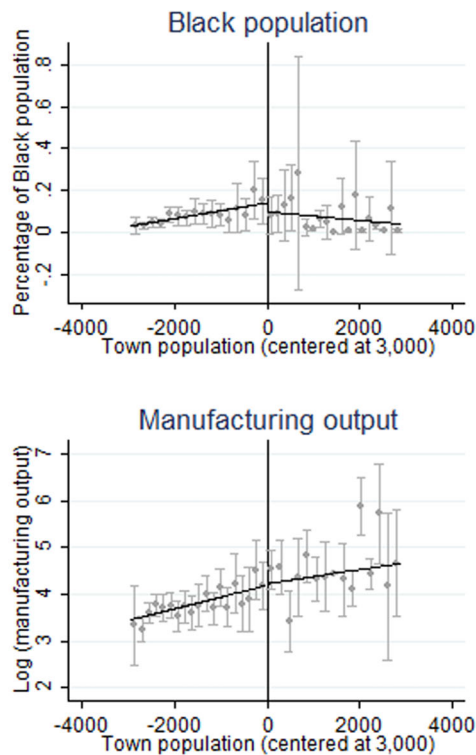


Figure 8. Conditional means of covariates conditional on town population.

are provided in Appendix D.3. Figure 7 presents the histogram of the town population (left) and the log frequency of the town population within each bin of 200 population (right). Superimposed on the right graph is the estimated log density along with the 95% confidence interval. Figure 8 plots the mean of the covariate in a bin of town population against the midpoint of the bin. The bars mark the 95% confidence intervals. Overall we do not find evidence that banks took advantage of the lower capital requirement and hence were more likely to operate in towns with populations just under 3000. Results of our validity checks strongly support the plausibility of our assumptions.

7. Conclusion

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



Supplementary Materials

The supplemental appendix provides proofs of the theoretical results presented in the article, preliminary lemmas, alternative inference based on undersmoothing, and details of estimating the biases and variances of the proposed estimators as well as the AMSE optimal bandwidths. It also provides data description and additional results for the empirical application.

References

- Almond, D., Doyle, J. J., Kowalski, A. E., and Williams, H. (2010), "Estimating Marginal Returns to Medical Care: Evidence From At-Risk Newborns," *The Quarterly Journal of Economics*, 125, 591–634. [1,4]
- Angrist, J. D., Imbens, G., and Graddy, K. (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. [4,5]
- Arai, Y., and Ichimura, H. (2018), "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," *Quantitative Economics*, 9, 441–482. [8,9]
- Arkhangelsky D., and Imbens, G. W. (2021), "Double-Robust Identification for Causal Panel Data Models," NBER Working Paper No. w28364. [5]
- Caetano, C., Caetano, G., and Escanciano, J. C. (2020), "Regression Discontinuity Design With Multivalued Treatments," Working paper. [2]
- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014), "Robust Nonparametric Bias Corrected Inference in Regression Discontinuity Design," *Econometrica*, 82, 2295–2326. [7,9]
- Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018), "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113, 767–779. [7]
- (2019), "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," arXiv:1808.01398. [7,9]

- (2020), “Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression-Discontinuity Designs,” *Econometrics Journal*, 23, 192–210. [7,11]
- Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015), “Inference on Causal Effects in a Generalized Regression Kink Design,” *Econometrica*, 83, 2453–2483. [2]
- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2020), *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge: Cambridge University Press. [2]
- (2021), *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge: Cambridge University Press, to appear. [2]
- Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2020), “The Regression Discontinuity Design,” *Handbook of Research Methods in Political Science and International Relations*, eds. L. Curini and R. J. Franzese, Sage Publications, Ch. 44, pp. 835–857. [2]
- Chen, Y., Ebenstein, A., Greenstone, M., and Li, H. (2013), “Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy From China’s Huai River Policy,” *PNAS*, 110, 12936–12941. [1,4]
- Chernozhukov, V., and Hansen, C. (2005), “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73, 245–261. [3,5]
- D’haultfoeuille, X., and Février, P. (2015), “Identification of Nonseparable Triangular Models With Discrete Instruments,” *Econometrica*, 83, 1199–1210. [2]
- Ebenstein, A., Fan, M., Greenstone, M., He, G., and Zhou, M. (2017), “New Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China’s Huai River Policy,” *PNAS*, 114, 10384–10389. [1,4]
- Fan, M., He, G., and Zhou, M. (2020), “The Winter Choke: Coal-Fired Heating, Air Pollution, and Mortality in China,” *Journal of Health Economics*, 71, 102316. [1,4]
- Frandsen B., Frölich, M., and Melly, B. (2012), “Quantile Treatment Effects in the Regression Discontinuity Design,” *Journal of Econometrics*, 168, 382–395. [2]
- Giuntella, O., and Mazzonna, F. (2019), “Sunset Time and the Economic Effects of Social Jetlag: Evidence From US Time Zone Borders,” *Journal of Health Economics*, 65, 210–226. [1]
- Hahn, J., Todd, P., and van der Klaauw, W. (2001), “Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design,” *Econometrica*, 69, 201–209. [1,3]
- Horowitz, J. (2001), “Nonparametric Estimation of a Generalized Additive Model with an Unknown Link Function,” *Econometrica*, 69, 499–513. [9]
- Imbens, G., and Kalaynaraman, K. (2012), “Optimal Bandwidth Choice for the Regression Discontinuity Estimator,” *The Review of Economic Studies*, 79, 933–959. [8,9]
- Imbens, G. W., and Lemieux, T. (2008), “Regression Discontinuity Designs: A Guide to Practice,” *Journal of Econometrics*, 142, 615–635. [2]
- Imbens, G., and Newey, W. (2009), “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” *Econometrica*, 77, 1481–1512. [4]
- Kong, E., Linton, O., and Xia, Y. (2010), “Uniform Bahadur Representation for Local Polynomial Estimates of M-Regression and Its Application to the Additive Model,” *Econometric Theory*, 26, 1529–1564. [7]
- Litschig, S., and Morrison, K. (2010), “Government Spending and Re-Election: Quasi-Experimental Evidence from Brazilian Municipalities,” UPF Discussion Paper. [1]
- Newey, W. (1994), “Kernel Estimation of Partial Means and a General Variance Estimator,” *Econometric Theory*, 10, 233–253. [9]
- Qu, Z., and Yoon, J. (2015), “Nonparametric Estimation and Inference on Conditional Quantile Processes,” *Journal of Econometrics*, 185, 1–19. [7]
- Shen, S., and Zhang, X. (2016), “Distributional Tests for Regression Discontinuity: Theory and Empirical Examples,” *The Review of Economics and Statistics*, 98, 685–700. [7]
- Torgovitsky, A. (2015), “Identification of Nonseparable Models Using Instruments With Small Support,” *Econometrica*, 83, 1185–1197. [2]