3-2019

Working Paper Number 19007

Wild bootstrap for fuzzy regression discontinuity designs: obtaining robust bias-corrected confidence intervals

Yang He
Freddie Mac

Otávio Bartalotti
Iowa State University, bartalot@iastate.edu

Wild bootstrap for fuzzy regression discontinuity designs: obtaining robust bias-corrected confidence intervals

Abstract
This paper develops a novel wild bootstrap procedure to construct robust bias- corrected valid confidence intervals (CIs) for fuzzy regression discontinuity designs, providing an intuitive alternative to existing analytical methods. The CIs generated by this procedure are valid under conditions similar to the standard analytical procedures used in the empirical literature. Simulations provide evidence that this new method is at least as accurate as the analytical corrections when applied to a variety of data generating processes featuring heteroskedasticity, endogeneity and clustering. Finally, we demonstrate its empirical relevance by revisiting Angrist and Lavy (1999) analysis of class size on student outcomes.

Keywords
Fuzzy Regression Discontinuity, Robust Confidence Intervals, Wild Boot- strap, Average Treatment Effect

Disciplines
Econometrics | Economic Theory

Wild bootstrap for fuzzy regression discontinuity designs: obtaining robust bias-corrected confidence intervals.

Yang He and Otávio Bartalotti

First Draft: December, 2016.
This Draft: March, 2019.

Abstract

This paper develops a novel wild bootstrap procedure to construct robust bias-corrected valid confidence intervals (CIs) for fuzzy regression discontinuity designs, providing an intuitive alternative to existing analytical methods. The CIs generated by this procedure are valid under conditions similar to the standard analytical procedures used in the empirical literature. Simulations provide evidence that this new method is at least as accurate as the analytical corrections when applied to a variety of data generating processes featuring heteroskedasticity, endogeneity and clustering. Finally, we demonstrate its empirical relevance by revisiting Angrist and Lavy (1999) analysis of class size on student outcomes.
Key Words: Fuzzy Regression Discontinuity, Robust Confidence Intervals, Wild Bootstrap, Average Treatment Effect.

1 Introduction

Regression discontinuity (RD) designs are one of the leading empirical approaches in economics, political science, and public policy evaluation, being extensively used to estimate the causal effects of treatments or policies under transparent assumptions.[1] The identification in RD designs exploits the fact that many policies and programs use a threshold based on a score, also called a "running variable" to assign treatment to individuals or firms. In that case, if the researcher credibly believes that subject's position relative to the threshold is not related to unobserved characteristics driving the outcome of interest, we can attribute the differences between units slightly above and below the cutoff as caused by treatment alone. When the running variable does not

*He: Freddie Mac, 1551 Park Run Dr, McLean, VA 22102. Email: y.he0802@gmail.com. Bartalotti: Department of Economics, Iowa State University and IZA. 260 Heady Hall, Ames, IA 50011. Email: bartalot@iastate.edu.

[1]Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide reviews of this literature with many examples.

entirely determine the treatment, there are both treated and untreated units on each side of the cutoff, a situation referred to as the fuzzy RD design. Directly comparing the outcomes on both sides of the cutoff results in an intent-to-treat effect, and the average treatment effect at the cutoff can be recovered by taking the ratio of difference in outcomes and difference in treatment probabilities at the threshold, as in a Wald formulation of the treatment effect in the instrumental variable setting. Even when units are self-selected to treatment based on anticipated gains, Hahn et al. (2001) show that this ratio can be interpreted as the local average treatment effect (LATE) under proper assumptions.

The identification of RD designs occurs exactly at the cutoff, and in practice the treatment effect is typically estimated by fitting local linear models above and below the threshold, which are extrapolated to the exact point of discontinuity.[2][3] The choice of the bandwidth, h, in these nonparametric estimators, is an important econometric issue which controls the trade-off between bias and variance. One popular bandwidth selector proposed by Imbens and Kalyanaraman (2012) and extended by Calonico et al. (2014), henceforth "CCT," minimizes the asymptotic mean squared error (AMSE) of the treatment effect estimator. However, this bandwidth selector has the form $h = O_p(n^{-1/5})$, where n is the number of observations. As pointed out by CCT, the AMSE-optimal bandwidth shrinks slowly enough that the leading bias term in the local polynomial estimators will be non-negligible, affecting the asymptotic distribution of the estimator. Consequently, the usual confidence intervals (CIs) for the RD treatment effects are invalid, and simulation studies on sharp RD designs in CCT, confirm that conventional CIs have empirical coverage well below their nominal levels.

CCT solve this problem by obtaining a valid estimate of the leading bias-term and re-centring the conventional point estimator. Furthermore, the additional variability introduced by the bias estimation needs to be considered when constructing CIs. This approach results in a bias-corrected point estimator which is asymptotically normal with weaker assumptions on the bandwidth. CIs based on this method are valid even when AMSE optimal bandwidths are used.

In this paper, a wild bootstrap procedure is proposed as an alternative to the analytical methods for bias-corrected robust inference for fuzzy RD designs. We theoretically show that the new bootstrap procedure is asymptotically equivalent to CCT's and provide simulation evidence that it performs well in finite samples. Compared with the analytical method, the bootstrap procedure is straightforward and does not require intensive derivations. Additionally, since the bootstrap is motivated by mimicking the true data generating process, it has the flexibility to accommodate dependent (clustered) data by adjusting the resampling algorithm accordingly. We demonstrate how the proposed bootstrap procedure can be applied to clustered data.

The wild bootstrap procedure exploits CCT's theoretical insight by resampling from higher order local polynomials. In particular, the local linear models are estimated as usual for both outcome and treatment, resulting in a conventional biased estimator. To estimate the bias, additional local quadratic models are used, and the potentially correlated residuals on both the outcome and treatment equations serve as the "true"

[2]See Fan (1992); Hahn et al. (2001) for a detailed discussion on local polynomial estimator properties and its use in RD designs.

[3]The following discussion is similar to the description in Bartalotti et al. (2017).

data generating process (DGP) for the bootstrap. The bias of the conventional estimator is therefore known under this bootstrap DGP and can be calculated by averaging the error of the linear model's estimates across bootstrap replications. Any remaining bias converges to zero at a faster rate, allowing the bias of the local linear model to be estimated. This approach is described in Algorithm 3.1 and the resulting bias-corrected estimator is shown to be asymptotically normal with mean zero in Theorem 3.1.

Following Bartalotti et al. (2017) we propose an iterated bootstrap procedure to account for the additional variability introduced by the bias correction: generate many bootstrap datasets from local quadratic models and calculate the bias-corrected estimate for each of them. The resulting empirical distribution of bias-corrected estimator is then used to construct CIs. This procedure is in line with CCT's approach, where the variance of estimated bias term and the covariance between estimated bias and original point estimator are derived analytically. This complex adjustment to the original variance is automatically embedded in the iterated bootstrap. The bootstrap implementation is described in Algorithm 3.2, and the resulting CIs are shown to be asymptotically valid in Theorem 3.2.

Relative to Bartalotti et al. (2017), which developed a similar iterated bootstrap procedure for robust inference in special case of sharp RD designs, the current paper provides important generalizations in several dimensions. First, it connects the idea of bootstrapping IV models and adapts that to a more general fuzzy RD design. Second, its validity is extended and theoretically proved to general local polynomials and higher order of derivatives of interests, which could be used in the context of "Kink" RD designs, for example. Last, its flexibility and capability to accommodate clustered data is discussed and confirmed by simulation studies.

Concomitantly and independently, Chiang et al. (2017) proposed a multiplier bootstrap procedure that could be used in fuzzy RD and many related general settings based on a Bahadur representation of a general class of Wald estimators. Both the procedure and proofs in that paper differ from the ones proposed here and could potentially serve as alternatives in the cases covered by both approaches. Nevertheless, our procedure benefits from its very intuitive nature, easy implementation and flexibility as exemplified in Section 5 when dealing with dependent (clustered) data.

The paper is organized as follows. Section 2 describes the basic fuzzy RD approach, its usual implementation, and the CCT's robust inference method. Section 3 presents the proposed bootstrap procedures to estimate bias and construct confidence interval. Their asymptotic properties are discussed and summarized in two theorems. Section B provides simulation evidence that the bootstrap procedure effectively reduces bias and generates valid CIs. Implementation of the bootstrap to clustered data is discussed in Section 5. Section 6 demonstrates the applied relevance of this bootstrap procedure by applying it to the scholastic achievement data used by Angrist and Lavy (1999). Finally, Section 7 concludes.

2 Background

This section provides additional details of identification assumptions and traditional estimation methods in fuzzy RD designs. It also briefly introduces the robust confi-

dence interval proposed by CCT. Notations defined in this and following sections are consistent with CCT.

In a typical fuzzy RD setting, researchers are interested in the local causal effect of treatment at a given cutoff. For any unit i, (X_i, T_i, Y_i) is observed, where X_i is a continuous running variable which determines treatment assignment, T_i is a binary variable which indicates actual treatment status and Y_i is the outcome. In sharp RD designs, the treatment actually received is the same as the assigned treatment, i.e., $T_i = \mathbb{1}(X_i \geq c)$, with c being the cutoff. In fuzzy RD designs, however, the received treatment is not a deterministic function of running variable X_i. Instead, the probability $\Pr(T_i = 1 \mid X_i)$ is between zero and one in both sides but experiences a sudden change at the cutoff. Without loss of generality, the cutoff c can be reset to zero. If assigned to treatment $(X_i \geq 0)$, unit i's actual treatment status and outcome are represented by functions $T_i(1)$ and $Y_i(1)$, otherwise $T_i(0)$ and $Y_i(0)$. Thus the observed treatment status and outcome are

$$\begin{aligned} T_i &= T_i(0)\,\mathbb{1}(X_i < 0) + T_i(1)\,\mathbb{1}(X_i \geq 0) \\ Y_i &= Y_i(0)\,\mathbb{1}(X_i < 0) + Y_i(1)\,\mathbb{1}(X_i \geq 0). \end{aligned}$$

For each unit i's outcome, either $Y_i(0)$ or $Y_i(1)$ is observed. The data itself is uninformative in terms of treatment effect because the counterfactual outcome could be arbitrary. However, under continuity and smoothness conditions on $T_i(0), Y_i(0), T_i(1)$ and $Y_i(1)$ around the cutoff $X_i = 0$, it is possible to identify the treatment effect for units just at the cutoff and the estimand of interest is

$$\zeta = \frac{\tau_Y}{\tau_T} = \frac{\mathrm{E}[Y_i(1) \mid X_i = 0] - \mathrm{E}[Y_i(0) \mid X_i = 0]}{\mathrm{E}[T_i(1) \mid X_i = 0] - \mathrm{E}[T_i(0) \mid X_i = 0]}, \tag{2.1}$$

where the symbol E represents the expectation and τ_Y and τ_T represent the sharp RD estimators, i.e., difference in expectations at the cutoff. Intuitively, this is a Wald estimator in the limit where the assigned treatment serves as an instrument. The reduced-form difference in expected outcome, τ_Y, reveals the "intent-to-treat" (ITT) effect. The treatment effect is recovered by dividing ITT effect by the difference in treatment probabilities. When the treatment effect is not constant across units, ζ should be interpreted with caution. If treatment status is independent of treatment effects at the cutoff, ζ is the average treatment effect (ATE) at the cutoff. This assumption rules out self-selection based on anticipated gain. Hahn et al. (2001) show that under a less restrictive assumption that the running variable is independent of the joint distribution of treatment effect and treatment status at the cutoff, the LATE is identified.

Equation 2.1 presents ζ as a ratio of two sharp RD estimators. Due to this symmetry, we use "Z" as a placeholder for either outcome variable Y or treatment variable T to ease the notation. In addition, denote the conditional expectations $\mu_{Z+}(x)$ and $\mu_{Z-}(x)$, conditional variances $\sigma^2_{Z+}(x)$ and $\sigma^2_{Z-}(x)$, the η-th order derivative of conditional expectations $\mu^{(\eta)}_{Z+}(x)$ and $\mu^{(\eta)}_{Z-}(x)$ and their limits. Formally, they are defined

as

$$\mu_{Z+}(x) = \mathrm{E}[Z_i(1) \mid X_i = x] \qquad \mu_{Z-}(x) = \mathrm{E}[Z_i(0) \mid X_i = x]$$
$$\sigma^2_{Z+}(x) = \mathbb{V}[Z_i(1) \mid X_i = x] \qquad \sigma^2_{Z-}(x) = \mathbb{V}[Z_i(0) \mid X_i = x]$$
$$\mu^{(\eta)}_{Z+}(x) = \frac{d^\eta \mu_{Z+}(x)}{dx^\eta} \qquad \mu^{(\eta)}_{Z-}(x) = \frac{d^\eta \mu_{Z-}(x)}{dx^\eta}$$
$$\mu^{(\eta)}_{Z+} = \lim_{x \to 0^+} \mu^{(\eta)}_{Z+}(x) \qquad \mu^{(\eta)}_{Z-} = \lim_{x \to 0^-} \mu^{(\eta)}_{Z-}(x)$$

where the symbol $\mathbb{V}(\cdot)$ represents variance. The treatment effect ζ is nonparametrically estimable because μ_{Z-} and μ_{Z+} can be estimated consistently under Assumption 2.1, which lists standard conditions in the fuzzy RD literature.[4] (See, in particular, Hahn et al., 2001, Porter, 2003 and CCT.)

Assumption 2.1 (Behavior of the DGP near the cutoff) *The random variables $\{X_i, T_i, Y_i\}_{i=1}^n$ form a random sample of size n. There exists a positive number κ_0 such that the following conditions hold for all x in the neighbourhood $(-\kappa_0,\ \kappa_0)$ around zero: (a) The density of X_i is continuous and bounded away from zero at x; (b) $\mathrm{E}[Z_i^4 \mid X_i = x]$ is bounded; (c) $\mu_{Z-}(x)$ and $\mu_{Z+}(x)$ are three times continuously differentiable; (d) $\sigma^2_{Z-}(x)$ and $\sigma^2_{Z+}(x)$ are continuous and bounded away from zero; (e) $\mu_{T-}(0) \neq \mu_{T+}(0)$.*

Assumption 2.1(a) ensures that the number of data points arbitrarily close to the cutoff increases as the sample size grows. Part (c) imposes necessary smoothness condition to allow an approximation by second order polynomials. Parts (b) and (d) put standard restrictions on moments to ensure that the estimated local polynomials are well behaved. Part (e) requires that the treatment assignment as an instrument is valid, in the sense that it induces a first stage difference in treatment probability. In practice, local polynomial regression is widely used to estimate RD designs because of nice boundary properties.[5] As an illustration, consider the local linear regression using kernel function $K(\cdot)$ with a common bandwidth, h, used for both the outcome and the treatment at both sides of the cutoff. The estimated treatment effect is

$$\hat{\zeta}(h) = \frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)} = \frac{\hat{\mu}_{Y+}(h) - \hat{\mu}_{Y-}(h)}{\hat{\mu}_{T+}(h) - \hat{\mu}_{T-}(h)}, \tag{2.2}$$

[4] Throughout the main text we focus on the case where the researcher implements a local linear model to estimate τ_Z and a quadratic model to approximate the bias term. The proofs presented in the appendix for the validity of the bootstraps proposed include the general case in which higher-order polynomials can be used to obtain τ_Z or a higher-order bias correction is implemented, e.g., Bartalotti (2018).

[5] See Fan and Gijbels (1996) for discussions on the boundary properties of local polynomial regression. See Gelman and Imbens (2018) for discussions on the choices of global and local polynomial regression and its order.

with

$$\hat{\mu}_{Z+}(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{X_i \geq 0\}(Z_i - \beta_0 - X_i\beta_1)^2 \frac{1}{h} K\left(\frac{X_i}{h}\right)$$
$$\hat{\mu}_{Z-}(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{X_i < 0\}(Z_i - \beta_0 - X_i\beta_1)^2 \frac{1}{h} K\left(\frac{X_i}{h}\right).$$

The conditional expectations μ_{Z+} and μ_{Z-} are consistently estimated by $\hat{\mu}_{Z+}(h)$ and $\hat{\mu}_{Z+}(h)$ when $h \to 0$.[6] The asymptotic distribution of the quotient estimator $\frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)}$ can be derived by applying the delta method. Let V_Z be the asymptotic variance of $\hat{\tau}_Z(h)$ and C_{YT} be the asymptotic covariance between $\hat{\tau}_Y(h)$ and $\hat{\tau}_T(h)$, i.e.,

$$\begin{pmatrix} \sqrt{nh}(\hat{\tau}_Y(h) - \tau_Y) \\ \sqrt{nh}(\hat{\tau}_T(h) - \tau_T) \end{pmatrix} \xrightarrow{d} N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_Y & C_{YT} \\ C_{YT} & V_T \end{pmatrix}\Big),$$
$$\sqrt{nh}(\hat{\zeta}(h) - \zeta) \xrightarrow{d} N\Big(0, \frac{1}{\tau_T^2} V_Y - \frac{2\tau_Y}{\tau_T^3} C_{YT} + \frac{\tau_Y^2}{\tau_T^4} V_T\Big).$$

Let $V(h) = \mathbb{V}(\hat{\zeta}(h) \mid X_1, ..., X_n)$, then $\frac{\hat{\zeta}(h) - \zeta}{\sqrt{V(h)}} \xrightarrow{d} N(0,1)$ and the CIs can be constructed as

$$\hat{\zeta}(h) \pm q_{1-\alpha/2} V(h)^{1/2} \tag{2.3}$$

where $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

The above asymptotic distribution is valid only when bandwidth h shrinks fast enough such that the bias of $\hat{\zeta}_Z(h)$ is negligible relative to $\sqrt{V(h)}$. Formally, $h = o_p(n^{-1/5})$ is required. With a bandwidth of order $O_p(n^{-1/5})$, Hahn et al. (2001) show that the asymptotic distribution is normal but not centred at zero. Using (2.3) to construct CIs without considering this first-order bias in distributional approximation leads to coverage rates that differ from the nominal level. Imbens and Kalyanaraman (2012) develop plug-in bandwidth selector for RD estimators, which is optimal in the sense that AMSE of the point estimator is minimized.

Two different approaches are commonly adopted in empirical studies. One is undersmoothing. In this case, instead of using the AMSE-optimal bandwidth, researchers may want to choose a smaller bandwidth in order to meet the requirement of $h = o_p(n^{-1/5})$. However, this often leads to a series of ad-hoc bandwidths without theoretical basis. Another approach is bias correction, in which the leading bias is consistently estimated to remove the distortion of the asymptotic approximation. However, this approach does not perform well initially because the estimated bias introduces additional variability. CCT's approach is based on bias correction, but derives an alternative asymptotic variance component for normalization so that the additional variability is accounted for.

For any bandwidth $h \to 0$, the first-order bias of fuzzy RD estimator from local linear regression is

$$\mathrm{E}[\hat{\zeta}(h) \mid X_1, ..., X_n] - \zeta = h^2\Big(\frac{1}{\tau_T} \boldsymbol{B}_Y(h) - \frac{\tau_Y}{\tau_T^2} \boldsymbol{B}_T(h)\Big)(1 + o_p(1)), \tag{2.4}$$

[6] Unless otherwise stated, all limits in this paper are assumed to hold as $n \to \infty$.

with

$$\boldsymbol{B}_Z(h) = \frac{\mu_{Z+}^{(2)}}{2}\mathfrak{B}_+(h) - \frac{\mu_{Z-}^{(2)}}{2}\mathfrak{B}_-(h).$$

The terms $\mathfrak{B}_+(h)$ and $\mathfrak{B}_-(h)$, explicitly defined in appendix, are observed quantities that depend on the kernel, bandwidth and running variable. To explicitly calculate the first-order bias, one needs to estimate τ_Z, $\mu_{Z+}^{(2)}$ and $\mu_{Z-}^{(2)}$. Among them τ_Z is consistently estimated by the local linear regression with bandwidth h. CCT propose a local second-order regression with a (potentially) different bandwidth, b, to estimate the second order derivatives $\mu_{Z+}^{(2)}$ and $\mu_{Z-}^{(2)}$. This produces the bias-corrected estimator

$$\hat{\zeta}^{bc}(h,b) = \hat{\zeta}(h) - \Delta(h,b),$$

with

$$\Delta(h,b) = h^2\Big(\frac{1}{\hat{\tau}_T(h)}\hat{\boldsymbol{B}}_Y(h,b) - \frac{\hat{\tau}_Y(h)}{\hat{\tau}_T^2(h)}\hat{\boldsymbol{B}}_T(h,b)\Big),$$

$$\hat{\boldsymbol{B}}_Z(h,b) = \frac{\hat{\mu}_{Z+}^{(2)}(b)}{2}\mathfrak{B}_+(h) - \frac{\hat{\mu}_{Z-}^{(2)}(b)}{2}\mathfrak{B}_-(h).$$

Notice that the bias $\Delta(h,b)$ is estimated with uncertainty. As a result, the variance of bias-corrected estimator $\hat{\zeta}^{bc}(h,b)$ is different from $V(h)$. CCT propose a new formula for the variance of bias-corrected estimator and use it for normalization:

$$\frac{\hat{\zeta}^{bc}(h,b) - \zeta}{V^{bc}(h,b)^{1/2}} \xrightarrow{d} N(0,1), \tag{2.5}$$

where $V^{bc}(h,b) = V(h) + C(h,b)$ and $C(h,b)$ captures the adjustment to variance introduced by the bias-correction term. This distributional approximation is valid even when $h = O_p(n^{-1/5})$, as long as certain conditions on h and b are satisfied. Assumption 2.2 specifies the bandwidth and kernel conditions assumed by CCT, which are also used in this paper.

Assumption 2.2 (Bandwidth and kernel) *Let h be the bandwidth used to estimate the local linear model and let b be the bandwidth used to estimate the local quadratic model used to estimate the bias correction. Then $nh \to \infty$, $nb \to \infty$, and $n \times \min(h,b)^5 \times \max(h,b)^2 \to 0$ as $n \to \infty$. The kernel function $K(\cdot)$ is positive, bounded, and continuous on the interval $[-\kappa,\ \kappa]$ and zero outside that interval for some $\kappa > 0$.*

Assumption 2.2 does not require $nh^{1/5} \to 0$. Instead, it only requires that $nh^{1/5}b^{1/2} \to 0$ when $h < b$ or $nb^{1/5}h^{1/2} \to 0$ when $h > b$. This assumption also allows for the vast majority kernel functions commonly used in practice.

To simplify notation, let $m = \min(h,b)$ and define the scaled and truncated kernel functions

$$K_{+,h}(x) = \tfrac{1}{h}K(x/h)\,\mathbb{1}\{x \geq 0\} \qquad K_{-,h}(x) = \tfrac{1}{h}K(x/h)\,\mathbb{1}\{x < 0\}$$

and

$$K_{+,b}(x) = \tfrac{1}{b}K(x/b)\,\mathbb{1}\{x \geq 0\} \qquad K_{-,b}(x) = \tfrac{1}{b}K(x/b)\,\mathbb{1}\{x < 0\}.$$

In the next section, a simple bootstrap procedure is proposed to construct robust CIs based on the insight provided by CCT's bias-corrected estimator. This bootstrap procedure is straightforward in the sense that no derivation of analytical formulas for the bias, variance and covariance terms is required. The bias-corrected estimator and its confidence interval are numerically different from CCT's but asymptotically equivalent.

3 Bootstrap Algorithm and Validity

In this section, two bootstrap algorithms are presented to obtain bias-corrected point estimator and its CIs in the fuzzy RD design, extending the results in Bartalotti et al. (2017). Their validity is justified in two theorems and proved in the appendix. The idea behind both algorithms is to use higher-order local polynomials to approximate the joint behaviour of (X_i, T_i, Y_i) around the cutoff. These polynomials, together with the empirical variance structure, serve as the "true" DGP in the bootstrap under which we evaluate the bias of the local linear estimator employed by the researcher when implementing RD. Assumption 2.2 guarantees that the estimated "true" DGP is close to the unknown DGP in the sense that distributional approximation derived from the "true" DGP is asymptotically valid. This can be best illustrated from the special case where the bandwidths used for estimating τ and the bias are the same, $h = b$, which translates to the bandwidth condition $nb^7 \to 0$ under Assumption 2.2. By the same argument that $h = o_p(n^{-1/5})$ is required for valid inference in a RD design estimated by local linear regression, $b = o_p(n^{-1/7})$ is required in a RD design estimated by local quadratic regression.

Algorithm 3.1 consistently estimates the bias term.

Algorithm 3.1 (Bias estimation) Assume h and b are bandwidths as defined by Assumption 2.2.

STEP *1.* Estimate local second order polynomials $\hat{g}_{Z-}$ and $\hat{g}_{Z+}$ with least squares using $K_{-,b}$ and $K_{+,b}$ for weights:

$$\hat{g}_{Z-}(x) = \hat{\beta}_{Z-,0} + \hat{\beta}_{Z-,1}x + \hat{\beta}_{Z-,2}x^2, \quad \hat{g}_{Z+}(x) = \hat{\beta}_{Z+,0} + \hat{\beta}_{Z+,1}x + \hat{\beta}_{Z+,2}x^2$$

with

$$(\hat{\beta}_{Z-,0}, \hat{\beta}_{Z-,1}, \hat{\beta}_{Z-,2})' = \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \sum_{i=1}^{n} (Z_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2 K_{-,b}(X_i)$$

$$(\hat{\beta}_{Z+,0}, \hat{\beta}_{Z+,1}, \hat{\beta}_{Z+,2})' = \underset{\beta_0,\beta_1,\beta_2}{\arg\min} \sum_{i=1}^{n} (Z_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2 K_{+,b}(X_i).$$

Let

$$\hat{g}_Z(x) = \begin{cases} \hat{g}_{Z-}(x) & \textit{if } x < 0 \\ \hat{g}_{Z+}(x) & \textit{otherwise} \end{cases}$$

and calculate the residuals $\hat{\varepsilon}_{Zi} = Z_i - \hat{g}_Z(X_i)$ for all i.

STEP *2.* Repeat the following steps B_1 times to produce the bootstrap estimates $\hat{\eta}_1^*(h), \ldots, \hat{\eta}_{B_1}^*(h)$. For the kth replication:

2.1. Draw i.i.d. random variables e_i^* with mean zero, variance one, and bounded fourth moments independent of the original data and construct

$$\begin{aligned} \varepsilon_{Zi}^* &= \hat{\varepsilon}_{Zi} e_i^*, \\ Z_i^* &= \hat{g}_Z(X_i) + \varepsilon_{Zi}^* \end{aligned}$$

for all i.

2.2. Calculate $\hat{\mu}_{Z+}^*(h)$ and $\hat{\mu}_{Z-}^*(h)$ by estimating the local linear model on the bootstrap data set using $K_{+,h}$ and $K_{-,h}$ for weights:

$$\hat{\mu}_{Z-}^*(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{-,h}(X_i)$$

$$\hat{\mu}_{Z+}^*(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{+,h}(X_i).$$

2.3. Save $\hat{\zeta}_k^*(h) = \frac{\hat{\mu}_{Y+}^*(h) - \hat{\mu}_{Y-}^*(h)}{\hat{\mu}_{T+}^*(h) - \hat{\mu}_{T-}^*(h)}$.

STEP *3.* Estimate the bias as

$$\Delta^*(h,b) = \tfrac{1}{B_1} \sum_{k=1}^{B_1} \hat{\zeta}_k^*(h) - \frac{\hat{g}_{Y+}(0) - \hat{g}_{Y-}(0)}{\hat{g}_{T+}(0) - \hat{g}_{T-}(0)}. \tag{3.1}$$

Algorithm 3.1 consists of three steps. The first step estimates the bootstrap DGP, which is captured by second order local polynomials. The second step creates a series of new samples through wild bootstrap and finds the traditional fuzzy RD estimate for each sample. Crucial for the procedure is that pairs of residuals are multiplied by the same realization of random number e^* to preserve the correlation between Y_i and T_i. The last step calculates the bias from local linear estimator in the bootstrap by definition. Under Assumptions 2.1, 2.2 and B_1 large enough, the procedure described by Algorithm 3.1 consistently estimates the bias component, resulting in a bias-corrected estimator that has the same asymptotic distribution as in Equation (2.5). This conclusion is formally given in Theorem 3.1.

Theorem 3.1 *Under Assumptions 2.1 and 2.2,*

$$\frac{\hat{\zeta}(h) - \Delta^*(h,b) - \zeta}{V^{bc}(h,b)^{1/2}} \to^d N(0,1), \tag{3.2}$$

where $\Delta^*(h,b)$ *is defined by Equation* (3.1).

Theorem 3.1 enables one to construct valid confidence interval in the form of $\hat{\zeta}(h) - \Delta^*(h,b) \pm V^{bc}(h,b)^{1/2}$. However, the term $V^{bc}(h,b)$ still needs to be calculated. The second algorithm circumvents the analytical derivation of $V^{bc}(h,b)$ through an iterated

bootstrap. In particular, the first layer bootstrap is designed to mimic the randomness due to sampling error and the second layer bootstrap, as described in Algorithm 3.1, is designed to estimate bias due to model misspecification. The additional variability introduced by the bias correction term will be automatically accounted for by this iterated bootstrap. The detailed procedure is given in Algorithm 3.2.

Algorithm 3.2 (Distribution) Assume h and b are bandwidths as defined by Assumption 2.2 and Algorithm 3.1.

STEP *1.* Estimate $\hat{g}_{Z+}$ and $\hat{g}_{Z-}$ and generate $\hat{g}_Z(\cdot)$ and the residuals $\hat{\varepsilon}_{Zi}$ just as in Algorithm 3.1.

STEP *2.* Repeat the following steps B_2 times to produce bootstrap estimates of the bias-corrected estimate. For the kth replication:

2.1. Draw i.i.d. random variables e_i^* with mean zero, variance one, and bounded fourth moments independent of the original data and construct

$$\varepsilon^*_{Zi} = \hat{\varepsilon}_{Zi} e_i^*, \quad Z_i^* = \hat{g}_Z(X_i) + \varepsilon^*_{Zi}.$$

for all $i = 1, \ldots, n$.

2.2. Calculate $\hat{\mu}^*_{Z+}(h)$ and $\hat{\mu}^*_{Z-}(h)$ by estimating the local linear model on the bootstrap data set using $K_{+,h}$ and $K_{-,h}$ for weights:

$$\hat{\mu}^*_{Z-}(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{-,h}(X_i),$$

$$\hat{\mu}^*_{Z+}(h) = \arg\min_{\mu} \min_{\beta} \sum_{i=1}^{n} (Z_i^* - \mu - \beta X_i)^2 K_{+,h}(X_i).$$

2.3. Apply Algorithm 3.1 to the bootstrapped data set $(X_1, T_1^*, Y_1^*), \ldots, (X_n, T_n^*, Y_n^*)$ using the same bandwidths h and b that are used in the rest of this algorithm but reestimating all of the local polynomials on the bootstrap data. Generate B_1 new bootstrap samples and let $\Delta^{**}(h, b)$ represent the bias estimator returned by Algorithm 3.1.

2.4. Save the estimator $\hat{\zeta}_k^*(h) = \frac{\hat{\mu}^*_{Y+}(h) - \hat{\mu}^*_{Y-}(h)}{\hat{\mu}^*_{T+}(h) - \hat{\mu}^*_{T-}(h)}$, and its bias $\Delta_k^{**}(h, b)$.

STEP *3.* Define $\zeta^* = \frac{\hat{g}_{Y+}(0) - \hat{g}_{Y-}(0)}{\hat{g}_{T+}(0) - \hat{g}_{T-}(0)}$ and use the empirical CDF of $\hat{\zeta}_1^*(h) - \Delta_1^{**}(h, b) - \zeta^*, \ldots, \hat{\zeta}_{B_2}^*(h) - \Delta_{B_2}^{**}(h, b) - \zeta^*$ as the sampling distribution of $\hat{\zeta}(h) - \Delta^*(h, b) - \zeta$.

Algorithm 3.2 also consists of three steps. The first step estimates the bootstrap DGP, which is captured by second order local polynomials. The second step creates a series of new samples, to which the Algorithm 3.1 is applied. The last step uses the empirical distribution of bias-corrected estimator to construct CIs. As before, B_2 is assumed large enough so that simulation error can be ignored. The validity of Algorithm 3.2 is established in the following theorem.

Theorem 3.2 *Under Assumptions 2.1 and 2.2,*

$$\mathbb{V}^*(\hat{\zeta}^*(h) - \Delta^{**}(h, b))/V^{bc}(h, b) \to^p 1$$

and

$$\sup_x \left| \Pr^* \left[\frac{\hat{\zeta}^*(h) - \Delta^{**}(h,b) - \zeta^*}{\mathbb{V}^*(\hat{\zeta}^*(h) - \Delta^{**}(h,b))^{1/2}} \leq x \right] - \Pr \left[\frac{\hat{\zeta}(h) - \Delta^*(h,b) - \zeta}{V^{bc}(h,b)^{1/2}} \leq x \right] \right| \to^p 0.$$

Theorem 3.2 enables one to construct CIs in the following form:

$$\big(\hat{\zeta}(h) - \Delta^*(h,b) + \zeta^* - (\hat{\zeta}^*(h) - \Delta^{**}(h,b))_{1-\alpha/2}, \hat{\zeta}(h) - \Delta^*(h,b) + \zeta^* + (\hat{\zeta}^*(h) - \Delta^{**}(h,b))_{\alpha/2}\big),$$

where all the terms with superscript $*$ are defined in Algorithm 3.2. Different from the analytical one, this CI is not centred at the bias-corrected point estimator.

Remark 3.1 The proposed bias correction differs from CCT's analytical formula in finite samples. While the analytical bias is obtained by linearizing $\mathrm{E}\left[\frac{\hat{\tau}_Y(h)}{\hat{\tau}_T(h)} - \frac{\tau_Y}{\tau_T}\right]$ and then only evaluating its first order terms, Algorithm 3.1 directly estimates $\mathrm{E}\left[\frac{\hat{\tau}_Y^*(h)}{\hat{\tau}_T^*(h)} - \frac{\tau_Y^*}{\tau_T^*}\right]$ through bootstrap. Both methods consistently estimate the bias.

Remark 3.2 When the original treatment is binary, the bootstrap sample will no longer have binary treatment. Though it creates some difficulty for interpretation, it does not invalidate the estimation and inference because its conditional expectation and covariance with outcome variable remain unchanged.

Remark 3.3 The iterated bootstrap is less computationally intensive than it might initially appear due to two reasons. First, the wild bootstrap creates new residuals but leaves the regressors unchanged, which means the design matrices only need to be computed once even when they are repeatedly used in fitting local polynomials.[7] Second, the number of data points actually used in estimation is a lot smaller than the full sample due to the local nature of the estimation.

4 Monte Carlo Simulations

This section summarizes the result of Monte Carlo experiments designed to evaluate the finite sample performance of the bootstrap procedures proposed in Section 3 relative to the existing analytical alternatives. The details about the data generating processes (DGP) used and implementation are provided in the appendix.

The conditional mean functions used in the simulations are similar to the ones used by CCT, adapted to the fuzzy RD context. For concreteness, the first mean function (DGP 1) is designed to match features of U.S. congressional election data from Lee (2008). The second mean function (DGP 2) matches the relationship between children mortality rate and county poverty rate from analysis of Head Start data in Ludwig and

[7] To fit local polynomials is equivalent to estimate weighted least square, i.e., the estimated parameter is $(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{Y}$, where $\mathbf{X}$ is matrix of regressors and $\mathbf{K}$ is weighting matrix determined by kernel function. Both $\mathbf{X}$ and $\mathbf{K}$ are not affected by the bootstrap so one just need to compute $(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}$ once and then reuse it in the bootstrap calculations. Then each bootstrap replication just requires a single matrix-vector multiplication.

Miller (2007). The last mean function (DGP 3) is similar to the first one except for some coefficients. CCT motivates this as an attempt to generate plausible model with sizable distortion when conventional t-test is performed. The true treatment effects for these three models are $\zeta_1 = 0.04, \zeta_2 = -3.45, \zeta_3 = 0.04$, respectively.

To accommodate different endogeneity structures found in empirical data, we consider three cases. In the baseline case the treatment status is exogenous, i.e., there is no correlation between treatment assignment and the outcome. In the two endogenous cases the treatment status is correlated with unobserved characteristics which affect the outcome. This is modelled by the correlation, ρ, between the error terms on the outcome and treatment status equations as described in the appendix.

Besides the proposed bootstrap, two additional approaches are estimated for comparison: CCT's robust estimator and the conventional estimator. Simulation results are presented in Table 4.1. For the estimated treatment effect, its bias, standard error and root MSE are reported in the first three columns. For the CI, its empirical coverage and average length are reported in the fourth and fifth columns. The last three columns list the average bandwidths used in the two robust methods (h_{CCT}, b_{CCT}) and the conventional method (h_{IK}).

The baseline case is listed in Panel A. The two robust methods, wild bootstrap and CCT's approach, generate point estimates with very similar bias and standard errors. In contrast, the conventional approach reports 3-5 times larger bias. This is not surprising since the two robust methods explicitly correct the bias. The conventional method also fails to deliver a valid CI (coverage rates are 68.1%, 2.6% and 87.2% for the three DGPs respectively). Robust methods achieve improvements by reducing bias and increasing interval length. Except for DGP 2, they both generate intervals with empirical coverage close to the nominal level and the wild bootstrap performs well for DGPs 1 and 3. However, for DGP 2, even the robust methods report significant size distortion. This is because DGP 2 has great curvature around the cutoff and makes precise fitting challenging. Still, the two robust methods improve significantly on the coverage obtained by the conventional method (from 2.6% to around 87%) at the sacrifice of slightly longer intervals (from 0.186 to around 0.21).

Panels B and C present results when the treatment is endogenous, which is likely the primary reason to choose RD designs as the identification strategy. The case with positive (negative) ρ is listed in Panel B(C). Again, the estimate from the conventional method has significantly larger bias than the two robust methods. As for CIs, the wild bootstrap and CCT's approach work reasonably well in all cases. The conventional method performs significantly worse, with empirical coverage rate as low as 1.7%. The sign of correlation has little effect on the bias because the bias is caused by model misspecification rather than imperfect instrumental variable.

To summarize, the wild bootstrap approach proposed in this paper performs significantly better than the conventional method and is at least on par with the CCT's analytical methods. This wild bootstrap procedure automatically accommodates various types of covariance structure and thus is a simple alternative to obtain valid CIs in RD designs.

5 Extension: Clustered Data

This section explores the application of the bootstrap procedure to clustered data in RD designs and provides evidence for its usefulness. Clustered data are very common in empirical studies. Units within the same cluster are usually dependent and ignoring this dependence is likely to invalidate statistical inference. There is a vast literature on handling clustered data.[8] In short, one can either explicitly estimate the dependence structure with some additional specifications, such as random coefficient models, or account for the dependence after estimation, such as using cluster-robust variance estimator as discussed in Liang and Zeger (1986); Arellano (1987).

Cluster-robust variance estimators are very popular partly because they do not require assumptions on the dependence structure and partly because of its availability in most statistical softwares. Its validity is based on asymptotics when the number of clusters, G, grows to infinity, which is, unfortunately, not trivial to establish in nonparametric models. The main obstacle is that shrinking bandwidths is likely to destroy the dependence structure. For local polynomial regressions, Wang (2003) and Chen et al. (2008) point out that the existence of joint density of running variable and clustering variable ensures that cluster dependence vanishes asymptotically, not being captured by the usual approximations.[9]

Bartalotti and Brummet (2017) develop analytical approximations for the distribution and optimal bandwidth selector for the traditional RD estimator in a fixed-G setting, sidestepping the issue. Recently available software provides options to take this dependence into consideration. Both the *rdrobust* and *RDD* packages used in this paper offer the option to specify a clustering variable as explained by Calonico et al. (2017).

Naturally, a bootstrap approach could offer an alternative to the analytical approximations described. In fact, Cameron et al. (2008) provide a comprehensive survey of bootstrap methods and show that proper bootstrap procedures outperform the conventional cluster-robust variance estimator when the number of clusters is small (five to thirty).

To highlight the flexibility and robustness of the wild bootstrap procedure proposed in this paper, we revise the resampling algorithm to accommodate clustering and test its performance with pseudo clustered data. Following Brownstone and Valletta (2001) and Cameron et al. (2008), the wild bootstrap procedure for clustered data is quite straightforward: for units in the same cluster, their residuals are multiplied by the same random number drawn from the auxiliary distribution. For example,

$$Z^*_{gi} = \hat{g}_Z(X_{gi}) + \hat{\varepsilon}_{Z_{gi}} e^*_g,$$

where e^*_g, a random number from distribution with zero mean and unit variance, is shared by all units in the same group. For the purpose of simulation, it is assumed

[8] See Wooldridge (2003); Cameron et al. (2011); Cameron and Miller (2015) for an overview on this topic.

[9] A special case where this does not happen is that clustering occurs at the running variable level as discussed by Chen and Jin (2005).For example, in panel data where each individual are observed for multiple times and the running variable is at individual level, each individual is a cluster and will not vanish with shrinking bandwidth. Lee and Card (2008) consider another example in RD designs where clustering occurs at the running variable level and cluster-robust variance estimator is recommended for inference.

that errors in the outcome equation are clustered according to a random effect model, in particular, $u_{ygi} = u^*_{yg} + u^*_{yi}$ with $g = 1, 2, \ldots, G$ being a cluster indicator.[10]

Simulation results for $G = 5, 10, 25$ are reported in Table 5.1.[11] The wild bootstrap approach consistently outperforms the conventional method, closely matching the coverage from CCT's robust approach. This simple experiment shows that the wild bootstrap procedure proposed can also be easily applied to clustered data with slight adjustment to its resampling algorithm.

6 Empirical Illustration

In this section, we apply the bootstrap procedure to the data used in Angrist and Lavy (1999).[12] In that paper, the effects of class size on scholastic achievement are estimated using the "Maimonides' rule" as instrument.

As described by Angrist and Lavy (1999) the "Maimonides' rule" holds that the maximum class size is 40, and has been adopted by Israeli public schools to determine the division of enrollment cohorts into classes since 1969. Following this rule, when enrollment increases and passes multiples of 40, an additional class is required. Since the total enrollment is roughly evenly divided into all classes, an additional class causes a sudden drop in class sizes. Ideally, when the enrollment grows from 40 to 41, class size will drop by almost half. Because of student turnover and imperfect enforcement of this rule, the empirical data fits into a fuzzy RD design.

We consider the first discontinuity in class size for the 4^{th} grade. The sample used in this application includes 1164 classes from schools with enrollments no larger than 80. The outcome variables are average verbal and math test scores at class level. The discontinuities in class size and outcomes against enrollment are visualized in Figure 1. Each dot in these plots represents a class and the regression lines are fitted by fourth order polynomials. The shaded areas indicate CIs. The first plot clearly shows the discontinuity in class size, which is exploited for identification of the class size effect. The second plot suggests a discontinuity in average verbal score, but not as important as that in class size. The last plot does not provide much evidence for a discontinuity in average math score.

Three methods are applied to estimate the effect of class size on average verbal/math scores and results are shown in Table 6.1. The first column lists the original point estimates from local linear regression, which depends only on the bandwidth choice. The second column lists the bias-corrected point estimates based on bootstrap and analytical bias corrections. The estimates are very close to each other but differ meaningfully from the original estimates: the magnitude increases from 0.449 to 0.575 for average verbal score and 0.185 to 0.263∼0.272 for math score.

Consistent with what Figure 1 shows, only one out of three intervals for the treatment effect on average verbal score excludes zero and all three intervals for the treatment on average math score include zero. The interval from wild bootstrap is wider

[10] The design ensures that the individual errors have the same standard errors as the baseline case presented in Section B for easy comparison.

[11] G denotes the number of clusters on each side of the cutoff.

[12] The data is available at http://economics.mit.edu/faculty/angrist/data1/data/anglavy99.

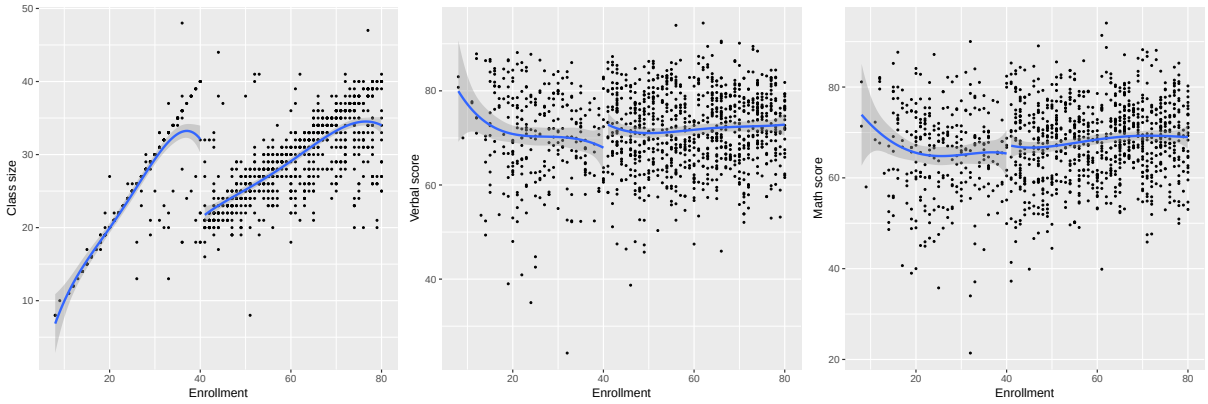

Figure 1: Class size, average verbal and math scores

than that from robust analytical approach, suggesting that it is more conservative, which is in line with the simulations presented.

7 Conclusion

A new wild bootstrap procedure is proposed to correct bias and construct valid CIs in fuzzy RD designs. This new method provides an easy to implement alternative to the analytical results in Calonico et al. (2014) to obtain robust bias-corrected CIs, and is implemented through a novel iterated bootstrap that extends the results and procedures described in Bartalotti et al. (2017). This new procedure is proved to be theoretically valid and empirically supported by simulation studies, performing as well as analytical alternatives, including in the presence of clustered data which has not been previously studied. An empirical illustration is provided, confirming the procedure's applied relevance.

Acknowledgements

This paper is derived from Yang He's work on his doctoral dissertation while at Iowa State University. The authors are grateful to Gray Calhoun for his invaluable insight, help, and support during the preparation of the original version of this manuscript. We are also thankful to Brent Kreider, Cindy Yu for valuable comments.

References

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics 114*, 533–75.

Arellano, M. (1987). Practitioners' corner: computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics 49*, 431–34.

Bartalotti, O. (2018). Regression discontinuity and heteroskedasticity robust standard errors: Evidence from a fixed-bandwidth approximation. *Journal of Econometric Methods 8*(1).

Bartalotti, O. and Q. Brummet (2017). Regression discontinuity designs with clustered data. In M. D. Cattaneo and J. C. Escanciano (Eds.), *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, Volume 38, pp. 383–420. Emerald Publishing Limited.

Bartalotti, O., G. Calhoun, and Y. He (2017). Bootstrap confidence intervals for sharp regression discontinuity designs. In M. D. Cattaneo and J. C. Escanciano (Eds.), *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, Volume 38, pp. 421–53. Emerald Publishing Limited.

Brownstone, D. and R. Valletta (2001). The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. *Journal of Economic Perspectives 15*(4), 129–41.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017). rdrobust: Software for regression-discontinuity designs. *Stata Journal 17*, 372–404.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*, 2295–326.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–27.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics 29*, 238–49.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*, 317–72.

Chen, K., J. Fan, and Z. Jin (2008). Design-adaptive minimax local linear regression for longitudinal/clustered data. *Statistica Sinica*, 515–34.

Chen, K. and Z. Jin (2005). Local polynomial regression analysis of clustered data. *Biometrika 92*, 59–74.

Chiang, H. D., Y.-C. Hsu, and Y. Sasaki (2017). A unified robust bootstrap method for sharp/fuzzy mean/quantile regression discontinuity/kink designs. *arXiv preprint arXiv:1702.04430*.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*, 162–69.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87*, 998–1004.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press, New York, NY.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis 49*, 361–76.

Gelman, A. and G. Imbens (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics 0*, 1–10.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*, 201–09.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies 79*, 933–59.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*, 615–35.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics 142*, 675–97.

Lee, D. S. and D. Card (2008). Regression discontinuity inference with specification error. *Journal of Econometrics 142*, 655–74.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *Quarterly Journal of Economics 122*, 159–208.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis*, pp. 437–61. Springer.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–25.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 255–85.

Porter, J. (2003). Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, 5–19.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika 90*, 43–52.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review 93*(2), 133–38.

Table 4.1: Empirical coverage and average interval length

DGP	Method	Bias	SD	RMSE	EC(%)	IL	h_{CCT}	b_{CCT}	h_{IK}
					Panel A: $\rho = 0$				
1	Wild bootstrap	0.015	0.054	0.056	93.1	0.197	0.197	0.323	
	CCT robust	0.015	0.054	0.056	91.5	0.191	0.197	0.323	
	Conventional	0.042	0.032	0.053	68.1	0.116			0.400
2	Wild bootstrap	0.037	0.058	0.069	86.9	0.210	0.165	0.299	
	CCT robust	0.039	0.060	0.071	86.6	0.212	0.165	0.299	
	Conventional	0.215	0.079	0.229	2.6	0.186			0.216
3	Wild bootstrap	0.005	0.053	0.053	95.3	0.205	0.162	0.317	
	CCT robust	0.005	0.053	0.054	94.1	0.200	0.162	0.317	
	Conventional	-0.025	0.044	0.050	87.3	0.157			0.205
					Panel B: $\rho = 0.9$				
1	Wild bootstrap	0.016	0.054	0.056	95.7	0.203	0.197	0.323	
	CCT robust	0.017	0.055	0.057	93.1	0.196	0.197	0.323	
	Conventional	0.043	0.033	0.054	70.7	0.121			0.398
2	Wild bootstrap	0.037	0.064	0.074	90.4	0.220	0.168	0.302	
	CCT robust	0.041	0.067	0.078	89.7	0.233	0.168	0.302	
	Conventional	0.226	0.092	0.244	3.0	0.207			0.222
3	Wild bootstrap	0.004	0.062	0.062	95.9	0.214	0.161	0.316	
	CCT robust	0.007	0.055	0.056	94.8	0.202	0.161	0.316	
	Conventional	-0.024	0.043	0.049	86.5	0.156			0.204
					Panel C: $\rho = -0.9$				
1	Wild bootstrap	0.015	0.053	0.056	91.3	0.198	0.199	0.324	
	CCT robust	0.013	0.055	0.056	91.1	0.190	0.199	0.324	
	Conventional	0.042	0.031	0.052	65.7	0.113			0.402
2	Wild bootstrap	0.037	0.052	0.064	85.5	0.205	0.161	0.296	
	CCT robust	0.038	0.052	0.064	84.4	0.190	0.161	0.296	
	Conventional	0.201	0.064	0.211	1.7	0.165			0.208
3	Wild bootstrap	0.005	0.053	0.053	95.6	0.206	0.163	0.317	
	CCT robust	0.003	0.054	0.054	94.5	0.203	0.163	0.317	
	Conventional	-0.027	0.045	0.052	89.1	0.160			0.207

Note: EC denotes empirical coverage and IL denote average interval length based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. Sample size is 1000. The triangular kernel is used. The columns h_{CCT} and b_{CCT} list average optimal bandwidths following CCT's method.The column h_{IK} lists average optimal bandwidth minimizing MSE. The bootstrap approach uses $B_1 = 500$ replications to compute bias and $B_2 = 999$ replications to obtain empirical distribution of bias-corrected estimator.

Table 5.1: Empirical coverage and average interval length (clustered data).

DGP	Method	Bias	SD	RMSE	EC(%)	IL	h_{CCT}	b_{CCT}	h_{IK}
					Panel A: $G = 5$				
1	Wild bootstrap	0.018	0.081	0.083	87.0	0.268	0.251	0.318	
	CCT robust	0.018	0.081	0.083	86.8	0.274	0.251	0.318	
	Conventional	0.043	0.071	0.083	83.7	0.249			0.392
2	Wild bootstrap	0.037	0.085	0.093	83.4	0.274	0.165	0.297	
	CCT robust	0.039	0.086	0.094	84.0	0.289	0.165	0.297	
	Conventional	0.214	0.101	0.237	22.5	0.275			0.216
3	Wild bootstrap	0.007	0.080	0.080	88.6	0.270	0.200	0.312	
	CCT robust	0.007	0.080	0.081	89.0	0.276	0.200	0.312	
	Conventional	-0.023	0.076	0.080	87.5	0.261			0.202
					Panel B: $G = 10$				
1	Wild bootstrap	0.017	0.068	0.070	90.2	0.240	0.230	0.321	
	CCT robust	0.018	0.068	0.071	88.9	0.236	0.230	0.321	
	Conventional	0.043	0.055	0.070	83.8	0.200			0.396
2	Wild bootstrap	0.036	0.071	0.079	87.1	0.250	0.166	0.299	
	CCT robust	0.038	0.071	0.081	86.2	0.253	0.166	0.299	
	Conventional	0.213	0.089	0.231	12.9	0.239			0.216
3	Wild bootstrap	0.005	0.067	0.067	92.7	0.243	0.186	0.316	
	CCT robust	0.005	0.068	0.068	91.5	0.240	0.186	0.316	
	Conventional	-0.025	0.062	0.067	88.0	0.220			0.204
					Panel C: $G = 25$				
1	Wild bootstrap	0.016	0.061	0.063	91.7	0.216	0.213	0.323	
	CCT robust	0.016	0.061	0.063	89.6	0.210	0.213	0.323	
	Conventional	0.043	0.043	0.060	78.7	0.157			0.399
2	Wild bootstrap	0.038	0.065	0.075	86.8	0.228	0.165	0.300	
	CCT robust	0.040	0.066	0.077	86.6	0.230	0.165	0.300	
	Conventional	0.214	0.084	0.230	6.4	0.210			0.216
3	Wild bootstrap	0.004	0.060	0.060	94.1	0.221	0.174	0.317	
	CCT robust	0.004	0.060	0.061	92.6	0.216	0.174	0.317	
	Conventional	-0.025	0.053	0.059	86.6	0.186			0.205

Note: EC denotes empirical coverage and IL denote average interval length based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. Sample size is 1000. The triangular kernel is used. The columns h_{CCT} and b_{CCT} list average optimal bandwidths following CCT's method.The column h_{IK} lists average optimal bandwidth minimizing MSE. The bootstrap approach uses $B_1 = 500$ replications to compute bias and $B_2 = 999$ replications to obtain empirical distribution of bias-corrected estimator.

Table 6.1: The effect of class size on average verbal score and average math score.

	ATE		95% CI		h_{CCT}	b_{CCT}	h_{IK}
	Original	Corrected					
	Panel A: Average verbal score						
Wild bootstrap	-0.449	-0.575	(-1.100	0.131)	12.391	18.278	
CCT robust	-0.449	-0.575	(-1.111	-0.040)	12.391	18.278	
Conventional	-0.488		(-1.104	0.129)			7.952
	Panel B: Average math score						
Wild bootstrap	-0.185	-0.263	(-0.924	0.466)	11.612	17.683	
CCT robust	-0.185	-0.272	(-0.884	0.340)	11.612	17.683	
Conventional	-0.202		(-0.802	0.398)			9.200

A Proofs

This appendix adopts Calonico et al. (2014), henceforth "CCT," notation where possible and utilizes some conclusions from that paper. Let e_p be the selection vector with 1 in element $p+1$ and 0 everywhere else and assume, with some abuse of notation, that the dimension of e_p adapts to make matrix and vector operations conformable. Much of the theory in this appendix applies to both sides of the cutoff symmetrically, so we use "$\bullet$" as a placeholder for either $+$ or $-$ in equations. Further let $r_p(x) = (1,\ x, \ldots,\ x^p)'$, $\mathbb{1}_+(x) = \mathbb{1}\{x \geq 0\}$, $\mathbb{1}_-(x) = \mathbb{1}\{x < 0\}$, $m = \min(h, b)$ and $\nu \leq p < q$. Define the following terms related to local polynomial regression:

$$\Gamma_{\bullet,p}(h) = \tfrac{1}{n}\sum_{i=1}^{n} r_p(X_i/h) r_p(X_i/h)' K_{\bullet,h}(X_i)$$

$$\Gamma_{\bullet,q}(b) = \tfrac{1}{n}\sum_{i=1}^{n} r_q(X_i/b) r_q(X_i/b)' K_{\bullet,b}(X_i)$$

$$\mathfrak{B}_{\bullet,\nu,p,q}(h) = \nu! e_\nu' \big(\Gamma_{\bullet,p}(h)\big)^{-1} \tfrac{1}{n}\sum_{i=1}^{n} (X_i/h)^q r_p(X_i/h) K_{\bullet,h}(X_i).$$

When $nh \to \infty$, $nm \to \infty$ and $h \to 0$, CCT's Lemma SA.1 and SA.2 imply that these terms have well-defined limits under Assumptions 2.1 and 2.2.

Let $\hat{\beta}_{Z\bullet,p}(h)$ be the coefficient estimators from the weighted regression of Z_i on $r_p(X_i)$:

$$\hat{\beta}_{Z\bullet,p}(h) = H_p(h)\Gamma_{\bullet,p}(h)^{-1} \tfrac{1}{n}\sum_{i=1}^{n} r_p(X_i/h) Z_i K_{\bullet,h}(X_i)$$

with $H_p(h) = diag(1, h^{-1}, \ldots, h^{-p})$. These coefficients are related to the quantities of interest by

$$\hat{\mu}^{(\nu)}_{Z\bullet,p}(h) = \nu! e_\nu' \hat{\beta}_{Z\bullet,p}(h)$$

and

$$\hat{\zeta}_{\nu,p}(h) = \frac{\hat{\mu}^{(\nu)}_{Y+,p}(h) - \hat{\mu}^{(\nu)}_{Y-,p}(h)}{\hat{\mu}^{(\nu)}_{T+,p}(h) - \hat{\mu}^{(\nu)}_{T-,p}(h)}$$

for $\nu = 0, \ldots, p$.

Proof of Theorem 3.1.

Based on the bias calculated from Algorithm 3.1, the difference between the bias-corrected estimator and the true treatment effect is

$$\hat{\zeta}_{\nu,p}(h) - \Delta^*_{\nu,p,q}(h,b) - \zeta_\nu = (\hat{\zeta}_{\nu,p}(h) - \zeta) - \left(\mathrm{E}^* \left[\frac{\hat{\tau}^*_{Y,\nu,p}(h)}{\hat{\tau}^*_{T,\nu,p}(h)} \right] - \frac{\tau^*_{Y,\nu}}{\tau^*_{T,\nu}} \right).$$

The first two terms on the right side can be written as

$$\begin{aligned}\hat{\zeta}_{\nu,p}(h)-\zeta_\nu =&\frac{1}{\tau_{T,\nu}}(\hat{\tau}_{Y,\nu,p}(h)-\tau_{Y,\nu})-\frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2}(\hat{\tau}_{T,\nu,p}(h)-\tau_{T,\nu})\\&+\frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2\hat{\tau}_{T,\nu,p}}(\hat{\tau}_{T,\nu,p}(h)-\tau_{T,\nu})^2-\frac{1}{\tau_{T,\nu}\hat{\tau}_{T,\nu,p}}(\hat{\tau}_{Y,\nu,p}(h)-\tau_{Y,\nu})(\hat{\tau}_{T,\nu,p}(h)-\tau_{T,\nu})\\=&\frac{1}{\tau_{T,\nu}}(\hat{\tau}_{Y,\nu,p}(h)-\tau_{Y,\nu})-\frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2}(\hat{\tau}_{T,\nu,p}(h)-\tau_{T,\nu})+R_n,\end{aligned}$$

with $R_n=O_p(\frac{1}{nh^{1+2\nu}}+h^{2(p+1-\nu)})$ (CCT's Lemma A.2). Similarly, the last two terms on the right side can be written as

$$\mathrm{E}^*\left[\frac{\hat{\tau}^*_{Y,\nu,p}(h)}{\hat{\tau}^*_{T,\nu,p}(h)}\right]-\frac{\tau^*_{Y,\nu}}{\tau^*_{T,\nu}}=\frac{1}{\tau^*_{T,\nu}}\Big(\mathrm{E}^*\left[\hat{\tau}^*_{Y,\nu,p}(h)\right]-\tau^*_{Y,\nu}\Big)-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}\Big(\mathrm{E}^*\left[\hat{\tau}^*_{T,\nu,p}(h)\right]-\tau^*_{T,\nu}\Big)+R^*_n,$$

with $R^*_n=O_p(\frac{1}{nh^{1+2\nu}}+h^{2(p+1-\nu)})$. By construction of the wild bootstrap DGP,

$$Z^*_i=\begin{cases}r_q(X_i/b)'H_q(b)^{-1}\beta^*_{Z+,q}+\varepsilon^*_i & X_i\geq 0\\ r_q(X_i/b)'H_q(b)^{-1}\beta^*_{Z-,q}+\varepsilon^*_i & X_i<0,\end{cases}$$

with $\beta^*_{Z+,q}$ and $\beta^*_{Z-,q}$ being the true parameters in the bootstrap data. Equivalently, $\mu^{*(\nu)}_{Z^*_\bullet}=\nu!e_\nu'\beta^*_{Z^*_\bullet,q}$ is the true treatment effect in the bootstrap data. CCT's Lemma SA.3 indicates that

$$\mathrm{E}^*\,\hat{\mu}^{*(\nu)}_{Z^*_\bullet,p}(h)-\mu^{*(\nu)}_{Z^*_\bullet}=h^{1+p-\nu}\mu^{*(1+p)}_{Z^*_\bullet}\mathfrak{B}_{\bullet,\nu,p,1+p}(h)/(1+p)!+O_p(h^{2+p-\nu}),$$

which allows for an analytical form of the bias in the bootstrap data:

$$\mathrm{E}^*\,\hat{\tau}^*_{Z,\nu,p}(h)-\tau^*_{Z,\nu}=h^{1+p-\nu}\big(\mu^{*(1+p)}_{Z+}\mathfrak{B}_{+,\nu,p,p+1}(h)-\mu^{*(1+p)}_{Z-}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!+O_p(h^{2+p-\nu}).$$

Notice that CCT's bias term is only slightly different from this. They use the following formula for bias correction:

$$\hat{\tau}^{bc}_{Z,\nu,p,q}(h,b)=\hat{\tau}_{Z,\nu,p}(h)-h^{1+p-\nu}\big(\hat{\mu}^{(1+p)}_{Z+,q}\mathfrak{B}_{+,\nu,p,p+1}(h)-\hat{\mu}^{(1+p)}_{Z-,q}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)!.$$

Built on above preparations, it can be shown that

$$\hat{\zeta}_{\nu,p}(h)-\Delta^*_{\nu,p,q}(h,b)-\zeta_\nu=\frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b)-\tau_{Y,\nu})-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b)-\tau_{T,\nu})+R_n-R^*_n-R^{*bc}_n+O_p(h^{2+p-\nu}),\tag{1.1}$$

where R_n^{*bc} is defined by:

$$\begin{aligned}
R_n^{*bc} =& \frac{1}{\tau_{T,\nu}^*} h^{1+p-\nu}\big(\mu_{Y+}^{*(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \mu_{Y-}^{*(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& - \frac{\tau_{Y,\nu}^*}{\tau_{T,\nu}^{*2}} h^{1+p-\nu}\big(\mu_{T+}^{*(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \mu_{T-}^{*(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& - \frac{1}{\tau_{T,\nu}} h^{1+p-\nu}\big(\hat{\mu}_{Y+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{Y-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& + \frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2} h^{1+p-\nu}\big(\hat{\mu}_{T+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{T-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
=& \frac{1}{\hat{\tau}_{T,\nu,q}(b)} h^{1+p-\nu}\big(\hat{\mu}_{Y+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{Y-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& - \frac{\hat{\tau}_{Y,\nu,q}(b)}{\hat{\tau}_{T,\nu,q}^2(b)} h^{1+p-\nu}\big(\hat{\mu}_{T+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{T-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& - \frac{1}{\tau_{T,\nu}} h^{1+p-\nu}\big(\hat{\mu}_{Y+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{Y-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& + \frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2} h^{1+p-\nu}\big(\hat{\mu}_{T+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{T-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
=& \Big(\frac{1}{\hat{\tau}_{T,\nu,q}(b)} - \frac{1}{\tau_{T,\nu}}\Big) h^{1+p-\nu}\big(\hat{\mu}_{Y+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{Y-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
& - \Big(\frac{\hat{\tau}_{Y,\nu,q}(b)}{\hat{\tau}_{T,\nu,q}^2(b)} - \frac{\tau_{Y,\nu}}{\tau_{T,\nu}^2}\Big) h^{1+p-\nu}\big(\hat{\mu}_{T+,q}^{(1+p)}\mathfrak{B}_{+,\nu,p,p+1}(h) - \hat{\mu}_{T-,q}^{(1+p)}\mathfrak{B}_{-,\nu,p,p+1}(h)\big)/(1+p)! \\
=& h^{1+p-\nu} O_p\Big(\frac{1}{\sqrt{nb^{1+2\nu}}} + b^{1+q-\nu}\Big) O_p\Big(1 + \frac{1}{\sqrt{nb^{3+2p}}}\Big).
\end{aligned}$$

The second equality holds because $\mu_{Z\bullet}^{*(1+p)} = \hat{\mu}_{Z\bullet,q}^{(1+p)}(b)$ and $\tau_{Z,\nu}^* = \hat{\tau}_{Z,\nu,q}(b)$ almost surely because the bootstrap DGP is obtained by fitting a local polynomials of order q. The last equality holds because of similar argument in CCT's Theorem A.2. Asymptotic normality of $\hat{\zeta}_{\nu,p}(h) - \Delta_{\nu,p,q}^*(h,b) - \zeta_\nu$ then follows from normality of $\hat{\tau}_{Y,\nu,p,q}^{bc}(h,b) - \tau_{Y,\nu}$, $\hat{\tau}_{T,\nu,p,q}^{bc}(h,b) - \tau_{T,\nu}$ (CCT's Theorem 1) and the fact that remaining terms R_n, R_n^*, R_n^{*bc} and $O_p(h^{2+p-\nu})$ are negligible.

CCT have shown that $V^{bc}(h,b) = O_p(\frac{1}{nh^{1+2\nu}} + \frac{h^{2(1+p-\nu)}}{nb^{3+2p}})$ (Lemma SA.4) and $R_n^2 = o_p(V^{bc}(h,b))$ (Theorem A.2). In addition, because $O_p(h^{2+p-\nu}) = o_p(R_n^{*bc})$, it suffices

to show that

$$\begin{aligned}
\frac{R_n^{*bc2}}{V^{bc}(h,b)} =& O_p\big(\min\{nh^{1+\nu}, \frac{nb^{3+2p}}{h^{2(1+p-\nu)}}\}\big)h^{2(1+p-\nu)}O_p\big(\frac{1}{nb^{1+2\nu}} + b^{2(1+q-\nu)}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(\min\{nh^{3+2p}, nb^{3+2p}\}\big)O_p\big(\frac{1}{nb^{1+2\nu}} + b^{2(1+q-\nu)}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(b^{2+2(p-\nu)}\min\{(\frac{h}{b})^{3+2p},1\} + nb^{2(1+q-\nu)}\min\{nh^{3+2p}, nb^{3+2p}\}\big)O_p\big(1+\frac{1}{nb^{3+2p}}\big)\\
=& O_p\big(b^{2+2(p-\nu)}\min\{(\frac{h}{b})^{3+2p},1\} + nb^{2(q-p)}b^{2(1+p-\nu)}\min\{nh^{3+2p}, nb^{3+2p}\}\big)\\
&+ O_p\bigg(\frac{1}{nb^{1+2\nu}}\min\{(\frac{h}{b})^{3+2p},1\} + b^{2(1+q-\nu)}\min\{(\frac{h}{b})^{3+2p},1\}\bigg)\\
=& o_p(1),
\end{aligned}$$

provided that $n\min\{h^{3+2p}, b^{3+2p}\}\max\{h^2, b^{2(q-p)}\} \to 0$ and $n\min\{h, b^{1+2\nu}\} \to \infty$. □

.

Proof of Theorem 3.2.

Repeat the steps from the proof for Theorem 3.1's proof for the iterated bootstrap to get

$$\hat{\zeta}^*_{\nu,p}(h) - \Delta^{**}_{\nu,p,q}(h,b) - \zeta^*_\nu = \frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b) - \tau^*_{T,\nu}) + R^*_n - R^{**}_n - R^{**bc}_n + O_p(h^{2+p-\nu}),$$

As is proved in previous section, the higher order terms do not contribute to its asymptotic variance and can be ignored. It will be firstly shown that the variance of $\frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b) - \tau^*_{Y,\nu}) - \frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b) - \tau^*_{T,\nu})$ converges to that of $\frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b) - \tau_{Y,\nu}) - \frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b) - \tau_{T,\nu})$, then its asymptotic normality will be proved.

For the variance convergence in probability result, rewrite bias-corrected estimator

for Z:

$$\begin{aligned}
\hat{\tau}^{bc}_{Z,\nu,p,q}(h,b)-\tau_{Z,\nu} =&(\hat{\tau}_{Z,\nu,p}(h)-\mathrm{E}\,\hat{\tau}_{Z,\nu,p}(h))+(\mathrm{E}\,\hat{\tau}_{Z,\nu,p}(h)-\tau_{Z,\nu})-(\mathrm{E}^*\,\hat{\tau}^*_{Z,\nu,p}(h)-\tau^*_{Z,\nu})\\
=&\hat{\tau}_{Z,\nu,p}(h)-\mathrm{E}\,\hat{\tau}_{Z,\nu,p}(h)\\
&+h^{1+p-\nu}(\hat{\mu}^{(q)}_{Z-,q}(b)-\mu^{(q)}_{Z-})\mathfrak{B}_{-,\nu,p,p+1}(h)/(1+p)!\\
&-h^{1+p-\nu}(\hat{\mu}^{(q)}_{Z+,q}(b)-\mu^{(q)}_{Z+})\mathfrak{B}_{+,\nu,p,p+1}(h)/(1+p)!\\
&+O_p(h^{2+p-\nu})\\
=&\nu!e_\nu'\Gamma_{+,p}(h)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_p(X_i/h)K_{+,h}(X_i)\varepsilon_{Zi}\Big)\\
&-\nu!e_\nu'\Gamma_{-,p}(h)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_p(X_i/h)K_{-,h}(X_i)\varepsilon_{Zi}\Big)\\
&+\frac{q!e_q'h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{-,q}(b)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_q(X_i/b)K_{-,b}(X_i)\varepsilon_{Zi}\Big)\mathfrak{B}_{-,\nu,p,p+1}(h)\\
&-\frac{q!e_q'h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{+,q}(b)^{-1}\Big(\tfrac{1}{n}\sum_{i=1}^n r_q(X_i/b)K_{+,b}(X_i)\varepsilon_{Zi}\Big)\mathfrak{B}_{+,\nu,p,p+1}(h)\\
&+O_p(h^{2+p-\nu})\\
=&\sum_{i=1}^n W(X_i)\varepsilon_{Zi}+O_p(h^{2+p-\nu})
\end{aligned}$$

with

$$W(X_i)=W_+(X_i)-W_-(X_i)$$

$$W_\bullet(X_i)=\tfrac{1}{n}\nu!e_\nu'\Gamma_{\bullet,p}(h)^{-1}r_p(X_i/h)K_{\bullet,h}(X_i)-\tfrac{1}{n}\frac{q!e_q'h^{1+p-\nu}}{(1+p)!b^q}\Gamma_{\bullet,q}(b)^{-1}r_q(X_i/b)K_{\bullet,b}(X_i).$$

With this simplified notation, we have

$$\frac{1}{\tau_{T,\nu}}(\hat{\tau}^{bc}_{Y,\nu,p,q}(h,b)-\tau_{Y,\nu})-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}(\hat{\tau}^{bc}_{T,\nu,p,q}(h,b)-\tau_{T,\nu})=\sum_{i=1}^n W(X_i)\left(\frac{1}{\tau_{T,\nu}}\varepsilon_{Yi}-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}\varepsilon_{Ti}\right)+O_p(h^{2+p-\nu}),$$

which has variance

$$\mathbb{V}\left(\sum_{i=1}^n W(X_i)\left(\frac{1}{\tau_{T,\nu}}\varepsilon_{Yi}-\frac{\tau_{Y,\nu}}{\tau^2_{T,\nu}}\varepsilon_{Ti}\right)\right)=\sum_{i=1}^n W(X_i)^2\left(\frac{1}{\tau^2_{T,\nu}}\sigma^2_{Yi}+\frac{\tau^2_{Y,\nu}}{\tau^4_{T,\nu}}\sigma^2_{Ti}-\frac{2\tau_{Y,\nu}}{\tau^3_{T,\nu}}\sigma_{Yi,Ti}\right).$$

Apply similar steps to the iterated bootstrap, we have

$$\frac{1}{\tau^*_{T,\nu}}(\hat{\tau}^{*bc}_{Y,\nu,p,q}(h,b)-\tau^*_{Y,\nu})-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}(\hat{\tau}^{*bc}_{T,\nu,p,q}(h,b)-\tau^*_{T,\nu})=\sum_{i=1}^n W(X_i)\left(\frac{1}{\tau^*_{T,\nu}}\varepsilon^*_{Yi}-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}\varepsilon^*_{Ti}\right),$$

which, by the construction of wild bootstrap, has variance

$$\mathbb{V}^*\left(\sum_{i=1}^n W(X_i)\left(\frac{1}{\tau^*_{T,\nu}}\varepsilon^*_{Yi}-\frac{\tau^*_{Y,\nu}}{\tau^{*2}_{T,\nu}}\varepsilon^*_{Ti}\right)\right)=\sum_{i=1}^n W(X_i)^2\left(\frac{1}{\tau^{*2}_{T,\nu}}\hat{\varepsilon}^2_{Yi}+\frac{\tau^{*2}_{Y,\nu}}{\tau^{*4}_{T,\nu}}\hat{\varepsilon}^2_{Ti}-\frac{2\tau^*_{Y,\nu}}{\tau^{*3}_{T,\nu}}\hat{\varepsilon}_{Yi}\hat{\varepsilon}_{Ti}\right).$$

By the standard argument on the convergence of residuals to the population error, it is ensured that $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}_{Yi}^2 \to^p \sum_{i=1}^n W(X_i)^2\sigma_{Yi}^2$, $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}_{Ti}^2 \to^p \sum_{i=1}^n W(X_i)^2\sigma_{Ti}^2$ and $\sum_{i=1}^n W(X_i)^2\hat{\varepsilon}_{Yi}\hat{\varepsilon}_{Ti} \to^p \sum_{i=1}^n W(X_i)^2\sigma_{Yi,T_i}$. Combined with the fact that $\tau^*_{Z,\nu} = \hat{\tau}_{Z,q}(b) \to^p \tau_Z$, the proof for convergence of variance is complete.

Finally, for concluding the asymptotic normality argument, note that, conditional on the regressors and residuals, $\{W(X_i)(\frac{1}{\tau_T^*}\hat{\varepsilon}_{Yi} - \frac{\tau_Y^*}{\tau_T^{*2}}\hat{\varepsilon}_{Ti})e_i^*\}$ is a sequence of independent and mean zero random variables. In addition, it consists of four parts based on the definition of $W(X_i)$. It can be shown that each part is asymptotically normal by Lindeberg-Feller CLT. The proof below is an example showing that the first part $\frac{1}{n}\nu!e_\nu'\Gamma_{\bullet,p}(h)^{-1}r_p(X_i/h)K_{\bullet,h}(X_i)\Big(\frac{1}{\tau_T^*}\hat{\varepsilon}_{Yi} - \frac{\tau_Y^*}{\tau_T^{*2}}\hat{\varepsilon}_{Ti}\Big)e_i^*$ is asymptotically normal.

The Liapunov's condition requires that

$$\frac{1}{s_n^{2+\delta}}\sum_{i=1}^n \mathrm{E}|H_i(Xi)|^{2+\delta} \to^p 0$$

with

$$H_i(X_i) = \tfrac{1}{n}\nu!e_\nu'\Gamma_{\bullet,p}(h)^{-1}r_p(X_i/h)K_{\bullet,h}(X_i)\left(\frac{1}{\tau_T^*}\hat{\varepsilon}_{Yi} - \frac{\tau_Y^*}{\tau_T^{*2}}\hat{\varepsilon}_{Ti}\right)e_i^*; \quad s_n^2 = \sum_{i=1}^n \mathbb{V}(H_i).$$

Based on CCT's Lemma SA.1, we know that

$$\sum_{i=1}^n \mathrm{E}|H_i(Xi)|^{2+\delta} = O_p\left(\frac{1}{(nh)^{1+\delta}}\right),$$
$$s_n^2 = O_p\left(\frac{1}{nh}\right),$$

which verifies the Liapunov's condition given that $nh \to \infty$. Similar arguments can be applied to other three parts. □

B Monte Carlo Simulations: Additional Details

The proposed bootstrap algorithms are applied to a variety of data generating processes (DGP). The baseline DGP is similar to CCT but re-designed to fit the fuzzy RD designs:

$$\begin{aligned}
&X_i \sim 2 \times beta(2,4) - 1\\
&T_i = \mathbb{1}\{u_{ti} \le \Phi^{-1}(0.5 - \frac{c}{2})\}\,\mathbb{1}\{X_i < 0\} + \mathbb{1}\{u_{ti} \le \Phi^{-1}(0.5 + \frac{c}{2})\}\,\mathbb{1}\{X_i \ge 0\}\\
&Y_i = \mu_j(X_i) + T_i\zeta_j + u_{yi},
\end{aligned}$$

where $u_{ti} \sim N(0,1)$ and $c = 0.9$. The equation for T_i indicates that $\mu_{T-} = 0.5 - c/2$ and $\mu_{T+} = 0.5 + c/2$. As a result, the expected treatment conditional on running variable is constant on both sides but the discontinuity at the cutoff is exactly c. In the equation for Y_i, the first term, $\mu_j(X_i)$ with $j = 1, 2, 3$, is the conditional expected

outcome without treatment, which is continuous at the cutoff. The second term, $T_i\zeta_j$, captures the additive treatment effect. In particular, the conditional expected outcome takes the following forms for each DGP:

$$\mu_1(x) = \begin{cases} 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise,} \end{cases}$$

$$\mu_2(x) = \begin{cases} 2.30x + 3.28x^2 + 1.45x^3 + 0.23x^4 + 0.03x^5 & \text{if } x < 0, \\ 18.49x - 54.81x^2 + 74.30x^3 - 45.02x^4 + 9.83x^5 & \text{otherwise,} \end{cases}$$

$$\mu_3(x) = \begin{cases} 1.27x + 3.59x^2 + 14.147x^3 + 23.694x^4 + 10.995x^5 & \text{if } x < 0 \\ 0.84x - 0.30x^2 + 2.397x^3 - 0.901x^4 + 3.56x^5 & \text{otherwise.} \end{cases}$$

These conditional mean functions are adapted from DGPs used by CCT for sharp RD designs by preserving the curvature but removing the discontinuity at the cutoff. The first mean function is designed to match features of U.S. congressional election data from Lee (2008). The second mean function is designed to match the relation between children mortality rate and county poverty rate from analysis of Head Start data in Ludwig and Miller (2007). The last mean function is similar to the first one except for some coefficients. CCT motivates this in an attempt to generate plausible model with sizable distortion when conventional t-test is performed. The true treatment effects for these three models are $\zeta_1 = 0.04, \zeta_2 = -3.45, \zeta_3 = 0.04$.

To accommodate the variety of different error structures in empirical data, the following cases are considered.

1. Baseline case. The simplest case where errors are i.i.d.:

$$\begin{pmatrix} u_{ti} \\ u^*_{yi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad \rho = 0, \quad u_{yi} = 0.1295 u^*_{yi}.$$

2. Endogeneity. The treatment status is correlated with unobserved characteristics which affect the outcome. This is modelled by the correlation between u_{ti} and u_{yi}, i.e., everything being the same as in the baseline case except $\rho \in \{-0.9, 0.9\}$.

In the implementation of Algorithm 3.1 and 3.2, the two-point distribution proposed in Mammen (1993) is adopted for creating bootstrap samples. This auxiliary distribution is

$$e^*_i = \begin{cases} \frac{1+\sqrt{5}}{2} & \text{with probability } \frac{\sqrt{5}-1}{2\sqrt{5}}, \\ \frac{1-\sqrt{5}}{2} & \text{otherwise,} \end{cases}$$

with zero mean and unit second and third moments. Its properties ensures that the skewness of the bootstrap error terms is the same as the skewness of the residuals, which is a desirable condition not imposed in Algorithm 3.1 and 3.2.[13] In addition, the residuals are transformed before applying the bootstrap because they are on average underestimated by least squares. Specifically, instead of directly using $\hat{\varepsilon}_{Zi}$, the

[13] Some later studies also show good properties of the simpler Rademacher distribution, see Flachaire (2005); Davidson and Flachaire (2008).

"HC3" type transformation $\hat{\varepsilon}_{Zi}/(1-H_{ii})$ is applied, with H_{ii} being the diagonal element of projection matrix.[14] This is based on jackknife covariance estimator and is shown to outperform the original heteroskedasticity-robust covariance estimator in MacKinnon and White (1985). Simulation studies by Davidson and Flachaire (2008) and MacKinnon (2013) also provide some evidence in favor of "HC3" transformation.

Besides the bootstrap approach, two additional approaches are estimated for comparison: CCT's robust estimator and the conventional estimator. All simulations are conducted with R software. Packages *rdrobust (V0.94)* and *RDD (V0.57)* are used to estimate the CCT's robust estimator and conventional RD estimator respectively. By default, the former one uses the nearest neighbour variance estimator and the latter one uses "HC1" type heteroskedasticity-robust variance estimator. The two bandwidths for the bootstrap and CCT's approaches are the same and are obtained by utilizing bandwidth selectors from CCT. The bandwidth used in the conventional approach is the MSE-optimal bandwidth proposed by Imbens and Kalyanaraman (2012).[15] These three approaches are applied to a total number of 5000 simulated samples with a sample size of 1000. Triangular kernel is used throughout all the simulations.[16]

The simulation results are presented in Table 4.1 in the main text. Part of the discussion about the results is repeated below, with additional details.

The baseline case is listed in Panel A. The two robust methods, wild bootstrap and CCT's approach, generate point estimates with very similar bias and standard error (identical for DGP 1 and 3 and slightly different for DGP 2). In contrast, the conventional approach reports 3-5 times larger bias. This is not surprising since the two robust methods explicitly correct the bias. The conventional method also fails to deliver a valid CI (coverage rates are 68.1%, 2.6% and 87.2% for the three DGPs respectively). Robust methods achieve improvements by reducing bias and increasing interval length. Except for DGP 2, they both generate intervals with empirical coverage close to the nominal level and the wild bootstrap performs well for DGPs 1 and 3. However, for DGP 2, even the robust methods report significant size distortion. This is because DGP 2 has great curvature around the cutoff and makes precise fitting very difficult.In particular, the DGP 2 shows great curvature just right to the cutoff. On the right side, its second derivative at the cutoff is -109.62, so local linear regression is likely to create large bias. Its third derivative at the cutoff is 445.8, so local quadratic regression is likely to create large bias. Still, the two robust methods improve significantly from the conventional method in coverage (from 2.6% to around 87%) at the sacrifice of slightly longer intervals (from 0.186 to around 0.21).

Panels B and C present results when the treatment is endogenous, which is almost always true and probably the primary reason to choose RD designs as the identification strategy. The case with positive (negative) ρ is listed in Panel B(C). Again, the estimate from conventional method has significantly larger bias than the other two robust methods. As for interval estimates, the wild bootstrap and CCT's approach work

[14] Local regressions project $\mathbf{K}^{1/2}\mathbf{Y}$ onto space of $\mathbf{K}^{1/2}\mathbf{X}$, with $\mathbf{K}$ being the weighting matrix determined by kernel function. So the projection matrix will be $\mathbf{K}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}^{1/2}$.

[15] As is suggested by Imbens and Kalyanaraman (2012), the optimal bandwidth choices in fuzzy RD designs are often similar to those based on the optimal bandwidth for the numerator only. For simplicity, all bandwidths are calculated ignoring the fact that the RD design is fuzzy.

[16] Results with other kernel functions are similar.

reasonably well in all cases. The conventional method performs significantly worse, with empirical coverage rate as low as 1.7% (DGP 2 with negative self-selection). The sign of correlation has little effect on the bias because the bias is caused by model misspecification rather than imperfect instrumental variable.