Bootstrap Confidence Intervals for Sharp Regression Discontinuity Designs with the Uniform Kernel

Otávio Bartalotti Gray Calhoun Yang He*

May 1, 2016

Abstract

This paper develops a novel bootstrap procedure to obtain robust bias-corrected confidence intervals in regression discontinuity (RD) designs using the uniform kernel. The procedure uses a residual bootstrap from a second order local polynomial to estimate the bias of the local linear RD estimator; the bias is then subtracted from the original estimator. The bias-corrected estimator is then bootstrapped itself to generate valid confidence intervals. The confidence intervals generated by this procedure are valid under conditions similar to Calonico, Cattaneo and Titiunik's (2014, *Econometrica*) analytical correction—i.e. when the bias of the naive regression discontinuity estimator would otherwise prevent valid inference. This paper also provides simulation evidence that our method is as accurate as the analytical corrections and we demonstrate its use through a reanalysis of Ludwig and Miller's (2008) Head Start dataset.

1 Introduction

Regression Discontinuity (RD) designs have emerged in the last decade as an important and popular research design strategy for analyzing the causal impact of policies and interventions in several fields of the social sciences, including economics, political science, public policy, and sociology. This research strategy exploits the fact that many programs use a threshold based on a numeric score to determine whether or not to provide a treatment.[1] In its basic version, *sharp RD*, individuals or groups with score above the threshold are treated while those below the threshold are left untreated. The identification of the treatment effect at the threshold is then based on comparing

*All authors: Department of Economics, Iowa State University. 260 Heady Hall, Ames, IA 50011. Bartalotti: bartalot@iastate.edu; Calhoun: gcalhoun@iastate.edu and http://gray.clhn.org; He: yanghe@iastate.edu.

[1]This score is often referred to as the *running variable* in this literature.

treated and untreated units at the cutoff. When a subject's position just above or below the cutoff is credibly not related to unobserved characteristics that would affect the outcome of interest, differences between treated and untreated individuals at the cutoff can be plausibly attributed to the treatment alone. As a practical matter this involves comparing units within a bandwidth just above and just below the threshold. Other RD strategies exist that can exploit different forms of discontinuities as well.

The RD design strategy was introduced by Thistlethwaite and Campbell (1960) to study educational outcomes and many of its recent applications in economics were to estimate the effects of other educational policies: evaluating the impact of investments in school facilities, class sizes, remedial education, early childhood education, and financial aid effects on student achievement and later outcomes, for example.[2] But the underlying identification strategy has proven to apply much more widely and RD has been used in health economics,[3] political science,[4] and labor economics,[5] among other fields. Imbens and Lemieux (2008) and Lee and Lemieux (2010) provide recent overviews of this literature with many more examples.

In these studies, identification occurs exactly at the cutoff, so the treatment effect is typically estimated by fitting separate local linear models above and below the cutoff, then extrapolating the models to the exact point of discontinuity. The difference between the estimated outcomes at that point is taken to be an estimate of the treatment effect. As a practical matter, a key econometric issue is determining the bandwidth for the local linear models. One very popular choice is the bandwidth estimator proposed by Imbens and Kalyanaraman (2012) and extended by Calonico, Cattaneo, and Titiunik (2014), which minimizes the Asymptotic Mean Squared Error (AMSE) of difference in the models' point estimators at the cutoff. But, as observed by Calonico, Cattaneo, and Titiunik (2014), henceforth "CCT," the AMSE-optimal bandwidth has the serious drawback that it produces invalid confidence intervals and hypothesis tests. Local polynomial estimators of the treatment effect are generally biased in finite samples because the functional form of the local conditional expectation that they need to approximate is unknown. The unmodeled component of the conditional expectation becomes smaller as the bandwidth itself becomes smaller, so the estimator's bias vanishes asymptotically as long as the bandwidth shrinks as the sample size increases. AMSE-optimal bandwidth shrinks as the sample size increases, so its estimator of the treatment effect is consistent, but the bandwidth shrinks slowly enough that the remaining bias term is large enough to affect the asymptotic distribution of the estimator. Consequently the usual "naive" con-

[2]See, for example, Van der Klaauw (2002), Jacob and Lefgren (2004), Ludwig and Miller (2007), Urquiola and Verhoogen (2009), Cellini, Ferreira, and Rothstein (2010)

[3]Card, Dobkin, and Maestas (2009); Barreca et al. (2011)

[4]Lee and Card (2008), Caughey and Sekhon (2011), Keele and Titiunik (2014), Erikson and Titiunik (2015), Fujiwara (2011, 2015)

[5]Schmieder, Von Wachter, and Bender (2012).

fidence intervals for the RD treatment effects are invalid and can have coverage well below their nominal level.

CCT show that the bias resulting from undersmoothing can be estimated and they provide a bias-corrected treatment effect estimator that remains asymptotically unbiased even when the bandwidth converges to zero at the AMSE-optimal rate. They also show that this bias-correction term contributes to the asymptotic variance of the resulting treatment effect estimator and provide a new formula for the asymptotic variance of the bias-corrected estimator. The resulting confidence intervals have accurate coverage even when the naive RD interval does not.

In this paper, we propose a bootstrap alternative to CCT's analytical corrections. CCT motivate their estimator by showing that the bias and variance components for the local linear estimator can be accounted for by estimating a local second order polynomial with bandwidth of the same order.[6] They use a Taylor expansion around the cutoff to show that the bias associated with the second order polynomial converges to zero at a faster rate, fast enough that the bias of the local linear model can be estimated and removed using the second order polynomial. Additionally, that approximation provides fast enough convergence that it can be used to estimate the correct variance correction as well.

Our approach exploits CCT's theoretical insight through a new residual bootstrap. In particular, we propose estimating the local linear model as usual, then estimating a local second order polynomial and generating bootstrap datasets by resampling the residuals of that polynomial. Since the second order polynomial is the true Data Generating Process (DGP) for the bootstrapped data, its estimate of the treatment effect is the true value of the treatment effect under the distribution induced by this bootstrap. The bias of the linear model is therefore known under this distribution and can be calculated by averaging the error of the linear model's estimates across many bootstrap replications. This approach is described in detail by our Algorithm 1 and the resulting bias corrected estimator is shown to be asymptotically normal with mean zero in our Theorem 1 under AMSE-optimal bandwidth rates.

Just as in CCT, our bias correction step introduces additional variability. However, the second order polynomial again adequately estimates the features of the true DGP that are necessary for estimating and accommodating that additional variability. So we propose an iterated bootstrap procedure (Hall and Martin, 1988): use the second order polynomial residual bootstrap to produce many bootstrap replications of the bias corrected estimator, and then use the resulting bootstrap distribution to produce confidence intervals. This procedure, which requires bootstrapping the datasets produced by an initial bootstrap, is described in Algorithm 2, and the resulting confidence intervals

[6]More generally, they show that the bias and variance of a local polynomial of order p can be accounted for by estimating the $p + 1$ local polynomial. We will restrict our analysis to the case with $p = 1$ in this paper because of its widespread use.

are shown to be asymptotically valid in Theorem 2.

This bootstrap procedure offers some advantages over analytical methods. In particular, both this paper and CCT assume that the observations are generated independently of each other; however, extending these bias correction methods to other forms of dependence is relatively straightforward for the bootstrap but can be substantially more complicated for analytical corrections, which have to be explicitly derived by the researcher. Similarly, although we only provide results for the local linear model in this paper, it is trivial to implement this procedure for higher order polynomials, for covariate-adjusted estimators (Frölich, 2007, Calonico et al., 2015), or for other local smoothers. (Loader, 2006) However, in this paper we focus on the baseline case of sharp RD with a local linear model and uniform kernel. Extensions to fuzzy RD designs[7] and nonuniform kernels, which require nontrivial changes to the underlying bootstrap algorithm, as well as developments to address cross sectional dependence, are the subject of ongoing research.

The paper is organized as follows. Section 2 describes the basic RD approach, its usual implementation, and the explicit analytical bias correction approach in the literature. Section 3 presents our proposed bootstrap bias corrected RD algorithm and discusses its asymptotic properties. Simulation evidence that the bootstrap procedure provides valid CIs and its relative performance to the analytical bias correction are presented in Section 4 and Section 5 demonstrates the estimator's usage by applying it to the Head Start dataset used by Ludwig and Miller (2007).[8] Finally, Section 6 concludes.

2 Background

This section provides additional details of RD estimators in general and of CCT's proposed bias correction. It also defines some of the notation and presents the assumptions that will be used for our theoretical analysis in Section 3. We have adopted CCT's notation where possible to aid readers familiar with that paper.

In the typical sharp RD setting, a researcher wishes to estimate the local causal effect of treatment at a given threshold. A running variable, X_i, determines treatment assignment. Given a known threshold, which we will set to zero without loss of generality, the ith subject receives the treatment of interest if $X_i \geq 0$ and does not receive treatment if $X_i < 0$.

[7] "Fuzzy" regression discontinuity design (Hahn, Todd, and Van der Klaauw, 2001), as opposed to "sharp" RD, describes situations where the probability of treatment changes discontinuously at a known threshold, but by less than 100%. Then there are treated and untreated subjects above and below the discontinuity but the treatment effect remains identified.

[8] The simulations and empirical analysis were carried out in the R programming language (R Core Team, 2015) and rely on the *rdrobust* (Calonico, Cattaneo, and Titiunik, 2015), *doParallel*, *foreach* (Revolution Analytics and Weston, 2015a,b), and *doRNG* (Gaujoux, 2014) packages.

Subject i's potential outcomes are denoted by the variable $Y_i(\cdot)$; $Y_i(1)$ is the subject's outcome under treatment and $Y_i(0)$ is the outcome without treatment. Since only one of the two outcomes is observed, the sample is comprised of the running variable, X_i, and the observed outcome Y_i, where

$$Y_i = Y_i(0)\,\mathbb{1}\{X_i < 0\} + Y_i(1)\,\mathbb{1}\{X_i \geq 0\}$$

and $\mathbb{1}\{\cdot\}$ denotes the indicator function.

In most cases, the population parameter of interest is the Average Treatment Effect (ATE) at the cutoff, which we will denote τ. This parameter is the difference in expected potential outcomes given $X_i = 0$; formally,

$$\tau = \mathbb{E}(Y(1) - Y(0) \mid X = 0).$$

Hahn, Todd, and Van der Klaauw (2001) show that the effect τ is identified under continuity and smoothness conditions on the joint distribution of X_i, $Y_i(0)$, and $Y_i(1)$ around the cutoff $X_i = 0$. Under these conditions, which are made precise in our Assumption 1, τ is equal to

$$\tau = \lim_{x \to 0+} \mu(x) - \lim_{x \to 0-} \mu(x)$$

where

$$\mu(x) = \mathbb{E}(Y_i \mid X_i = x).$$

For later convenience, also define the derivatives

$$\mu^{(\eta)}(x) = \frac{d^\eta \mu(x)}{dx^\eta}$$

and let

$$\mu_+(x) = \mathbb{E}(Y_i(1) \mid X_i = x) \qquad \mu_-(x) = \mathbb{E}(Y_i(0) \mid X_i = x)$$
$$\sigma^2_+(x) = \mathbb{V}(Y_i(1) \mid X_i = x) \qquad \sigma^2_-(x) = \mathbb{V}(Y_i(0) \mid X_i = x)$$

and

$$\mu^{(\eta)}_+ = \lim_{x \to 0^+} \mu^{(\eta)}(x), \qquad \mu^{(\eta)}_- = \lim_{x \to 0^-} \mu^{(\eta)}(x),$$

where the symbol $\mathbb{V}(\cdot)$ represents the variance. The effect τ is nonparametrically identified because both μ_- and μ_+ can be estimated consistently under Assumption 1, which lists standard conditions

in the RD literature. (See, in particular, Hahn, Todd, and Van der Klaauw, 2001, Porter, 2003, and CCT.)

Assumption 1 (Behavior of the DGP near the cutoff)**.** *The random variables Y_i, X_i 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 neighborhood $(-\kappa_0, \kappa_0)$ around zero:*

1. *The density of each X_i is continuous and bounded away from zero.*
2. *$\mathbb{E}(Y_i^4 \mid X_i = x)$ is bounded.*
3. *$\mu_+(x)$ and $\mu_-(x)$ are both 3 times continuously differentiable.*
4. *$\sigma_+^2(x)$ and $\sigma_-^2(x)$ are both continuous and bounded away from zero.*

Since the conditions for identification only need to hold in a neighborhood around the cutoff, μ_+ and μ_- can be estimated by extrapolating from a local polynomial regression. We will focus here on local linear regression.[9] For this model, if h represents a bandwidth parameter, the estimator of τ, $\hat{\tau}(h)$, is defined as

$$\hat{\tau}(h) = \hat{\mu}_+(h) - \hat{\mu}_-(h)$$

with

$$\hat{\mu}_+(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{h > X_i \geq 0\}(Y_i - \beta_0 - X_i\beta_1)^2$$

and

$$\hat{\mu}_-(h) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{n} \mathbb{1}\{0 > X_i > -h\}(Y_i - \beta_0 - X_i\beta_1)^2.$$

Conventional (naive) confidence intervals can be calculated by using an asymptotic approximation for $\hat{\tau}(h)$. In particular, if

$$\frac{\hat{\tau}(h) - \tau}{\sqrt{V(h)}} \to^d N(0,1), \tag{1}$$

[9] See Hahn, Todd, and Van der Klaauw (2001), Porter (2003) or Fan and Gijbels (1992) for discussions of the properties of local polynomial regressions for boundary problems. The bootstrap algorithm proposed in this paper can be extended to accommodate higher order polynomials discontinuities in the derivatives of the conditional expectation, like "Kink RD" design (Card et al., 2009).

with

$$V(h)/\mathbb{V}(\hat{\tau}(h) \mid X_1,\ldots,X_n) \to^p 1$$

then valid confidence intervals can be constructed through the usual method of inverting the t-test.[10] This procedure gives the widely-used interval estimator

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

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

The statistical properties of these estimators, however, clearly depend on the bandwidth parameter h, and bandwidths that have desirable properties for point estimation may not have desirable properties for hypothesis testing or interval estimation. In particular, for (1) to hold, h must satisfy $nh \to \infty$ and $nh^5 \to 0$. (Hahn, Todd, and Van der Klaauw, 2001; Porter, 2003) Otherwise, the finite-sample bias of $\hat{\tau}(h)$ does not converge in probability to zero quickly enough and it contributes non-negligibly to the asymptotic distribution in (1). This result holds even though $\hat{\tau}(h)$ can be consistent under those conditions. These bandwidth issues are relevant in practice because many widely-used bandwidth selection procedures, most notably the AMSE-optimal bandwidth and cross-validation bandwidth (Imbens and Kalyanaraman, 2012) do not produce $o_p(n^{-1/5})$ bandwidths.

CCT solve this problem by deriving the analytical form of the first-order bias and explicitly recentering $\hat{\tau}(h)$. Under weaker assumptions on the asymptotic behavior of the bandwidth, which we will specify in Assumption 2, CCT show that the approximate bias of $\hat{\tau}(h)$ has the form

$$\mathbb{E}(\hat{\tau}(h) \mid X_1,\ldots,X_n) - \tau = h^2\Big[\tfrac{\mu_+^{(2)}}{2}\mathcal{B}_+(h) - \tfrac{\mu_-^{(2)}}{2}\mathcal{B}_-(h)\Big](1+o_p(1))$$

where $\mathcal{B}_+(h)$ and $\mathcal{B}_-(h)$ are observed quantities that depend on the kernel, bandwidth, and running variables $X_1,\ldots,X_n$; formal definitions of these terms are given in the Mathematical Appendix. The plug-in bias-corrected estimator then requires estimates for the second derivatives of the conditional mean from above and below the cutoff, $\mu_+^{(2)}$ and $\mu_-^{(2)}$, and CCT show that these derivatives can be estimated by fitting a second order local polynomial, i.e. one order higher than the polynomial used to obtain $\hat{\tau}$, using a (potentially) different pilot bandwidth b. Their procedure gives the bias-corrected estimator

$$\hat{\tau}'(h,b) = \hat{\tau}(h) - h^2\Big[\tfrac{\hat{\mu}_+^{(2)}(b)}{2}\mathcal{B}_+(h) - \tfrac{\hat{\mu}_-^{(2)}(b)}{2}\mathcal{B}_-(h)\Big]$$

[10] The mathematical appendix of this paper gives a precise definition for $V(h)$.

The variance introduced by the bias-correction term does not vanish, so the naive confidence interval needs not only to be re-centered to correct the bias, but also rescaled to allow for the additional variability introduced by the bias correction, resulting in the following asymptotic approximation:

$$\frac{\hat{\tau}'(h,b) - \tau}{V'(h,b)^{1/2}} \to^d N(0,1)$$

with $V'(h,b) = V(h) + C(h,b)$ and $C(h,b)$ an additional variance component generated by the bias-correction term.[11] This new approximation can be used instead of (1) to construct "bandwidth robust" confidence intervals, and CCT provide simulation evidence that their intervals perform well in finite samples even when the naive interval performs badly.

Assumption 2 specifies the bandwidth conditions assumed by CCT, which we will also use in this paper.

Assumption 2 (Bandwidth)**.** *Let h be the bandwidth used to estimate the local linear model and let b be the bandwidth used to estimate a second local quadratic model. Then $nh \to \infty$, $nb \to \infty$, $nh^5b^2 \to 0$, and $nb^5h^2 \to 0$ as $n \to \infty$.[12] The relationship $h \leq b$ also holds for all n.*

In the next section, we build upon the insight provided by CCT bias-corrected estimator and propose a simple bootstrap procedure that can directly construct the robust CIs without requiring the derivation of analytical formulas and direct estimators for the bias, variance and covariance terms, while relying on the same first-order bias correction approximation. The requirement $h < b$ is one additional restriction that we impose in this paper because the bootstrap can not be implemented for the uniform kernel without it, but the other parts of Assumption 2 are identical to CCT.[13]

3 Bootstrap Bias Correction

This section presents our theoretical contributions. We propose two algorithms in this section. The first uses a residual bootstrap based on a second order local polynomial to estimate the bias of the local linear model. That estimate can be subtracted from the biased original estimator to provide an asymptotically unbiased estimator with the same asymptotic distribution as CCT's. As in CCT, this bias correction term introduces a new source of variance, invalidating standard (naive) critical values. Consequently, the second algorithm we propose uses an iterated bootstrap to estimate the correct critical values of the bias corrected estimator. The intuition behind both

[11] $C(h,b)$ and $V(h)$ are defined precisely in the mathematical appendix.

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

[13] This assumption can be relaxed by considering other kernels, which is the subject of current research by the authors.

procedures is straightforward. CCT show that a local second order polynomial captures the aspects of the DGP necessary for constructing valid confidence intervals. Our proposed algorithms estimate and embed the second order behavior in the bootstrap DGP through a residual bootstrap.

Throughout this section and the rest of the paper, we will let $\mathbb{E}^*$, $\Pr^*$, etc. denote expectations and probabilities taken with respect to the distribution induced by the bootstrap (which implicitly conditions on $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$) and let parameters with * superscripts be the parameter values under the distribution induced by the bootstrap. Two * superscripts indicate that the parameter or probability measure corresponds to a secondary bootstrap distribution.

Algorithm 1 explains the bias-correction steps in detail.

Algorithm 1 (Bias estimation)**.** *Assume h and b are bandwidths as defined by Assumption 2 and define*

$$I_-(h) = \{i : -h < X_i < 0\}, \qquad I_+(h) = \{i : 0 \le X_i < h\}.$$

Also define $M_-(h)$ and $M_+(h)$ to be the number of elements in $I_-(h)$ and $I_+(h)$ respectively, and $m_-(h,1),\dots,m_-(h,M_-(h))$ and $m_+(h,1),\dots,m_+(h,M_+(h))$ to be subsequences of $1,\dots,n$ that index $I_-(h)$ and $I_+(h)$, respectively.

1. *Estimate local second order polynomials $\hat{g}_-$ and $\hat{g}_+$ using the observations in $I_-(b)$ and $I_+(b)$:*

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

   with

$$\hat{\beta}_- = \arg\min_{\beta} \sum_{i\in I_-(b)} (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2$$
$$\hat{\beta}_+ = \arg\min_{\beta} \sum_{i\in I_+(b)} (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2)^2.$$

   Calculate the residuals

$$\hat{\varepsilon}_i = \begin{cases} Y_i - \hat{g}_-(X_i) & \text{if } X_i < 0 \\ Y_i - \hat{g}_+(X_i) & \text{otherwise.} \end{cases} \tag{3}$$

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

   (a) Draw an i.i.d. sample of size $M_-(b)$ from $\{\hat{\varepsilon}_i : i \in I_-(b)\}$ and one of size $M_+(b)$ from

$\{\hat{\varepsilon}_i : i \in I_+(b)\}$. *Let* $\varepsilon^*_{-,i}$ *and* $\varepsilon^*_{+,i}$ *denote the ith element of each sample and construct*

$$Y^*_{-,m_-(b,i)} = \hat{g}_-(X_{m_-(b,i)}) + \varepsilon^*_{-,i} \qquad Y^*_{+,m_+(b,i)} = \hat{g}_+(X_{m_-(b,i)}) + \varepsilon^*_{+,i}.$$

(b) Calculate $\hat{\mu}^*_+(h)$ *and* $\hat{\mu}^*_-(h)$ *by estimating the local linear model on the bootstrap data set:*[14]

$$\hat{\mu}^*_-(h) = \arg\min_{\mu} \min_{\beta} \sum_{i \in I_-(h)} (Y^*_i - \mu - \beta X^*_i)^2$$
$$\hat{\mu}^*_+(h) = \arg\min_{\mu} \min_{\beta} \sum_{i \in I_+(h)} (Y^*_i - \mu - \beta X^*_i)^2.$$

(c) Save $\hat{\tau}^*_k(h) = \hat{\mu}^*_+(h) - \hat{\mu}^*_-(h)$.

3. Estimate the bias as

$$\Delta^*(h,b) = \frac{1}{B_1} \sum_{k=1}^{B_1} \hat{\tau}^*_k(h) - \left[\hat{g}_+(0) - \hat{g}_-(0)\right]. \tag{4}$$

Note that $\hat{g}_+(0) - \hat{g}_-(0)$ is the true treatment effect under the distribution induced by this bootstrap. The bootstrap estimator works by constructing an approximate DGP with known properties. As the dataset gets larger, the approximate DGP mimics the unknown real DGP more closely, and the population parameter values in the bootstrap DGP can become accurate estimates of the true parameter values in the real DGP.

Under Assumptions 1 and 2 the procedure described by Algorithm 1 provides a consistent estimator of the bias component that converges fast enough in probability that it can be be used as a correction. As is standard in the bootstrap literature, we will assume that the number of bootstrap replications, B_1, is large enough that the simulation error can be ignored. Theorem 1 presents the result formally.

Theorem 1. *Under Assumptions 1 and 2,*

$$\frac{(\hat{\tau}(h) - \Delta^*(h,b) - \tau)}{V'(h,b)^{1/2}} \to^d N(0,1), \tag{5}$$

where $\Delta^*(h,b)$ *is defined by Equation 4.*

Note that the variance component of (5) is the same value used by CCT and introduced in (1). Theorem 1 implies that our bias-corrected estimator has the same asymptotic distribution as CCT's.

[14]Note that the indices of summation are chosen to correspond to the indices of the variables generated in the previous step.

This equivalence should be unsurprising; both estimators use a second order polynomial to directly estimate the bias of the local linear model, so they should behave very similarly.

Since the second order polynomial captures the relevant aspects of the DGP for estimating the variance as well as the bias, the asymptotic distribution of the bias corrected estimator $\hat{\tau}(h) - \Delta^*(h,b)$ can also be approximated with a bootstrap. We propose bootstrapping $\hat{\tau}(h) - \Delta^*(h,b)$ using the same residual bootstrap method used in Algorithm 1. Algorithm 2 provides the details of our procedure and Theorem 2 establishes its theoretical properties.

Algorithm 2 (Confidence intervals)**.** *Define the same notation as in Algorithm 1.*

1. *Estimate $\hat{g}_+$ and g_- and generate the residuals $\hat{\varepsilon}_i$ just as in Algorithm 1.*

2. *Repeat the following steps B_2 times to produce the bootstrap estimates $\hat{\tau}_1'^*(h,b), \ldots, \hat{\tau}_{B_2}'^*(h,b)$. For the kth value:*

   (a) *Draw an i.i.d. sample of size $M_-(b)$ from $\{\hat{\varepsilon}_i : i \in I_-(b)\}$ and one of size $M_+(b)$ from $\{\hat{\varepsilon}_i : i \in I_+(b)\}$. Let $\varepsilon^*_{-,i}$ and $\varepsilon^*_{+,i}$ denote the ith element of each sample and construct*

   $$Y^*_{-,m_-(b,i)} = \hat{g}_-(X_{m_-(b,i)}) + \varepsilon^*_{-,i} \qquad Y^*_{+,m_+(b,i)} = \hat{g}_+(X_{m_+(b,i)}) + \varepsilon^*_{+,i}.$$

   (b) *Calculate $\hat{\mu}^*_+(h)$ and $\hat{\mu}^*_-(h)$ by estimating the local linear model on the bootstrap data set:*

   $$\hat{\mu}^*_-(h) = \arg\min_{\mu} \min_{\beta} \sum_{i \in I_-(h)} (Y^*_i - \mu - \beta X^*_i)^2$$
   $$\hat{\mu}^*_+(h) = \arg\min_{\mu} \min_{\beta} \sum_{i \in I_+(h)} (Y^*_i - \mu - \beta X^*_i)^2.$$

   (c) *Apply Algorithm 1 to the bootstrapped data set,*

   $$(Y^*_{-,m_-(b,1)}, X_{-,m_-(b,1)}), \ldots, (Y^*_{-,m_-(b,M_-(b))}, X_{-,m_-(b,M_-(b))}),$$
   $$(Y^*_{+,m_+(b,1)}, X_{+,m_+(b,1)}), \ldots, (Y^*_{+,m_+(b,M_+(b))}, X_{+,m_+(b,M_+(b))})$$

   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 1.

   (d) *Save the bias-corrected estimator $\hat{\tau}'^*_k(h,b) = \hat{\mu}^*_+(h) - \hat{\mu}^*_i(h) - \Delta^{**}(h,b)$.*

3. *Use the empirical CDF of $\hat{\tau}'^*_1(h,b), \ldots \hat{\tau}'^*_{B_2}(h,b)$ to construct confidence intervals, etc.*

Theorem 2 establishes that this iterated bootstrap approximates the asymptotic distribution of the bias-corrected statistic proposed by Algorithm 1 and justifies this second algorithm. As before, we assume that B_1 and B_2 are large enough that simulation error can be ignored.

Theorem 2. *Under Assumptions 1 and 2,*

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

and

$$\sup_x \left| \Pr^*[\hat{\tau}^*(h) - \Delta^{**}(h,b) - \tau^* \leq x] - \Pr[\hat{\tau}(h) - \Delta^*(h,b) - \tau \leq x] \right| \to^p 0.$$

Evidence of the usefulness of the procedures proposed above and their relative performance to the analytical bias correction proposed in CCT are presented in a series of Monte Carlo simulations in Section 4.

4 Simulation Evidence

This section presents evidence from Monte Carlo simulations that the bootstrap procedures proposed in Section 3 produce valid, robust confidence intervals similar to those obtained by the analytical procedures established in CCT. The bootstrap CIs obtained compare favorably to the analytical alternative, with coverage slightly closer to nominal coverage and shorter length of the intervals in the specifications implemented.

The Monte Carlo experiments have a similar structure. For all of them, we generate 500 i.i.d. observations from the DGP

$$\begin{aligned} Y_i &= \mu_j(X_i) + \varepsilon_i \\ X_i &\sim 2 \times beta(2,4) - 1 \\ \varepsilon_i &\sim N(0, 0.1295^2), \end{aligned}$$

where j will index the specific DGP. This is the experimental design used by Imbens and Kalyanaraman (2012) and CCT, which we adopt here to make our simulation results directly comparable with theirs and with the rest of the literature. We will use the same three functional forms for μ_j as CCT as well.

The first DGP is designed to match features of Lee's (2008) analysis of U.S. congressional elections. Lee estimates the incumbency advantage in electoral races for the House of Representatives — candidates who received the largest vote share in the previous election are the incumbents,

which creates the discontinuity. The conditional expectation is a fifth order polynomial fit to that dataset, (after excluding a small number of extreme observations; see Imbens and Kalyanaraman, 2012, or CCT for further details) giving

$$\mu_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise.} \end{cases}$$

The population ATE for this DGP is 0.04 (= 0.52 − 0.48).

The second DGP is based on Ludwig and Miller's (2007) analysis of the Head Start program. Funding eligibility is determined at the county level using the county's historical poverty rate, with a sharp threshold that determines the provision of services. We use the fifth order polynomial estimated on Ludwig and Miller's dataset as the conditional expectation for the second DGP:

$$\mu_2(x) = \begin{cases} 3.71 + 2.30x + 3.28x^2 + 1.45x^3 + 0.23x^4 + 0.03x^5 & \text{if } x < 0, \\ 0.26 + 18.49x - 54.81x^2 + 74.30x^3 - 45.02x^4 + 9.83x^5 & \text{otherwise} \end{cases}$$

and the population ATE is −3.45 (= 0.26 − 3.71).

Finally, for the third DGP, we use CCT's modification of μ_1, given by

$$\mu_3(x) = \begin{cases} 0.48 + 1.27x + 3.59x^2 + 14.147x^3 + 23.694x^4 + 10.995x^5 & \text{if } x < 0 \\ 0.52 + 0.84x - 0.30x^2 + 2.397x^3 - 0.901x^4 + 3.56x^5 & \text{otherwise} \end{cases}$$

and the population ATE is again 0.04. CCT introduce this DGP because it has high curvature and local linear models are likely to exhibit high bias, making it a natural test case for both their analytical corrections and our bootstrap.

To estimate the finite-sample coverage of our new bootstrap based confidence interval, we simulate 5000 samples from each of the three DGPs and calculate nominal 95% two-sided confidence intervals. We use 999 bootstrap replications (B_2) to calculate the asymptotic distribution of the bias corrected estimator, and each of those replications uses an additional 500 replications (B_1) to estimate the bias. We use the 0.025 and 0.975 quantiles of the bootstrap distribution as the lower and upper bound of the interval, and the bandwidths, h and b, are chosen using the AMSE-optimal rule proposed by CCT.

We also show the coverage of two of CCT's interval estimators for comparison. We estimate their bias corrected procedure with AMSE-optimal bandwidths for both the uniform kernel and the triangular kernel. The uniform kernel provides a direct comparison with our bootstrap, while the triangular kernel may be more appealing to practitioners. Since CCT establish conclusively that the "naive" uncorrected interval estimator has poor coverage in these settings we do not report results

DGP	Method	Bias	SD	RMSE	CI Coverage (%)	CI Length
1	Resid. bootstrap	–0.014	0.067	0.069	93.4	0.242
	CCT (uniform)	–0.014	0.067	0.069	92.5	0.246
	CCT (triangular)	–0.014	0.067	0.068	91.4	0.239
2	Resid. bootstrap	–0.011	0.088	0.089	95.1	0.323
	CCT (uniform)	–0.011	0.088	0.089	93.7	0.353
	CCT (triangular)	–0.008	0.086	0.086	93.2	0.346
3	Resid. bootstrap	–0.004	0.065	0.065	95.9	0.247
	CCT (uniform)	–0.004	0.065	0.065	93.8	0.251
	CCT (triangular)	–0.005	0.065	0.065	93.4	0.244

Table 1: Experimental coverage probabilities for each interval estimator based on 5000 simulations; nominal coverage probabilities are 95% for each estimator. The column "CI Coverage" lists the coverage frequency in these simulations and "CI Length" lists the average length of the confidence interval across simulations.

for that estimator.

Table 1 presents the results of these simulations. The rows labeled "Resid. bootstrap" show results for our proposed bootstrap interval; "CCT (uniform)" and "CCT (triangular)" show results for the analytically-corrected intervals. The first three columns give the bias, standard deviation, and Root-MSE of the *bias corrected* point estimators corresponding to each confidence interval. As expected, these estimators are close to unbiased, with their standard deviation largely determining the estimators' RMSE. Moreover, as Theorem 1 suggests, the bootstrap and analytically-corrected estimators have essentially the same standard deviation across all of the DGPs.

The column "CI Coverage" lists the experimental coverage of each interval. All of the intervals are reasonably close to their nominal coverage, although the analytically corrected estimators seem to persistently under-cover the true ATE (especially the one using triangular kernel). Our proposed bootstrap procedure consistently performs well, as it is about a percentage point closer to nominal coverage than the other interval estimators in DGPs 1 and 2, and is slightly conservative (95.9% coverage) in DGP 3. The final column, "CI Length," indicates that all three of these procedures generate intervals that are approximately the same length on average.

These are DGPs where the naive confidence intervals are known to perform poorly, and this set of simulation results indicates that our proposed bootstrap approach is quite competitive with analytical methods for producing valid intervals. Overall, the bootstrap bias-correction procedure proposed in this paper provides a simple alternative to obtain valid robust confidence intervals in RD designs and performs well compared to the analytical bias correction procedures proposed by CCT.

5 Application

In this section, we apply the bootstrap procedure to the data used in Ludwig and Miller (2007).[15] In their paper, the effects of Head Start application assistance on health and schooling were investigated under a sharp RD design.

In 1965, the Head Start program was established to help poor children age three to five and their families. The program elements include parent involvement, nutrition, social services, mental health services and health services. To promote this program in the most needing area, the Office of Economic Opportunity provided grant-writing assistance to the poorest 300 counties in the United States based on the 1960 poverty rate. So the poverty rate of the 300th poorest county serves as a sharp cutoff of treatment. It is shown in Ludwig and Miller (2007) that the 228 "treatment" counties with poverty rates 10 percentage points above this cutoff have average Head Start spending per four-year-old as twice of that for 349 "control" counties with poverty rates 10 percentage points below this cutoff.

Ludwig and Miller (2007) utilize this fact to estimate the "intent-to-treat" effect: the effect of proposal developing assistance on health and schooling. They use mortality as their health outcome measure and use data from the National Vital Statistics System of the National Center for Health Statistics, which provide information on cause of death and age at death. Ludwig and Miller (2007) limited the causes of death to those which could be affected by Head Start health services and found a large drop in mortality rates of children five to nine years of age over the period of 1973–1983. They also found some evidence for a positive effect on schooling from decennial census data.

We reestimate the ATE on health and schooling with robust procedures. To be specific, we apply both the robust procedure proposed by CCT and the bootstrap procedure introduced in this paper. Our bootstrap estimator uses the AMSE-optimal bandwidths proposed by CCT and the uniform kernel. In the bootstrap procedure, we use 500 bootstraps for bias correction and 999 to calculate the confidence intervals. The analytical estimator using CCT's bias correction and variance estimator use their AMSE-optimal bandwidths as well.

For completeness, we also report the original results in Ludwig and Miller (2007).[16] Since our research focuses on confidence intervals, we calculate and report the conventional unadjusted RD confidence interval for each of the bandwidths used by Ludwig and Miller (2007). (Ludwig and Miller use a range of bandwidths for their analysis.) They also use a paired bootstrap algorithm to generate p-values, and we report those p-values for completeness.

Table 2 shows the results of the Head Start program on mortality of children five to nine years

[15]The data is publicly available from http://faculty.econ.ucdavis.edu/faculty/dlmiller/statafiles.

[16]One minor issue arose in reproducing Ludwig and Miller's results: their paper presents results for the *triangular* kernel, but we were only able to recover their ATE estimate using the *uniform* kernel. Results for the triangular and uniform kernel were essentially the same, and this discrepancy does not affect Ludwig and Miller's conclusions.

	ATE	95% CI	h	b	p-value
LM (2007)	–1.895	(–3.930, 0.139)	9		0.036
LM (2007)	–1.198	(–2.561, 0.165)	18		0.081
LM (2007)	–1.114	(–2.138, –0.090)	36		0.027
CCT	–3.795	(–7.037, –0.554)	3.888	6.807	
Resid. bootstrap	–3.792	(–6.512, –0.262)	3.888	6.807	

Table 2: The effect of Head Start assistance on mortality. The first three rows come from Table 3 in Ludwig and Miller (2007) except "95% CI," which is calculated using the conventional asymptotic interval estimator. The last two rows list results from two robust procedures.

of age.[17] Instead of choosing an optimal bandwidth, Ludwig and Miller (2007) adopted three candidate bandwidths 9, 18, and 36. Their estimates indicate that Head Start assistance lowers the targeted mortality rate by –1.895, –1.198 and –1.114 respectively, which are not very sensitive to the choice of bandwidths in this range. These ATEs are also significantly different from zero (with p-value 0.036, 0.081 and 0.027) based on Ludwig and Miller's percentile-t bootstrapped p-value. The statistical inference changes when conventional analytical confidence interval is used, which includes zero for bandwidth 9 and 18.

Results from the two robust procedures are similar to each other but greatly differ from the original estimates. The estimated ATEs are as high as –3.795 from CCT and –3.792 from the bootstrap. Both are significantly different from zero, though the confidence intervals are also very wide, which is likely to be due to the much smaller bandwidths used relative to the other estimators ($h = 3.888$, $b = 6.807$).

Table 3 presents the effect of the program on schooling for the cohort aged 18–24 in 1990. The measurement of schooling is the fraction of people with high school or more in Panel A and the fraction of people with some college or more in Panel B. A bandwidth of 7 is used in Ludwig and Miller (2007). Their estimates suggest that Head Start assistance increases the fraction of people with high school or more by 3% and the fraction of people with some college or more by 3.7%. Both are significantly different from zero based on both their p-value and the conventional analytical confidence interval.

The two robust procedures are again very similar to each other and give slightly larger point estimates in both panels. In Panel A, the ATE increases from 0.030 to 0.055 (CCT) and 0.054 (bootstrap). In Panel B, the ATE increases from 0.037 to 0.051 (CCT) and 0.052 (bootstrap). The confidence intervals tend to shift up as well. In contrast to the striking differences between the conventional estimates bias-corrected procedures for the mortality estimates, the differences in Table 3 are much smaller.

[17] We, like Ludwig and Miller, focus on the 1973–1983 period.

	ATE	95% CI	h	b	p-value
Fraction "high school or more" (Panel A)					
LM (2007)	0.030	(0.003, 0.057)	7		0.032
CCT	0.055	(0.014, 0.096)	3.671	8.618	
Bootstrap	0.054	(0.013, 0.096)	3.671	8.618	
Fraction "some college or more" (Panel B)					
LM (2007)	0.037	(0.002, 0.073)	7		0.032
CCT	0.051	(0.004, 0.099)	5.076	10.251	
Bootstrap	0.052	(0.001, 0.094)	5.076	10.251	

Table 3: The effect of Head Start assistance on education for cohort 18–24 in 1990. Panel A uses the fraction of people with high school or more as dependent variable. Panel B uses the fraction of people with some college or more as dependent variable. The first row in each panel comes from Table 4 in Ludwig and Miller (2007) except "95% CI," which is calculated using the conventional asymptotic interval estimator. The last two rows in each panel list results from two robust procedures.

To briefly summarize, the bootstrap procedure performs similarly to the robust estimator proposed by CCT in the above applications. Both provide somewhat dissimilar answers from the classical point and interval estimators that do not account for the estimator's bias, but our estimates largely support the direction and the statistical significance of Ludwig and Miller's (2007) empirical findings.

6 Conclusion

This paper proposes a novel bootstrap procedure to obtain robust bias-corrected confidence intervals in sharp regression discontinuity designs using a uniform kernel. The approach proposed builds upon the developments and intuition advanced by CCT and is based on a first-order bias correction. We exploit CCT's theoretical insight through a new residual bootstrap. In particular, we propose estimating the local linear model as usual, then estimating a local second order polynomial and generating bootstrap datasets by resampling the residuals of that polynomial. This bootstrap allows the bias of the linear model to be estimated and removed, and the bootstrap can be repeated to accurately estimate the sampling distribution of the bias-corrected estimator.

Bootstrap procedures are appealing in this setting because they are relatively easy for applied researchers to modify to account for new and unusual dependence structures or functional forms. The variance adjustment was proven to be correct under the assumption of i.i.d. data, but could easily be extended to handle cross-sectional or time series dependence by using the appropriate

resampling strategy on the second order polynomial's residuals. For the analytical methods, however, new formulas need to be derived to accommodate new features of the DGP. To fully take advantage of the bootstrap's flexibility, though, these results need to be extended to other kernels, design strategies (e.g. fuzzy and kink RD designs) and other more realistic dependence structures. All of these are the subject of ongoing research.

A Mathematical appendix

Let e_p be the selection vector with 1 in element $p+1$ and 0 everywhere else. Define the following additional notation:[18] $r_p(x) = (1,\ x,\ldots,\ x^p)'$,

$$(\hat{\mu}_{+,p}(h),\ \hat{\mu}^{(1)}_{+,p}(h),\ldots,\ \hat{\mu}^{(p)}_{+,p}(h))' = \arg\min_{\beta} \sum_{i\in I_+(h)} (Y_i - r_p(X_i/h)'\beta)^2$$
$$(\hat{\mu}_{-,p}(h),\ \hat{\mu}^{(1)}_{-,p}(h),\ldots,\ \hat{\mu}^{(p)}_{-,p}(h))' = \arg\min_{\beta} \sum_{i\in I_-(h)} (Y_i - r_p(X_i/h)'\beta)^2$$

and

$$\begin{aligned}
\Gamma_{+,p}(h) &= \tfrac{1}{nh} \sum_{i\in I_+(h)} r_p(X_i/h) r_p(X_i/h)' \\
\Gamma_{-,p}(h) &= \tfrac{1}{nh} \sum_{i\in I_-(h)} r_p(X_i/h) r_p(X_i/h)' \\
\Psi_{+,p,q}(h,b) &= \tfrac{1}{nhb} \sum_{i\in I_+(\min(h,b))} r_p(X_i/h) r_q(X_i/b)'\, \mathbb{V}(Y_i \mid X_i) \\
\Psi_{-,p,q}(h,b) &= \tfrac{1}{nhb} \sum_{i\in I_-(\min(h,b))} r_p(X_i/h) r_q(X_i/b)'\, \mathbb{V}(Y_i \mid X_i) \\
\mathfrak{B}_+(h) &= e_0'\Gamma_{+,1}^{-1} \sum_{i\in I_+(h)} r_1(X_i/h)\ X_i^2/h^2, \\
\mathfrak{B}_-(h) &= e_0'\Gamma_{-,1}^{-1} \sum_{i\in I_-(h)} r_1(X_i/h)\ X_i^2/h^2.
\end{aligned}$$

Also, for reference, define the variance terms

$$V(h) = e_0'\Big(\tfrac{1}{n}\Gamma_{-,1}^{-1}\Psi_{-,1}\Gamma_{-,1}^{-1} + \tfrac{1}{n}\Gamma_{+,1}^{-1}\Psi_{+,1}\Gamma_{+,1}^{-1}\Big)e_0$$

[18] As we mention in Section 2, we have adopted CCT's notation where possible and these terms originate in that paper.

and

$$C(h,b) = n^{-1}e_2'\Big[\Gamma_{+,2}^{-1}(b)\Psi_{+,2,2}(b,b)\Gamma_{+,2}^{-1}(b)\mathfrak{B}_+^2(h) + \Gamma_{-,2}^{-1}(b)\Psi_{-,2,2}(b,b)\Gamma_{-,2}^{-1}(b)\mathfrak{B}_-^2(h)\Big]e_2$$
$$- 2h^2n^{-1}b^{-2}e_0'\Big[\Gamma_{+,1}^{-1}(h)\Psi_{+,1,2}(h,b)\Gamma_{+,2}^{-1}(b)\mathfrak{B}_+(h) + \Gamma_{-,1}^{-1}(h)\Psi_{-,1,2}(h,b)\Gamma_{-,2}^{-1}(b)\mathfrak{B}_-(h)\Big]e_2.$$

See CCT for the motivation and derivation of these formulas.

A.1 Proof of Theorem 1

We have

$$\hat{\tau}(h) - \Delta^*(h,b) - \tau = (\hat{\tau}(h) - \mathbb{E}\,\hat{\tau}(h)) + (\mathbb{E}\,\hat{\tau}(h) - \tau) - (\mathbb{E}^*\,\hat{\tau}^*(h) - \tau^*).$$

The design of the bootstrap ensures that

$$\mathbb{E}^*\,\hat{\mu}_{+,1}^*(h) - \mu_+^* = h^2\mu_+^{*(2)}\mathfrak{B}_+(h)/2, \qquad \mathbb{E}^*\,\hat{\mu}_{-,1}^*(h) - \mu_-^* = h^2\mu_-^{*(2)}\mathfrak{B}_-(h)/2,$$

almost surely, implying that

$$\mathbb{E}^*\,\hat{\tau}^*(h) - \tau^* = h^2\,\mu_+^{*(2)}\mathfrak{B}_+(h)/2 - h^2\,\mu_-^{*(2)}\mathfrak{B}_-(h)/2.$$

CCT's Lemma A1 implies that

$$\mathbb{E}\,\hat{\tau}(h) - \tau = h^2\,\mu_+^{(2)}\mathfrak{B}_+(h)/2 - h^2\,\mu_-^{(2)}\mathfrak{B}_-(h)/2 + O_p(h^3)$$

as well, giving

$$\begin{aligned}\hat{\tau}(h) &- \mathbb{E}\,\hat{\tau}(h) + (\mathbb{E}\,\hat{\tau}(h) - \tau) - (\mathbb{E}^*\,\hat{\tau}^*(h) - \tau^*)\\ &= \hat{\tau}(h) - \mathbb{E}\,\hat{\tau}(h) + h^2((\mu_-^{*(2)} - \mu_-^{(2)})\mathfrak{B}_-(h)/2 - (\mu_+^{*(2)} - \mu_+^{(2)})\mathfrak{B}_+(h)/2) + O_p(h^3)\\ &= \hat{\tau}(h) - \mathbb{E}\,\hat{\tau}(h) + h^2(\hat{\mu}_{-,2}^{(2)}(b) - \mu_-^{(2)})\mathfrak{B}_-(h)/2 \\ &\qquad - h^2(\hat{\mu}_{+,2}^{(2)}(b) - \mu_+^{(2)})\mathfrak{B}_+(h)/2 + O_p(h^3)\end{aligned} \tag{6}$$

The second equality holds because $\mu_+^{*(2)} = \hat{\mu}_{+,2}^{(2)}(b)$ and $\mu_-^{*(2)} = \hat{\mu}_{-,2}^{(2)}(b)$ almost surely. Asymptotic normality then follows from normality of $\hat{\tau}(h) - \mathbb{E}\,\hat{\tau}(h)$, $\hat{\mu}_{+,2}^{(2)}(b) - \mu_+^{(2)}$, and $\hat{\mu}_{-,2}^{(2)}(b) - \mu_-^{(2)}$ using similar arguments to CCT's Lemma SA4.D. □

A.2 Proof of Theorem 2

Repeat the steps from Theorem 1's proof through (6) for the iterated bootstrap to get

$$\begin{aligned}\hat{\tau}(h)^* - \Delta^{**}(h,b) - \tau^* &= (\hat{\tau}^*(h) - \mathbb{E}^*\,\hat{\tau}^*(h)) + (\mathbb{E}^*\,\hat{\tau}^*(h) - \tau^*) - (\mathbb{E}^{**}\,\hat{\tau}^{**}(h) - \tau^{**}) \\ &= \hat{\tau}^*(h) - \mathbb{E}^*\,\hat{\tau}^*(h) + h^2(\hat{\mu}_{-,2}^{*(2)}(b) - \mu_-^{*(2)})\mathfrak{B}_-(h)/2 - h^2(\hat{\mu}_{+,2}^{*(2)}(b) - \mu_+^{*(2)})\mathfrak{B}_+(h)/2 \\ &= \Omega_+(h,b)'\boldsymbol{\varepsilon}_+^* - \Omega_-(h,b)'\boldsymbol{\varepsilon}_-^*,\end{aligned}$$

where $\boldsymbol{\varepsilon}_+^* = (\varepsilon_{+,1}^*, \ldots, \varepsilon_{+,M_+(b)}^*)'$ and $\Omega_+(h,b)$ is an $M_+(b)$-dimensional vector with ith element $\Omega_{+,1i}(h,b) - \Omega_{+,2i}(h,b)$, which are defined as

$$\Omega_{+,1i}(h,b) = (1 \quad 0)\Big(\sum_{j\in I_+(h)} r_1(X_j/h)r_1(X_j/h)'\Big)^{-1} r_1(X_{m_+(b,i)}/h)\,\mathbb{1}\{h > X_{m_+(b,i)} \geq 0\}$$

and

$$\Omega_{+,2i}(h,b) = (0 \quad 0 \quad h^2)\Big(\sum_{j\in I_+(b)} r_2(X_j/b)r_2(X_j/b)'\Big)^{-1} r_2(X_{m_+(b,i)}/b).$$

The definitions of $\boldsymbol{\varepsilon}_-^*$ and $\Omega_-(h,b)$ have the same definitions as $\boldsymbol{\varepsilon}_+^*$ and $\Omega_+(h,b)$ after making the obvious substitutions of "−" for "+."

Notice that (6) implies that

$$\hat{\tau}(h) - \Delta^*(h,b) - \tau =^d \Omega_+(h,b)'\boldsymbol{\varepsilon}_+ - \Omega_-(h,b)'\boldsymbol{\varepsilon}_- + O_p(h^3),$$

with $\boldsymbol{\varepsilon}_+$ an i.i.d. random vector of length $M_+(b)$, the ith element of which is distributed as

$$\varepsilon_{i,+} =^d \begin{cases} Y_{m_+(1,b)} - \mathbb{E}(Y_{m_+(1,b)} \mid X_{m_+(1,b)}) & \text{with probability } 1/M_+(b) \\ \vdots & \\ Y_{m_+(M_+(b),b)} - \mathbb{E}(Y_{m_+(M_+(b),b)} \mid X_{m_+(1,b)}) & \text{with probability } 1/M_+(b), \end{cases}$$

and $\boldsymbol{\varepsilon}_-$ the corresponding quantity after replacing "+" with "–." Consequently, it suffices to prove that

$$\rho(V'(h,b)^{-1/2}\Omega_+(h,b)'\boldsymbol{\varepsilon}_+^*,\ V'(h,b)^{-1/2}\Omega_+(h,b)'\boldsymbol{\varepsilon}_+) \to^p 0 \tag{7}$$

and

$$\rho(V'(h,b)^{-1/2}\Omega_-(b)'\boldsymbol{\varepsilon}_-^*,\ V'(h,b)^{-1/2}\Omega_-(b)'\boldsymbol{\varepsilon}_-) \to^p 0 \tag{8}$$

with ρ the "Mallows metric" used by Bickel and Freedman (1981).[19] Convergence in ρ is equivalent to convergence in both distribution and second moments. (Bickel and Freedman, 1981, Lemma 8.3.)

We will only prove (7) since the proof of (8) is identical. By Theorem 2.1 of Freedman (1981), we have

$$\begin{aligned}\rho(V'(h,b)^{-1/2}\Omega_+(h,b)'\boldsymbol{\varepsilon}^*_+, V'(h,b)^{-1/2}\Omega_+(h,b)'\boldsymbol{\varepsilon}_+) \\ &\leq \rho(\varepsilon^*_{+,1},\varepsilon_{+,1}) \times \text{tr}(V'(h,b)^{-1/2}\Omega_+(h,b)'\Omega_+(h,b)V'(h,b)^{-1/2}) \\ &= \rho(\varepsilon^*_{+,1},\varepsilon_{+,1}) \times O_{p^*}(1),\end{aligned}$$

the second line holding as a consequence of CCT's Lemma SA1. So it suffices to show that $\rho(\varepsilon^*_{+,1},\varepsilon_{+,1}) \to^{p^*} 0$. Let ε°_i be an i.i.d. sequence randomly drawn from the realized values of $\boldsymbol{\varepsilon}_+$ with replacement and let $\bar{\varepsilon}^\circ = (1/M_+(b)) \sum_{i=1}^{M_+(b)} \varepsilon^\circ_i$. Then, using the same arguments as Freedman (1981), we have the upper bound

$$\rho(\varepsilon^*_{+,1},\varepsilon_{+,1}) \leq \left[(\bar{\varepsilon}^\circ)^2 + (2/M_+(b)) \sum_{i=1}^{M_+(b)} (\hat{\varepsilon}_i - \varepsilon_i)^2\right]^{\frac{1}{2}} + \rho(\varepsilon^\circ_1,\, \varepsilon_{+,1}).$$

But $(\bar{\varepsilon}^\circ)^2 \to^p 0$ by the LLN,

$$(1/M_+(b)) \sum_{i=1}^{M_+(b)} \mathbb{E}((\hat{\varepsilon}_i - \varepsilon_i)^2 \mid X_i) \to 0$$

by CCT's Lemma SA1, and $\rho(\varepsilon^\circ_1,\, \varepsilon_{+,1}) \to^p 0$ by Lemma 8.4 of Bickel and Freedman (1981). □

[19] For two random vectors u and v with finite 2nd moments, $\rho(u,v)$ is defined as

$$\rho(u,v) = \inf_{(U,V)\ s.t.\ U=^d u,\ V=^d v} (\mathbb{E}\|U-V\|^2)^{1/2}.$$

References

Barreca, Alan I, Melanie Guldi, Jason M Lindo, and Glen R Waddell. 2011. "Saving Babies? Revisiting the effect of very low birth weight classification." *The Quarterly Journal of Economics* 126 (4):2117–2123.

Bickel, Peter J. and David A. Freedman. 1981. "Some asymptotic theory for the bootstrap." *Annals of Statistics* :1196–1217.

Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocıo Titiunik. 2015. "Regression Discontinuity Designs using Covariates." Unpublished working paper.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.

———. 2015. *rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs*. R package version 0.80.

Card, David, Carlos Dobkin, and Nicole Maestas. 2009. "Does Medicare Save Lives?" *Quarterly Journal of Economics* 124 (2):597–636.

Card, David, David S Lee, Zhuan Pei et al. 2009. "Quasi-experimental identification and estimation in the regression kink design." Working Paper, Princeton University.

Caughey, Devin and Jasjeet S Sekhon. 2011. "Elections and the regression discontinuity design: Lessons from close US house races, 1942–2008." *Political Analysis* 19 (4):385–408.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *Quarterly Journal of Economics* 125 (1):215–261.

Erikson, Robert S and Rocío Titiunik. 2015. "Using Regression Discontinuity to Uncover the Personal Incumbency Advantage." *Quarterly Journal of Political Science* 10:101–119.

Fan, Jianqing and Irene Gijbels. 1992. "Variable Bandwidth and Local Linear Regression Smoothers." *Annals of Statistics* 20 (4):1669–2195.

Freedman, David A. 1981. "Bootstrapping regression models." *The Annals of Statistics* 9 (6):1218–1228.

Frölich, Markus. 2007. "Regression discontinuity design with covariates." IZA Discussion Paper No. 3024.

Fujiwara, Thomas. 2011. "A Regression Discontinuity Test of Strategic Voting and Duverger's Law." *Quarterly Journal of Political Science* 6 (3-4):197–233.

———. 2015. "Voting Technology, Political Responsiveness, and Infant Health: Evidence From Brazil." *Econometrica* 83 (2):423–464.

Gaujoux, Renaud. 2014. *doRNG: Generic Reproducible Parallel Backend for foreach Loops*. R package version 1.6.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1):201–209.

Hall, Peter and Michael A Martin. 1988. "On bootstrap resampling and iteration." *Biometrika* 75 (4):661–671.

Imbens, Guido W. and Karthik Kalyanaraman. 2012. "Optimal Bandwitdth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3):933–959.

Imbens, Guido W and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2):615–635.

Jacob, Brian A. and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86 (1):226–244.

Keele, Luke J and Rocio Titiunik. 2014. "Geographic Boundaries as Regression Discontinuities." *Political Analysis* :1–29.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2):675–697.

Lee, David S. and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142 (2):655–674.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281–355.

Loader, Clive. 2006. *Local regression and likelihood*. Springer Science & Business Media.

Ludwig, Jens and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122 (1):159–208.

Porter, Jack. 2003. "Estimation in the Regression Discontinuity Model." Unpublished Manuscript, Department of Economics, University of Wisconsin, Madison.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revolution Analytics and Steve Weston. 2015a. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.10.

———. 2015b. *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.3.

Schmieder, Johannes F., Till Von Wachter, and Stefan Bender. 2012. "The Effects of Extended Unemployment Insurance over the Business Cycle: Evidence from Regression Discontinuity Estimates over 20 Years." *Quarterly Journal of Economics* 127 (2):701–752.

Thistlethwaite, Donald L. and Donald T. Campbell. 1960. "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational psychology* 51 (6):309.

Urquiola, Miguel and Eric Verhoogen. 2009. "Class-size caps, sorting, and the regression-discontinuity design." *The American Economic Review* 99 (1):179–215.

Van der Klaauw, Wilbert. 2002. "Estimating the effect of financial aid offers on college enrollment: A Regression–Discontinuity Approach*." *International Economic Review* 43 (4):1249–1287.