Bias-Aware Inference in Fuzzy Regression Discontinuity Designs

Claudia Noack[*] Christoph Rothe[†]

Abstract

Fuzzy regression discontinuity (FRD) designs occur frequently in many areas of applied economics. We argue that the confidence intervals based on nonparametric local linear regression that are commonly reported in empirical FRD studies can have poor finite sample coverage properties for reasons related to their gereneral construction based on the delta method, and to how they account for smoothing bias. We therefore propose new confidence sets, which are based on an Anderson-Rubin-type construction. These confidence sets are bias-aware, in the sense that they explicitly take into account the exact smoothing bias of local linear regression. They are simple to compute, highly efficient, have excellent coverage properties in finite samples. They are also valid under weak identification (that is, if the jump in treatment probabilities at the threshold is small) and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

This Version: June 11, 2019. We thank Tim Armstrong and Michal Kolesár and numerous seminar participants for helpful comments and suggestions. Financial support of the ERC through grant SH1-77202 is gratefully acknowledged.

[*] Department of Economics, University of Mannheim, 68131 Mannheim, Germany. E-Mail: cnoack@mail.uni-mannheim.de.

[†] Department of Economics, University of Mannheim, 68131 Mannheim, Germany. E-Mail: rothe@vwl.uni-mannheim.de. Website: http://www.christophrothe.net.

1. INTRODUCTION

Regression discontinuity (RD) designs (Hahn et al., 2001; Imbens and Lemieux, 2008; Lee and Lemieux, 2010) have become a popular empirical strategy for estimating causal treatment effects from observational data in economics. In sharp regression discontinuity (SRD) designs units receive a treatment if and only if a running variable falls above some known threshold value, whereas in fuzzy regression discontinuity (FRD) designs the probability of treatment jumps discontinuously at the threshold, but generally not from zero to one. Methods for inference based on local linear regression are widely used in empirical research with both types of designs, and their theoretical properties have been studied extensively in the econometrics literature (Hahn et al., 2001; Porter, 2003; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2018).

With local linear SRD inference, one first estimates the jump in the conditional expectation of the outcome given the running variable by fitting separate linear model on each side of the threshold via weighted least squares, using weights that decrease with the running variable's distance to the threshold relative to a bandwidth. Unless the true conditional expectation is linear on each side of the threshold, the resulting jump estimator is typically biased, but its standard error is straightforward to obtain.

To form a valid confidence interval (CI), one then has to address the estimator's possible smoothing bias. The most common ways to do this are undersmoothing (choosing a "small" bandwidth to make the bias negligible), robust bias correction (subtracting an estimate of the bias, and adjusting the standard error appropriately, cf. Calonico et al., 2014), or the recently proposed bias-aware approach (adjusting the critical value for the exact "worst case" bias, cf. Armstrong and Kolesár, 2018). Bias-aware CIs have been shown to have advantages relative to the other approaches in that they are more efficient, tend to have better finite sample coverage properties, and are also valid with a discrete running variable (Armstrong and Kolesár, 2018, 2019; Kolesár and Rothe, 2018).

Inference in FRD designs is conceptually more complicated because the usual point estimator is the ratio of local linear estimates of the jumps in expected outcomes and treatment probabilities. The most common way to deal with this additional nonlinearity is through a "delta method" type approach. This entails approximating the jump ratio with a linear combination of the two estimated jumps, and imposing conditions under which the resulting approximation error is "asymptotically small". Since the linear term in this expansion effectively behaves like an SRD estimator, one can then apply one of the three above-mentioned techniques for handling smoothing bias to construct a CI, with the bias-aware approach again having advantages over undersmoothing and robust bias correction in this context (Armstrong and Kolesár, 2019).

In this paper, we argue that there are at least three main problems with such delta method CIs. First, they do not take into account the actual bias of the treatment effect estimator, but only that of a first-order approximation. In finite samples, even bias-aware versions of these CIs are subject to distortions that are not present in the SRD case. Second, delta method CIs generally perform poorly under weak identification, i.e. when the jump in treatment probabilities at the threshold is rather small (see also Feir et al., 2016). This "small denominator" issue is analogous to that of a weak first stage in the instrumental variables literature (Staiger and Stock, 1997). Third, delta method CIs are generally not valid if the running variable is discrete, because continuous variation around the threshold is needed for the approximation error to be "asymptotically small". Since discrete running variables abound in the empirical literature, this is an important limitation.[1]

We therefore propose new confidence sets (CSs) for treatment effects in FRD designs based on local linear regression that are not subject to such shortcomings. We avoid issues related to "linearizing" the FRD point estimator by basing our CSs on alternative auxiliary parameters that can be estimated directly via a single local linear regression step. This construction was previously considered by Feir et al. (2016), and is conceptually similar to that of Anderson-Rubin CSs in the instrumental variables literature. We then combine this approach with methods for bias-aware inference developed in Armstrong and Kolesár (2018, 2019). Our resulting *bias-aware Anderson-Rubin-type CSs* are simple to compute, highly efficient, and have excellent coverage properties in finite samples because they explicitly take into account the exact smoothing bias from the local linear regression steps. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

Two functions of the running variable play a key role in our analysis: the conditional expectation of the outcome and the conditional treatment probability. The derivation of our CSs assumes that both functions are smooth on either side of the cutoff, in the sense that their second derivatives are bounded in absolute value by some constant that is specified explicitly. Our main theoretical result is that the resulting CSs are honest, in the sense that

[1] Following Lee and Card (2008), it has become common practice in the applied literature to use standard errors that are clustered at the level of the running variable whenever the latter is discrete. This practice does not alleviate the issues caused by a discrete running variable for delta method CIs in FRD designs. Indeed, Kolesár and Rothe (2018) show that clustering by the running variable tends to produce CIs with poor coverage properties. Such standard errors, and corresponding CIs, should therefore not be used.

have correct asymptotic coverage uniformly over the class of candidates for these two key functions. Roughly speaking, honesty requires the CSs to perform well across the entire range of plausible data generating processes, and is thus necessary for good finite-sample coverage. A lack of honesty, or uniform asymptotic validity, is why some widely used methods for RD inference often perform poorly in practice (Kamat, 2018; Armstrong and Kolesár, 2019).

The just-mentioned bounds on second derivatives are the main tuning parameters required for the construction of our bias-aware Anderson-Rubin-type CSs. Once they are specified, our CSs are valid for any bandwidth choice, and optimal bandwidths are determined automatically. Choosing a bound close to zero effectively imposes the assumption that the respective function is close to linear, while choosing a larger bound allows for functions with increasingly higher curvature. In general, subject-specific knowledge is necessary to determine whether a particular derivative bound is plausible, and in practice we recommend reporting our CSs for a range of bound values in order to assess the robustness of empirical findings. Note that we cannot avoid specifying the derivative bounds: any alternative method for inference that maintains correct coverage over the same class of functions must make this choice either explicitly or implicitly (Armstrong and Kolesár, 2019).

Our paper contributes to an extensive methodological literature on RD; see Imbens and Lemieux (2008) or Lee and Lemieux (2010) for survey articles, and Cattaneo et al. (2019) for a textbook treatment. It is particularly related to Feir et al. (2016), who also consider Anderson-Rubin-type CSs in an FRD context. The main difference is that Feir et al. (2016) use undersmoothing instead of a bias-aware approach, which means their CSs are subject to a number of practical limitations common to all methods based on undersmoothing. See Section 6.4 for a more detailed discussion, and Section 7 for simulation results regarding the relative merits of our approach compared to that in Feir et al. (2016).

The remainder of this paper is structured as follows. In the following section, we describe our setup and introduce some baseline notation. Section 3 describes our proposed CSs, and Section 4 establishes their theoretical properties. Section 5 discusses a number of possible extensions, and Section 6 compares our CSs to others that have been proposed in the literature. In Section 7, we present our simulation study, and Section 8 contains an empirical application. Finally, Section 9 concludes. Proofs are given in the Appendix.

2. SETUP

2.1. **Model and Parameter of Interest.** In an FRD design, we seek to infer the causal effect of a binary treatment from a simple random sample of size n from a large population.

Let $Y_i \in \mathbb{R}$ be the outcome, $T_i \in \{0,1\}$ be the actual treatment status, $Z_i \in \{0,1\}$ be the assigned treatment, and $X_i \in \mathbb{R}$ be the running variable of the ith unit in the sample, for $i = 1, \dots, n$. Units are assigned to treatment if and only if the running variable crosses a known threshold, which we normalize to zero. This means that $Z_i = \mathbf{1}\{X_i \geq 0\}$, and due to limited compliance in FRD designs we could have $Z_i \neq T_i$ for some units. We then write $\mu_Y(x) = \mathbb{E}(Y_i|X_i = x)$ and $\mu_T(x) = \mathbb{E}(T_i|X_i = x)$ for the conditional expectation functions of the outcome and the treatment status indicator, respectively, given the running variable; and let $\mu_+ = \lim_{x\downarrow 0} \mu(x)$ and $\mu_- = \lim_{x\uparrow 0} \mu(x)$ denote the right and left limit, respectively, of a generic function μ at zero. Our parameter of interest θ is the ratio of jumps in the functions μ_Y and μ_T at the cutoff:

$$\theta = \frac{\tau_Y}{\tau_T}, \quad \tau_Y = \mu_{Y+} - \mu_{Y-}, \quad \tau_T = \mu_{T+} - \mu_{T-}. \tag{2.1}$$

This parameter typically has a causal interpretation as the local average treatment effect among compliers at the cutoff. That is, let $Y_i(t)$ be the potential outcome of unit i under treatment $t \in \{0,1\}$, so that $Y_i = Y_i(T_i)$; and, following terminology introduced in Imbens and Angrist (1994), refer to units that receive the treatment if and only if their realization of the running variable falls above the cutoff value as "compliers". Then, under certain continuity and monotonicity conditions (e.g., Hahn et al., 2001; Dong, 2017), we have that $\theta = \mathbb{E}(Y_i(1) - Y_i(0)|X_i = 0, \text{unit } i \text{ is a complier})$.

2.2. **Goal of the Paper.** Our goal in this paper is to construct confidence sets (CSs) that asymptotically cover the parameter θ with at least some pre-specified probability, uniformly in (μ_Y, μ_T) over some suitably chosen function class $\mathcal{F}$ that embodies the shape restrictions that the analyst is willing to impose. That is, we want to construct data-dependent sets $\mathcal{C}^\alpha \subset \mathbb{R}$ that satisfy

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1 - \alpha \tag{2.2}$$

for some $\alpha > 0$. Note that here and throughout the paper we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on the functions μ_Y and μ_T implicit in our notation. Following Li (1989), we refer to such a set $\mathcal{C}^\alpha$ as $100(1-\alpha)\%$ CS that is honest with respect to $\mathcal{F}$. This type of uniform asymptotic validity is a much stronger requirement than correct pointwise asymptotic coverage

$$\lim_{n\to\infty} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1 - \alpha \text{ for all } (\mu_Y, \mu_T) \in \mathcal{F}. \tag{2.3}$$

The condition (2.2) implies that we can find a sample size n such that the coverage probability of $\mathcal{C}^\alpha$ does not subceed $1-\alpha$ by more than an arbitrarily small amount for every $(\mu_Y, \mu_T) \in \mathcal{F}$. With only (2.3) there is no such guarantee, and even in very large samples the coverage probability of $\mathcal{C}^\alpha$ could be poor for some $(\mu_Y, \mu_T) \in \mathcal{F}$. Since we do not know in advance which function pair is the correct one, honesty as in (2.2) is necessary for good finite sample coverage of $\mathcal{C}^\alpha$ across data generating processes. Of course, we also want our CSs to be rather efficient, in the sense that they should be "small" while still maintaining honesty.

2.3. **Smoothness Conditions.** Following Armstrong and Kolesár (2019, 2018), we specify the class $\mathcal{F}$ of plausible candidates for (μ_Y, μ_T) as a smoothness class. Specifically, let

$$\mathcal{F}_H(B) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w''\|_\infty \leq B, w = 0, 1\}$$

be the Hölder-type class of functions mapping from the real line to the real line, that are potentially discontinuous at zero, are twice differentiable almost everywhere on either side of the threshold, and have second derivatives uniformly bounded by some constant $B \geq 0$; and let

$$\mathcal{F}_H^\delta(B) = \{f \in \mathcal{F}_H(B) : |f_+ - f_-| > \delta\},$$

for some $\delta \geq 0$ be a similar Hölder-type class of functions whose discontinuity at zero exceeds δ in absolute magnitude. We then assume that

$$(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^0(B_T) \equiv \mathcal{F}, \tag{2.4}$$

where B_Y and B_T are constants chosen explicitly by the researcher based on her subject knowledge about the respective application. Intuitively, small values of these smoothness constants mean that the respective functions are "close" to being linear on either side of the cutoff, whereas for larger values the functions can be increasingly "wiggly". Without explicit bounds on the smoothness of μ_Y and μ_T, it is generally not possible to conduct inference on θ that is both valid and informative, even in large samples (Kamat, 2018; Bertanha and Moreira, 2018). This is because, roughly speaking, without such restrictions the true data generating process would always be arbitrarily close, in some appropriate sense, to one in which the parameter of interest takes an arbitrary value on the real line.

In addition to imposing smoothness on μ_Y and μ_T, the structure of (2.4) has two further implications. First, since $\mathcal{F}$ is a Cartesian product of two function classes, we rule out cross-restrictions between the shapes of μ_Y and μ_T. This seams reasonable for empirical applications in economics. Second, since we impose $\mu_T \in \mathcal{F}_H^0(B_T)$ instead of $\mu_T \in \mathcal{F}_H(B_T)$, we have $\tau_T \neq 0$. This is a technicality that is only required for $\theta = \tau_Y/\tau_T$ to be well-defined. Our setup still explicitly allows τ_T to be arbitrarily close to zero.

2.4. **Support of the Running Variable.** Conditional expectation functions are only well-defined over the support of the conditioning variable. The assumption (2.4) must thus be interpreted with some care if the running variable is discrete, or more generally such that there are gaps in its support. In such cases, the condition is formally understood to mean that there exists a pair of functions $(\mu_Y, \mu_T) \in \mathcal{F}$ that are well defined over the entire real line, and is such that $(\mu_Y(X_i), \mu_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i))$ with probability 1.

With this convention, the parameter θ is well-defined through equation (2.1) irrespective of whether the running variable has full support or not. It is point identified as long as the support of X_i contains an open neighborhood around the cutoff, and partially identified otherwise. The latter statement holds because (2.4) implies that $\theta \in \Theta$, where

$$\Theta = \left\{ \frac{m_{Y+} - m_{Y-}}{m_{T+} - m_{T-}} : (m_Y, m_T) \in \mathcal{F} \text{ and } \right. \\ \left. (m_Y(X_i), m_T(X_i)) = (\mu_Y(X_i), \mu_T(X_i)) \text{ with probability } 1 \right\}$$

is typically a true subset of the real line.[2] While it is generally not possible to consistently estimate a parameter that is only partially identified, it is possible to conduct valid inference in this case (Imbens and Manski, 2004). The question whether θ is point or partially identified is immaterial, however, for our CSs described below. See Kolesár and Rothe (2018) for a further discussion of inference with a discrete running variable in SRD designs.

2.5. **Local Linear Estimation.** Local linear regression (Fan and Gijbels, 1996) is arguably most popular empirical strategy for estimation and inference in RD designs. Formally, for some generic dependent variable W_i, that could be equal to Y_i or T_i, for example, the local linear estimator of $\tau_W = \mu_{W+} - \mu_{W-}$, the jump in the generic dependent variable's conditional

[2] More specifically, the identified set Θ always takes one of three general forms: a closed interval $[a_1, a_2]$; the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$, $a_1 < a_2$; or the entire real line. This is because the range of $(m_{Y+} - m_{Y-}, m_{T+} - m_{T-})$ over the functions $(m_Y, m_T) \in \mathcal{F}$ that are such that $P(m_Y(X_i), m_T(X_i)) = (\mu_Y(X_i), \mu_T(X_i)) = 1$ is the Carthesian product of two intervals. We then have that Θ is a closed interval if the range of $m_{T+} - m_{T-}$ does not contain zero. It is equal to the union of two disjoint half-lines if the range of $m_{T+} - m_{T-}$ contains zero, but the range of $m_{Y+} - m_{Y-}$ does not. Finally, Θ is equal to the real line if both of these ranges contain zero.

expectation given the running variable at zero, is

$$\widehat{\tau}_W(h) = e_1' \operatorname*{argmin}_{\beta\in\mathbb{R}^4} \sum_{i=1}^n K(X_i/h)(W_i - \beta'(Z_i, X_i, Z_iX_i, 1))^2, \tag{2.5}$$

where $K(\cdot)$ is a bounded kernel function that is zero outside $[-1, 1]$, $h > 0$ is a bandwidth, and $e_1 = (1, 0, 0, 0)'$ is the first unit vector. With this notation, a natural point estimator of θ is given by $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$, for example.

Estimators of the form in (2.5) are the building blocks of our honest CSs described below, and we refer to such estimators $\widehat{\tau}_W(h)$ as *SRD-type estimators of* τ_W in the following, as they are the usual estimators in the context of a hypothetical SRD in which W_i is the outcome variable. We exploit repeatedly that these estimators can be written as a weighted average of the realizations of the respective dependent variable, with weights that depend on the data through the realizations of the running variable only:

$$\begin{aligned}
&\widehat{\tau}_W(h) = \sum_{i=1}^n w_i(h)W_i, \quad w_i(h) = w_{i,+}(h) - w_{i-}(h),\\
&w_{i,+}(h) = e_1'Q_+^{-1}\tilde{X}_iK(X_i/h)\mathbf{1}\{X_i \geq 0\}, \quad Q_+ = \sum_{i=1}^n K(X_i/h)\tilde{X}_i\tilde{X}_i'\mathbf{1}\{X_i \geq 0\}\\
&w_{i-}(h) = e_1'Q_-^{-1}\tilde{X}_iK(X_i/h)\mathbf{1}\{X_i < 0\}, \quad Q_- = \sum_{i=1}^n K(X_i/h)\tilde{X}_i\tilde{X}_i'\mathbf{1}\{X_i < 0\},
\end{aligned}$$

with $\tilde{X}_i = (1, X_i)'$. Note that we use the same bandwidth on each side of the cutoff to keep the notation simple. It would be conceptually straightforward to work with different bandwidths above and below the treatment threshold; see Section 5.3 for details.

3. BIAS-AWARE ANDERSON-RUBIN-TYPE CONFIDENCE SETS

3.1. **General Approach.** The most commonly used methods for inference in FRD designs are based on approximating the point estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$ by a linear function of $\widehat{\tau}_Y(h)$ and $\widehat{\tau}_T(h)$, and imposing conditions under which the resulting approximation error is "small" in some appropriate sense. Since a linear combination of SRD-type estimators is again a SRD-type estimator, one can then use established arguments to construct a CI for θ. We discuss in Section 6.2 below why such a "linearization" or "delta method" approach is unattractive in many empirically relevant settings.

Our preferred approach to inference avoids linearization errors by directly considering an object that can be estimated by an SRD-type estimator. To describe the general idea, we introduce some notation. For any $c \in \mathbb{R}$, we define the "auxiliary" parameter $\tau_M(c)$ through

$$\tau_M(c) = \mu_{M+}(c) - \mu_{M-}(c), \quad \mu_M(x,c) = \mathbb{E}(M_i(c)|X_i = x), \quad M_i(c) = Y_i - cT_i;$$

so that $\tau_M(c)$ is the jump in the conditional expectation function $\mu_M(x,c)$ of the constructed outcome $M_i(c)$ given the running variable at $x = 0$. This jump can be estimated by an SRD-type estimator of the from

$$\widehat{\tau}_M(h,c) = \sum_{i=1}^{n} w_i(h) M_i(c),$$

and we can use methods developed in Armstrong and Kolesár (2018, 2019) to form a bias-aware CI for $\tau_M(c)$ based on this estimator. To obtain a CS for our *actual* parameter of interest $\theta = \tau_Y/\tau_T$, we simply note that

$$\tau_M(c) = \tau_Y - c \cdot \tau_T \overset{!}{=} 0 \text{ if and only if } c = \theta$$

by linearity of conditional expectations operators. We can thus construct a CS for θ by collecting all values of $c \in \mathbb{R}$ for which the auxiliary CI for $\tau_M(c)$ contains zero. This construction is analogous to Fieller's (1954) method for inference on ratios, which is also similar to Anderson and Rubin (1949) inference in just identified linear instrumental variable models with a single endogenous variable.[3] We refer to the resulting CS as a bias-aware Anderson-Rubin-type CS for θ in the following.

3.2. **Description.** To implement this approach, let $b_M(h,c) = \mathbb{E}(\widehat{\tau}_M(h,c)|\mathcal{X}_n) - \theta_M(c)$ and $s_M(h,c) = \mathbb{V}(\widehat{\tau}_M(h,c)|\mathcal{X}_n)^{1/2}$ denote the conditional bias and standard deviation, respectively, of $\widehat{\tau}_M(h,c)$, given $\mathcal{X}_n = (X_1, \ldots, X_n)'$. Since the weights $w_i(h) = w_{i,+}(h) - w_{i-}(h)$ depend on the data through the realizations of the running variable only, and since $\mu_M(x,c) =$

[3] One can think of delta method CIs as inverting a test that checks if a sample analogue of $\tau_Y/\tau_T - c$ is close to zero, whereas our CSs could be described as inverting a test that chekcs if a sample analogue of $\tau_Y - c\tau_T$ is close to zero. In a standard linear instrumental variable model with outcome Y_i, endogenous variable T_i and instrument Z_i, the parameter of interest is $\text{Cov}(Y_i, Z_i)/\text{Cov}(T_i, Z_i)$, and the Anderson-Rubin test checks if a sample analogue of $\text{Cov}(Y_i - cT_i, Z_i)$ is close to zero. However, this is equivalent to checking a sample analogue of $\text{Cov}(Y_i, Z_i) - c\text{Cov}(T_i, Z_i)$, and this form is analogous to the one we use in this paper.

$\mu_Y(x) - c\mu_T(x)$ and $\sum_{i=1}^n w_{i,+}(h) = \sum_{i=1}^n w_{i-}(h) = 1$, these quantities can be written as

$$\begin{aligned} b_M(h,c) &= \sum_{i=1}^n w_{i,+}(h)(\mu_Y(X_i) - c\mu_T(X_i) - (\mu_{Y+} - c\mu_{T+})) \\ &\qquad - \sum_{i=1}^n w_{i-}(h)(\mu_Y(X_i) - c\mu_T(X_i) - (\mu_{Y-} - c\mu_{T-})), \\ s_M(h,c) &= \left(\sum_{i=1}^n w_i(h)^2 \sigma^2_{M,i}(c)\right)^{1/2}, \end{aligned}$$

where $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ is the conditional variance of $M_i(c)$ given X_i. The conditional bias and standard deviation of $\widehat{\tau}_M(h,c)$ are generally unknown in applications, but can be bounded and estimated, respectively. First, a natural standard error, or estimate of $s_M(h,c)$, is of the form

$$\widehat{s}_M(h,c) = \left(\sum_{i=1}^n w_i(h)^2 \widehat{\sigma}^2_{M,i}(c)\right)^{1/2},$$

where $\widehat{\sigma}^2_{M,i}(c)$ is an appropriate estimate of $\sigma^2_{M,i}(c)$. We discuss in more detail how to construct such estimates below. Second, considering the bias, note that $b_M(h,c)$ depends on (μ_Y, μ_T) through the transformation $\mu_Y - c \cdot \mu_T$ only. Since $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}^0_H(B_T)$, linearity of the second derivatives operator implies that

$$\mu_Y - c \cdot \mu_T \in \mathcal{F}_H(B_Y + |c|B_T).$$

It then follows from Armstrong and Kolesár (2019) that the "worst case" absolute bias over the functions contained in $\mathcal{F}$, for any value of the bandwidth h, can be bounded as follows:

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h,c)| = \overline{b}_M(h,c) \equiv \frac{B_Y + |c|B_T}{2} \cdot \sum_{i=1}^n w_i(h) X_i^2.$$

with the supremum being achieved by a pair of piecewise quadratic functions with second derivatives equal to $(B_Y \cdot \text{sign}(x), B_T \cdot \text{sign}(x))$.[4] We now write the t-statistic for $\widehat{\tau}_M(h,c)$ as

$$\frac{\widehat{\tau}_M(h,c) - \tau_M(c)}{\widehat{s}_M(h,c)} = \frac{\widehat{\tau}_M(h,c) - \tau_M(c) - b_M(h,c)}{\widehat{s}_M(h,c)} + \frac{b_M(h,c)}{\widehat{s}_M(h,c)}. \tag{3.1}$$

[4] Note that the bound $\overline{b}_M(h,c)$ on the conditional bias of $\widehat{\tau}_M(h,c)$ may not be sharp if this pair of piecewise quadratic functions is not a feasible candidate for (μ_Y, μ_T). Since T_i is binary, and the range of any candidate for μ_T therefore must be a subset of the unit interval, this might be the case if $h > (2B_T)^{-1/2}$, as then there is no function μ_T with $\mu_T''(x) = B_T \cdot \text{sign}(x)$ and $\mu_T(x) \in [0,1]$ for all $x \in [-h,h]$. A similar point applies if the support of Y_i happens to be bounded.

Under standard regularity conditions, a Central Limit Theorem (CLT) implies that the first term on the right hand side of the previous equation is approximately standard normal conditional on $\mathcal{X}_n$ in large samples. The second term, on the other hand, is bounded in absolute value by

$$\widehat{r}_M(h,c) = \frac{\overline{b}_M(h,c)}{\widehat{s}_M(h,c)},$$

the "worst case" bias to standard error ratio. For every $c \in \mathbb{R}$ and bandwidth $h > 0$, we can thus construct an auxiliary CI for the pseudo parameter $\tau_M(c)$ as

$$\left(\widehat{\tau}_M(h,c) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}_M(h,c)) \cdot \widehat{s}_M(h,c)\right), \tag{3.2}$$

where the critical value $\mathrm{cv}_{1-\alpha}(r)$ is the $1-\alpha$ quantile of the $|N(r,1)|$ distribution, the distribution of the absolute value of a normal random variable with mean r and variance 1. In statistical software packages, this critical value can easily be computed as the square root of the $(1-\alpha)$-quantile of a non-central χ^2 distribution of one degree of freedom and non-centrality parameter r^2.

This construction is analogous to that of the bias-aware CI of Armstrong and Kolesár (2019, 2018) for SRD designs. Since it is conditional on the realizations of the running variable, it is valid irrespective of whether the distribution of the latter is continuous or discrete; and since it takes into account the exact conditional bias, the CI is also valid for any choice of bandwidth, including fixed ones that do not depend on the sample size. The bandwidth

$$\widehat{h}_{\mathrm{opt}}(c) = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}(\widehat{r}_M(h,c)) \cdot \widehat{s}_M(h,c)$$

minimizes the length of the auxiliary CI, and thus maximizes the efficiency of inference. The auxiliary CI can be shown to remain honest with this choice under standard conditions. Roughly, this is because, as long as the standard error satisfies a mild uniform consistency property, $\widehat{h}_{\mathrm{opt}}(c)$ is a consistent estimate of the infeasible optimal bandwidth

$$h_{\mathrm{opt}}(c) = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}(r_M(h,c)) \cdot s_M(h,c), \quad r_M(h,c) = \frac{\overline{b}_M(h,c)}{s_M(h,c)}$$

which depends on neither the outcomes $M_i(c)$ nor the functions μ_Y and μ_T. Our proposed CS for our actual parameter of interest θ is then given by the collection of all $c \in \mathbb{R}$ such that the auxiliary CI in (3.2) with the optimized bandwidth $\widehat{h}_{\text{opt}}(\mathrm{c})$ contains zero:

$$\mathcal{C}_{\text{ar}}^{\alpha}=\left\{c:|\widehat{\tau}_M(\widehat{h}_{\text{opt}}(c),c)|<\text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_{\text{opt}}(c),c))\cdot\widehat{s}_M(\widehat{h}_{\text{opt}}(c),c))\right\}. \tag{3.3}$$

Note that $\mathcal{C}_{\text{ar}}^{\alpha}$ is not necessarily an interval, although it will take this form in many applications. We provide details on its shape and computation in Section 5 below.

3.3. **Standard Errors.** There are several ways to construct the estimates $\widehat{\sigma}_{M,i}^2(c)$ of $\sigma_{M,i}^2(c)=\mathbb{V}(M_i(c)|X_i)$ that enter the standard error $\widehat{s}_M(h,c)$. Here we focus on a variant of the nearest-neighbor approach of Abadie et al. (2014). Specifically, let $R>0$ be a small, fixed integer, denote the rank of $|X_j-X_i|$ among the elements of the set $\{|X_s-X_i|:s\in\{1,\dots,n\}\setminus\{i\},X_sX_i>0\}$ by $r(j,i)$, let $\mathcal{R}_i$ be the set of indices such that $r(j,i)\leq Q_i$, where Q_i is the smallest integer such that $\mathcal{R}_i$ contains at least R elements corresponding to at least two distinct realizations of the running variable, and let $R_i=\#\mathcal{R}_i$ be the resulting cardinality of $\mathcal{R}_i$. If every realization of X_i is unique, then $R=Q_i=R_i$, and $\mathcal{R}_i$ is simply the set of unit i's R nearest neighbors' indices. With ties in the data, multiple units could be equally far from unit i, and hence R_i could be greater than R.

If the realization of X_i is observed at least R times, we then simply put $\widehat{\sigma}_i^2(c)$ equal to the sample variance of the outcomes of the units with that realization; and otherwise we put $\widehat{\sigma}_{M,i}^2(c)$ equal to a scaled version of the squared difference between $M_i(c)$ and its best linear predictor given its R_i nearest neighbors. That is,

$$\widehat{\sigma}_{M,i}^2(c)=\begin{cases}\dfrac{1}{R_i-1}\displaystyle\sum_{j:X_j=X_i}\left(M_j(c)-\frac{1}{R_i}\sum_{l:X_l=X_i}M_l(c)\right)^2 & \text{if } \#\{j:X_j=X_i\}\geq R,\\ \dfrac{R_i}{R_i+H_i}\left(M_i(c)-\widehat{M}_i(c)\right)^2 & \text{else,}\end{cases}$$

with

$$\widehat{M}_i(c)=\tilde{X}_i(\sum_{j\in\mathcal{R}_i}\tilde{X}_j'\tilde{X}_j)^{-1}\sum_{j\in\mathcal{R}_i}\tilde{X}_j'M_j(c),\quad H_i=R_i\tilde{X}_i(\sum_{j\in\mathcal{R}_i}\tilde{X}_i'\tilde{X}_j)^{-1}\tilde{X}_i{}',\quad \tilde{X}_i=(1,X_i)'.$$

The role of the adjustment term H_i is to ensure that $\widehat{\sigma}_{M,i}^2(c)$ is approximately unbiased in large samples. Its form follows from standard arguments for out-of-sample forecast error evaluation in linear regression models. In our simulations and empirical application in this paper, we implement the estimator $\widehat{\sigma}_{M,i}^2(c)$ with $R=5$.

The construction of $\widehat{\sigma}_{M,i}^2(c)$ differs from the conventional nearest-neighbor approach of Abadie et al. (2014), which is often recommend in the RD literature (e.g. Calonico et al., 2014;

Armstrong and Kolesár, 2018, 2019). The conventional estimator replaces the best linear prediction $\widehat{M_i}(c)$ with the sample average of the outcomes of the R nearest neighbors of unit i, and sets $H_i = 1$. While this works well in many simulations and empirical applications, it formally does not lead to a standard error that is consistent uniformly over $\mathcal{F}$. This is because the leading bias of the conventional estimator is proportional to the first derivative of $\mu_M(\cdot, c)$ at X_i, which is unbounded over $\mathcal{F}$. By using a best linear prediction, the bias of $\widehat{\sigma}^2_{M,i}(c)$ becomes proportional to the second derivative of $\mu_M(\cdot, c)$, which is bounded in absolute value over $\mathcal{F}$ by $B_Y + |c|B_T$.

4. THEORETICAL PROPERTIES

Our main theoretical result in this paper is that $\mathcal{C}^{\alpha}_{\text{ar}}$ is an honest CS for θ with respect to $\mathcal{F}$, as defined in (2.2), under rather weak conditions. We first derive this result under general "high level" assumptions, and then verify these conditions for two specific setups.

Assumption 1. *(i) The data* $\{(Y_i, T_i, X_i), i = 1, \ldots, n\}$ *are an i.i.d. sample from a fixed population; (ii)* $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^q|X_i = x)$ *exists and is bounded uniformly over* $x \in$ *supp*(X_i) *and* $(\mu_Y, \mu_T) \in \mathcal{F}$ *for some* $q > 2$ *and every* $c \in \mathbb{R}$*; (iii)* $\mathbb{V}(M_i(c)|X_i = x)$ *is bounded and bounded away from zero uniformly over* $x \in$ *supp*(X_i) *and* $(\mu_Y, \mu_T) \in \mathcal{F}$ *for every* $c \in \mathbb{R}$*; (iv) the kernel function* K *is a continuous, unimodal, symmetric density function that is equal to zero outside some compact set, say* $[-1, 1]$.

Assumption 1 is standard in the literature on local linear regression. Part (i) could be weakened to allow for certain forms of dependent sampling, such as cluster sampling. Parts (ii)–(iii) are standard moment conditions. Since $M_i(c) = Y_i - cT_i$ and T_i is binary, these conditions mainly restrict the conditional moments of the outcome variable. Part (iv) is satisfied by all kernel functions commonly used in applied RD analysis, such an the triangular or the Epanechnikov kernels.

Assumption 2. *(i)* $w_{\text{ratio}}(\widehat{h}_{\text{opt}}(c)) = w_{\text{ratio}}(h_{\text{opt}}(c))(1 + o_P(1))$ *, and* $w_{\text{ratio}}(h_{\text{opt}}(c)) = o_P(1)$ *uniformly over* $\mathcal{F}$*, where* $w_{\text{ratio}}(h) = \max_{i=1,\ldots,n} w_i(h)^2 / \sum_{i=1}^n w_i(h)^2$*; and (ii)* $\widehat{s}_M(\widehat{h}_{\text{opt}}(c), c) = s_M(h_{\text{opt}}(c), c)(1 + o_P(1))$ *uniformly over* $\mathcal{F}$.

Assumption 2 is a high-level condition that applies to a wide range of settings. We discuss more "low level" conditions for its validity below. Part (i) implies that the magnitude of each of the weights $w_i(\widehat{h}_{\text{opt}}(c))$ is arbitrarily small relative to the others' in large samples. Together with the moment conditions in Assumptions 1, this ensures that a CLT applies to an appropriately standardized version of the estimator $\widehat{\tau}_M(\widehat{h}_{\text{opt}}(c), c)$. Part (ii) states that the standard error $\widehat{s}_M(\widehat{h}_{\text{opt}}(c), c)$ is consistent for the true standard deviation $s_M(h_{\text{opt}}(c), c)$ at the infeasible optimal bandwidth, uniformly over the function class $\mathcal{F}$. We then have the following result.

Theorem 1. *Suppose that Assumptions 1–2 hold. Then*

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\theta \in \mathcal{C}_{\text{ar}}^{\alpha}) \geq 1-\alpha.$$

Again, we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on the functions μ_Y and μ_T implicit in our notation. The theorem shows that we can expect $\mathcal{C}_{\text{ar}}^{\alpha}$ to have accurate coverage in finite samples if a CLT applies to our estimates of $\tau_M(c)$, and we have a uniformly consistent standard error. For empirical practice, it is important to also give more low-level conditions for the validity of our approach. The two following ones cover the case of a discrete and a continuously distributed running variable, respectively.

Assumption LL1. The support of X_i is countable.

Assumption LL2. (i) The running variable X_i is continuously distributed with density f_X that is bounded and bounded away from zero over an open neighborhood of the cutoff; (ii) $\mathbb{V}(M_i(c)|X_i = x)$ is Lipschitz continuous uniformly over $x \in \mathbb{R}$ for every $c \in \mathbb{R}$; and (iii) $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^4|X_i = x)$ exists and is uniformly bounded over $x \in \mathbb{R}$ for every $c \in \mathbb{R}$.

Theorem 2. *Suppose that Assumption 1 and either Assumption LL1 or Assumption LL2 are satisfied. Then Assumption 2 holds with the standard error formula given in Section 3.3 above.*

It is also interesting to consider further implications of these two low-level conditions for the behavior of $\mathcal{C}_{\text{ar}}^{\alpha}$. Under Assumption LL1, it is easy to see $h_{\text{opt}}(c)$ approaches a positive constant as the sample size tends to infinity, and thus $\mathcal{C}_{\text{ar}}^{\alpha}$ does not shrink to a singleton asymptotically. This is to be expected given the analysis in Section 2.5 above. Under Assumption LL2, it follows from results in Armstrong and Kolesár (2019) that

$$h_{\text{opt}}(c) = \left(\frac{1}{n} \cdot \frac{\int K(u)^2 du}{(\int K(u)u^2 du)^2} \cdot \frac{\sigma_{M+}^2(c) + \sigma_{M-}^2(c)}{(B_Y + |c|B_T)^2 f_X(0)} \cdot r_*^2\right)^{1/5}$$

where $r_* = \operatorname{argmin}_{r>0} r^{-1/5}\text{cv}_{1-\alpha}(r)$. Their results also imply that $\text{cv}_{1-\alpha}(r_M(h_{\text{opt}}(c), c)) = \text{cv}_{1-\alpha}(r_*) + o_P(1)$. For the common choice $\alpha = 0.05$, for example, $r_* \approx 0.53$, and the corresponding critical value $\mathrm{cv}_{1-\alpha}(r_*) \approx 2.21$ is slightly larger than the usual critical value of 1.96 based on the normal distribution.

5. EXTENSIONS AND REMARKS

5.1. **Improving Finite-Sample Coverage Accuracy.** When constructing $\mathcal{C}^{\alpha}_{\mathrm{ar}}$, using the bandwidth $\widehat{h}_{\mathrm{opt}}(c)$ that minimizes the length of the auxiliary CI in (3.2) is attractive, as it balances bias and standard error in way that is optimal for inference. In finite samples, however, this choice can potentially cause some distortions. Too see why, recall, as discussed after Assumption 2, that for any bandwidth h asymptotic normality of $\widehat{\tau}_M(h,c) = \sum_{i=1}^n w_i(h)M_i(c)$ follows from a CLT under the assumption that $w_{\mathrm{ratio}}(h) = o_P(1)$. Normality should thus be a "good" approximation in finite samples if $w_{\mathrm{ratio}}(h)$ is "close" to zero. For $B_Y + |c|B_T$ large, however, the bandwidth $\widehat{h}_{\mathrm{opt}}(c)$ can be rather small. This in turn leads to weights $w_i(\widehat{h}_{\mathrm{opt}}(c))$ concentrating on a few observations close to the cutoff, to $w_{\mathrm{ratio}}(\widehat{h}_{\mathrm{opt}}(c))$ becoming large, and to $\widehat{\tau}_M(\widehat{h}_{\mathrm{opt}}(c), c)$ effectively behaving like a sample average of a small number of observations. CLT approximations could then be inaccurate in practice.

To address this issue, one can impose a lower bound on the bandwidth used in the construction of the auxiliary CI in (3.2), where the bound is chosen such that the resulting value of $w_{\mathrm{ratio}}(\cdot)$ remains below some reasonable threshold. Specifically, one can consider replacing $\widehat{h}_{\mathrm{opt}}(c)$ in (3.2) with

$$\widehat{h}^*_{\mathrm{opt}}(c) = \max\left\{\widehat{h}_{\mathrm{opt}}(c), h_{\min}(\eta)\right\}, \quad h_{\min}(\eta) = \min\left\{h : w_{\mathrm{ratio}}(h) < \eta\right\},$$

where $\eta > 0$ is a small constant. Note that using $\widehat{h}^*_{\mathrm{opt}}(c)$ instead of $\widehat{h}_{\mathrm{opt}}(c)$ does not affect the validity of the auxiliary CI in (3.2), as the latter is valid for *any* choice of bandwidth. It is easy to see that in standard setups like those described by Assumptions LL1 or LL2 the lower bound on the bandwidth never binds asymptotically, but in simulations we found that it can potentially improve the finite-sample coverage of our CSs.

To give some intuition for what could be a plausible choice of η, suppose that $\mathcal{X}_n = \{\pm.02, \pm.04, \ldots, \pm 1\}$, that $K(t) = (1-|t|)\mathbf{1}\{|t| < 1\}$ is the triangular kernel, and that $h = 1$. For such a setting a CLT approximation to the distribution of $\widehat{\theta}(h,c)$ should be reasonably accurate in finite samples, as the estimator corresponds to a weighted linear regression with 50 observations on each side of cutoff. Since $w_{\mathrm{ratio}}(h) \approx .075$ in this case, choosing $\eta \in [0.05, 0.1]$ seems reasonable; we actually use $\eta = .1$ in our simulations.

As $h^*_{\mathrm{opt}}(c) \geq h_{\mathrm{opt}}(c)$, in finite samples this bandwidth potentially over-smooths the data relative to the one that would be asymptotically optimal for inference. By using it, we accept the cost of a larger finite-sample bias in return for normality being a better approximation in finite samples. This idea could also be used in other settings where the finite sample accuracy of inference faces a similar bias-vs-normality trade-off, such as inference on average treatment effects under unconfoundedness with limited overlap (e.g. Rothe, 2017).

5.2. **Shape and Computation of CS.** It is difficult to give formal results regarding the shape of our proposed CS $\mathcal{C}_{\text{ar}}^{\alpha}$ in finite samples. To see why, recall the definition from (3.3) that $c \in \mathcal{C}_{\text{ar}}^{\alpha}$ if and only if

$$|\widehat{\tau}_M(\widehat{h}_{\text{opt}}(c), c)| - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_{\text{opt}}(c), c)) \cdot \widehat{s}_M(\widehat{h}_{\text{opt}}(c), c) < 0. \tag{5.1}$$

The left-hand-side of (5.1) depends on c both directly and indirectly through the bandwidth $\widehat{h}_{\text{opt}}(c)$. Finding the set of values of c that satisfy the above inequality is therefore generally not possible analytically. In practice, we compute $\mathcal{C}_{\text{ar}}^{\alpha}$ as follows. For every $c \in \mathbb{R}$, let $p(c) = \sup\{\alpha : c \in \mathcal{C}_{\text{ar}}^{\alpha}\}$, so that $1 - p(c)$ is the smallest nominal level at which our CS contains c. We then calculate function $p(c)$ exactly over a grid $\{c_1, \ldots, c_m\}$, and approximate it at intermediate points through piecewise linear interpolation. Denoting the resulting approximation function by $\widetilde{p}(c)$, we then compute a numerical approximation to $\mathcal{C}_{\text{ar}}^{\alpha}$ as $\widetilde{\mathcal{C}}_{\text{ar}}^{\alpha} = \{c : \widetilde{p}(c) < \alpha\}$. In simulations and empirical applications, we find that $\mathcal{C}_{\text{ar}}^{\alpha}$ almost always takes one of three general forms: a closed interval $[a_1, a_2]$; the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$, $a_1 < a_2$; or the entire real line.

While we do not have a formal result regarding the shape of $\mathcal{C}_{\text{ar}}^{\alpha}$ in finite samples, one can prove such a result for a variant of our CS that uses a bandwidth that does not depend on the parameter value under consideration. The result suggests that as long as the dependence of the terms on the left-hand-side of (5.1) on the value of c dominates the indirect dependence through the bandwidth $\widehat{h}_{\text{opt}}(c)$, our actual CS $\mathcal{C}_{\text{ar}}^{\alpha}$ should also take one of the three general shapes mention in the lemma.

Lemma 1. *Let $\mathcal{C}_{\text{ar}}^{\alpha}(h) = \{c : |\widehat{\tau}_M(h, c)| < \text{cv}_{1-\alpha}(\widehat{r}_M(h, c)) \cdot \widehat{s}_M(h, c))\}$ be a variant of $\mathcal{C}_{\text{ar}}^{\alpha}$, where $h > 0$ is an arbitrary bandwidth that does not depend on c. Then $\mathcal{C}_{\text{ar}}^{\alpha}(h) = [a_1, a_2]$, or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = (-\infty, a_1] \cup [a_2, \infty)$, or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = (-\infty, \infty)$.*

5.3. **Side-Specific Bandwidths and Smoothness Bounds.** So far, our local linear regressions use the same bandwidth on either side of the cutoff, and we have imposed that the second derivatives of μ_Y and μ_T are bounded in absolute value by the same respective constant on either side of the cutoff. Both conditions can easily be relaxed. Regarding the bandwidth, however, it follows from results in Armstrong and Kolesár (2019) that there is little to be gained from allowing for side-specific values unless there is a substantial shift in $\mathbb{V}(M_i(c)|X_i = x)$ at the cutoff.

Allowing for different smoothness constants on each side of the cutoff could be of more practical importance, at least for some applications. If the running variable measures time, for example, and the cutoff represents the introduction of a policy, researchers might know in advance that the shape of conditional expected outcomes and/or conditional treatment probabilities become much more "erratic" after the reform. For such scenarios, one could define a more general Hölder-type class as

$$\mathcal{F}_H(B_+, B_-) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_1''\|_\infty \leq B_+, \|f_0''\|_\infty \leq B_-\},$$

define the class $\mathcal{F}_H^\delta(B_+, B_-)$ analogously, and then seek to obtain CSs that are honest uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_{Y,+}, B_{Y-}) \times \mathcal{F}_H^0(B_{T,+}, B_{T-})$. It is easy to see that this would affect our analysis above by changing the explicit expression of the bound on the absolute value of the conditional bias of $\widehat{\tau}_m(h, c)$ to

$$\bar{b}_M(h, c) = \frac{B_{Y,+} + |c|B_{T,+}}{2} \cdot \sum_{i=1}^n w_{i,+}(h)X_i^2 - \frac{B_{Y-} + |c|B_{T-}}{2} \cdot \sum_{i=1}^n w_{i-}(h)X_i^2,$$

but every other step of our derivation would remain the same. Of course, in this case it would make sense to consider different bandwidths on either side of the cutoff.

6. COMPARISON WITH OTHER METHODS

6.1. **Accounting for Smoothing Bias.** Our construction of $\mathcal{C}_{\text{ar}}^\alpha$ involves creating a bias-aware CI for the auxiliary parameter $\tau_M(c)$ based on the estimator $\widehat{\tau}_M(h, c)$. Since the latter is an SRD-type estimator, such an auxiliary CI for $\tau_M(c)$ could in principle also be obtained through one of the several alternative approaches to handling smoothing bias from the literature on SRD inference. Armstrong and Kolesár (2019, Section 4) compare the theoretical properties of such methods to bias-aware inference in a more general context. We now briefly review their main findings.

Adapting notation appropriately to our context, the approaches discussed in Armstrong and Kolesár (2019, Section 4) are: *(i)* A naive approach that simply ignores the presence of the bias term. In practice, with this approach the bandwidth is often chosen as an estimate $\hat{h}_{\text{mse}}(c)$ of the value $h_{\text{mse}}(c)$ that minimizes the pointwise asymptotic MSE of $\widehat{\tau}_M(h, c)$ at the "true" function $\mu_Y - c \cdot \mu_T$, see Imbens and Kalyanaraman (2012); *(ii)* Undersmoothing, or using a "small" bandwidth that makes the "bias to standard error" ratio asymptotically negligible. This type of approach was considered by Feir et al. (2016) in the context of FRD Anderson-Rubin inference; see Section 6.4 below for details. In practice, undersmoothing bandwidths are often chosen in an ad-hoc way as $\widehat{h}_{\text{mse}}(c) \cdot n^{-\epsilon}$ for some $\epsilon > 0$; *(iii)* Robust bias correction (Calonico et al., 2014), which involves constructing a new estimate of $\theta_M(c)$ as the difference between $\widehat{\tau}_M(\hat{h}_{\text{mse}}(c), c)$ and an estimate of its bias, obtained via local quadratic regression with another estimated "pilot bandwidth", and adjusting the standard error appropriately.

Armstrong and Kolesár (2019) argue that these three methods have substantial shortcomings relative to bias-aware inference. One is their reliance on the empirical bandwidth selector $\hat{h}_{\text{mse}}(c)$. The issue is that the bandwidth $h_{\text{mse}}(c)$ can be very large in settings where the underlying functions are highly nonlinear, which in turn leads to large smoothing biases in finite samples. The estimator $\hat{h}_{\text{mse}}(c)$ therefore involves a regularization step that is supposed to prevent extreme bandwidth values, but in practice the result is often unstable and depends critically on the values of tuning parameters that are difficult to pick. Another issue is that, even with reasonable infeasible bandwidth choices, none of the three methods lead to honest CIs. Naive and undersmoothing CIs generally undercover in finite samples, as they are not correctly centered. The undercoverage of robust bias correction CIs is typically less pronounced,[5] but they are inefficient and tend to be much longer than bias-aware CIs.

On the other hand, the bias-aware auxiliary CI for $\tau_M(c)$, is highly efficient, in the sense that no other approach can produce substantially shorter CIs and still maintain uniform coverage. It is also valid when the running variable is discrete, and comes with a straightforward way to select the bandwidth. The bias-aware approach thus seems to be the most appropriate one to construct the auxiliary CI for $\tau_M(c)$.

6.2. **Delta Method Inference.** In addition to the issue of how to handle smoothing bias, one can also consider alternatives to the Anderson-Rubin-type CS construction that we use in this paper. Indeed, the arguably most widely used methods for inference in FRD designs are based on linearizing the point estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$. In particular, standard arguments yield that $\widehat{\theta}(h) - \theta = \widetilde{\theta}^L(h) + \widetilde{\theta}^R(h)$, where

$$\widetilde{\theta}^L(h) = \frac{\widehat{\tau}_Y(h) - \tau_Y}{\tau_T} - \frac{\tau_Y(\widehat{\tau}_T(h) - \tau_T)}{\tau_T^2}, \quad \widetilde{\theta}^R(h) = \widehat{\tau}_Y(h) \int_{\tau_T}^{\widehat{\tau}_T(h)} \frac{(\widehat{\tau}_T(h) - t)^2}{t^3} dt.$$

[5] Kamat (2018) shows that the robust bias correction CIs based on infeasible MSE-optimal bandwidths are honest with respect to a smaller function class that puts bounds on the absolute value of the third derivatives, instead of only the second.

Under certain regularity and bandwidth conditions, the term $\widetilde{\theta}^L(h)$ is the leading term in this expansion of $\widehat{\theta}(h)-\theta$, and $\widetilde{\theta}^R(h)$ is an asymptotically negligible remainder term. The first order asymptotic properties of $\widehat{\theta}(h)$ then coincide with those of $\widetilde{\theta}^L(h)$, which is again a SRD-type estimator:

$$\widetilde{\theta}^L(h)=\sum_{i=1}^{n} w_i(h)U_i, \quad U_i=\frac{Y_i-\tau_Y}{\tau_T}-\frac{\tau_Y(T_i-\tau_T)}{\tau_T^2}.$$

Inference can then in principle be carried out using any of the methods to handling smoothing bias in SRDs described above, and we refer to any CI based on such a construction as a *delta method CI*. Given our discussion in the previous subsection, bias-aware delta method CIs have clear advantages over competitors based on other bias handling schemes, and we discuss this approach and how it relates to ours in more detail in the next subsection.[6]

A more important principle issue for any type of delta method CI, however, is that the basic condition needed for validity of such an approach, namely that $\widetilde{\theta}^R(h)$ is asymptotically negligible relative to $\widetilde{\theta}^L(h)$, are not innocuous. Indeed, they generally rule out weakly identified settings in which τ_T is close to zero, and settings with discrete running variables, irrespective of the type of method chosen to control the bias. To see the first point, note that in order for any delta method CI to be honest with respect to $\mathcal{F}$, the term $\widetilde{\theta}^R(h)$ must be of smaller order than $\widetilde{\theta}^L(h)$ not only at the "true" function pair (μ_Y,μ_T), but uniformly over all $(\mu_Y,\mu_T)\in\mathcal{F}$. However, this is not possible since $\mathcal{F}$ contains functions under which the jump τ_T in the treatment probability at the cutoff can be arbitrarily close to zero, which means that

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}}\left|\widetilde{\theta}^R(h)\right|=\infty, \tag{6.1}$$

unless $\widehat{\tau}_Y(h)=0$. Any delta method CI can therefore potentially break down in this case.[7] To circumvent this issue, one could work with a reduced honesty requirement that only requires correct coverage over a smaller function class in which the value of τ_T is bounded away from zero. Such a criterion could be sufficient if the analyst knows in advance that τ_T is well-separated from zero for a particular application.

[6] While the constructed variable U_i is unobserved, any of the methods for handling smoothing bias can be made feasible by instead using an estimate $\widehat{U}_i=(Y_i-\widehat{\tau}_Y)/\widehat{\tau}_T-\widehat{\tau}_Y(T_i-\widehat{\tau}_T)/\widehat{\tau}_T^2$, with $\widehat{\tau}_Y$ and $\widehat{\tau}_T$ some suitable consistent estimator of τ_Y and τ_T, respectively.

[7] Feir et al. (2016) also point out the failure of delta method inference under weak identification, but do so using different technical arguments. Specifically, they show that delta method CIs with undersmoothing do not have correct asymptotic coverage under pointwise asymptotics when τ_T tends to zero with the sample size at an appropriate rate.

To see that the delta method CIs are generally not valid if the running variable is discrete, even if we only require honesty with respect to a function class that rules out weak identification, note that we can also write the remainder term $\widetilde{\theta}^R(h)$ as

$$\widetilde{\theta}^R(h) = \frac{\widehat{\tau}_Y(h)(\widehat{\tau}_T(h) - \tau_T)^2}{\widehat{\tau}_T^*(h)^3},$$

with $\widehat{\tau}_T^*(h)$ is an intermediate value between τ_T and $\widehat{\tau}_T(h)$. Consistent estimation of τ_T – and τ_Y, for that matter – is generally not possible with a discrete running variable. The term $\widetilde{\theta}^R(h)$ must therefore be of the same order as $\widetilde{\theta}^L(h)$ in large samples, and thus cannot be ignored for the purpose of inference on θ. As pointed out above, discrete running variables are ubiquitous in empirical applications, and they do not constitute a conceptual issue for our bias-aware Anderson-Rubin CSs.

6.3. **Bias-Aware Delta Method Inference.** In this subsection, we formally describe bias-aware delta method CIs for FRD designs, and compare them to ours based on the Anderson-Rubin principle. Bias-aware delta method CIs are obtained by applying the techniques of Armstrong and Kolesár (2018, 2019) to the leading term $\widetilde{\theta}^L(h)$. In order to derive formal results, we of course have to make assumptions that ensure that the delta method approach is valid in the first place. Specifically, we assume that Assumption LL2 holds, which implies, among other things, that the running variable is continuously distributed; and that $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^\delta(B_T) \equiv \mathcal{F}^\delta$ for some $\delta > 0$ to rule out weak identification.

To keep the notation similar to that in Section 3, we write $\widehat{\tau}_U(h)$ instead of $\widetilde{\theta}^L(h)$ in the following, and let $b_U(h) = \mathbb{E}(\widehat{\tau}_U(h)|\mathcal{X}_n)$ and $s_U(h) = \mathbb{V}(\widehat{\tau}_U(h)|\mathcal{X}_n)^{1/2}$ denote its conditional bias and standard deviation, respectively. Exploiting linearity, we write these quantities as

$$b_U(h) = \sum_{i=1}^n w_{i,+}(h)(\mu_U(X_i) - \mu_{U+}) - \sum_{i=1}^n w_{i-}(h)(\mu_U(X_i) - \mu_{U-}),$$
$$s_U(h) = \left(\sum_{i=1}^n w_i(h)^2 \sigma_{U,i}^2\right)^{1/2},$$

where $\mu_U(x) \equiv \mathbb{E}(U_i|X_i = x) = (\mu_Y(x) - \tau_Y)/\tau_T - \tau_Y(\mu_T(x) - \tau_T)/\tau_T^2$ is a linear combination of the functions μ_Y and μ_T, and $\sigma_{U,i}^2 = \mathbb{V}(U_i|X_i)$ is the conditional variance of U_i given X_i. Since the bias depends on (μ_Y, μ_T) through the function $\mu_U \in \mathcal{F}_H(B_Y/|\tau_T| + |\tau_Y|B_T/\tau_T^2)$ only, its "worst case" magnitude over the functions contained in $\mathcal{F}$, for any value of the bandwidth h, is given by

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_U(h)| = \bar{b}_U(h) \equiv \frac{1}{2}\left(\frac{B_Y}{|\tau_T|} + \frac{|\tau_Y|B_T}{\tau_T^2}\right)\sum_{i=1}^n w_i(h)X_i^2.$$

This bound on the bias involves the unknown population quantities τ_Y and τ_T, and thus needs to be estimated. An obvious candidate for such an estimate is

$$\widehat{\bar{b}}_U(h) = \frac{1}{2}\left(\frac{B_Y}{|\widehat{\tau}_T|} + \frac{|\widehat{\tau}_Y|B_T}{\widehat{\tau}_T^2}\right)\sum_{i=1}^n w_i(h)X_i^2,$$

where $\widehat{\tau}_Y = \widehat{\tau}_Y(g_Y)$ and $\widehat{\tau}_T = \widehat{\tau}_T(g_T)$ are local linear estimates based on some preliminary bandwidths g_Y and g_T, respectively. Under the regularity conditions we consider in this subsection, the preliminary bandwidths can be chosen such $\widehat{\tau}_Y$ and $\widehat{\tau}_T$ are uniformly consistent over $\mathcal{F}^\delta$, converging with the usual optimal rate of $n^{-2/5}$. Using such preliminary estimates, one can also construct a feasible standard error of the form

$$\widehat{s}_U(h) = \left(\sum_{i=1}^n w_i(h)^2\widehat{\sigma}^2_{\widehat{U},i}\right)^{1/2}$$

based on estimates $\widehat{U}_i = (Y_i - \widehat{\tau}_Y)/\widehat{\tau}_T - \widehat{\tau}_Y(T_i - \widehat{\tau}_T)/\widehat{\tau}_T^2$ of the realizations of the U_i. For every value of h, we then define the bias-aware delta method CI for θ with nominal level $1-\alpha$ as

$$\mathcal{C}^\alpha_\Delta(h) = \left(\widehat{\theta}(h) \pm \mathrm{cv}_{1-\alpha}\left(\frac{\widehat{\bar{b}}_U(h)}{\widehat{s}_U(h)}\right)\cdot\widehat{s}_U(h)\right),$$

The bandwidth value that minimizes the length of this CI is

$$\widehat{h}_U = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}\left(\widehat{\bar{b}}_U(h)/\widehat{s}_U(h)\right)\cdot\widehat{s}_U(h),$$

and we write $\mathcal{C}^\alpha_\Delta = \mathcal{C}^\alpha_\Delta(\widehat{h}_U)$ for the CI that corresponding to this bandwidth choice. Results in Armstrong and Kolesár (2019) then imply that this CS is honest with respect to $\mathcal{F}^\delta$, and that it is near-optimal, in the sense that no other method can substantially improve upon its length asymptotically.

There are two main downsides to bias-aware delta method CIs relative to our bias-aware Anderson-Rubin CSs. First, as mentioned above, validity of any delta method approach to FRD inference requires strong identification and a continuously distributed running variable. Neither is required for the Anderson-Rubin approach. Second, when combined with the delta method, the bias-aware approach does not account for the actual bias of the estimator of interest, but only for that of the leading term in a stochastic approximation. Moreover, even the bound on the approximate bias needs to be estimated. This naturally affects finite-sample coverage properties of the resulting CI. With the Anderson-Rubin approach, on the other hand, smoothing biases are controlled exactly even in finite samples.

The following theorem shows that bias-aware delta method CIs are also not more efficient than our bias-aware Anderson-Rubin-type CSs in settings where the former are valid. Specifically, we show that both procedures have the same nontrivial asymptotic coverage of a drifting parameter that is within an $n^{-2/5}$ neighborhood of θ.

Theorem 3. *Suppose that Assumptions 1 and LL2 hold, and put $\theta^{(n)} = \theta + \kappa \cdot n^{-2/5}$ for some fixed $\kappa \neq 0$. Then*

$$0 < \liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} \mathbb{P}(\theta^{(n)} \in \mathcal{C}^\alpha_{\text{ar}}) = \liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} \mathbb{P}(\theta^{(n)} \in \mathcal{C}^\alpha_{\Delta}) < 1-\alpha.$$

The robustness of bias-aware Anderson-Rubin CSs against weak identification and discrete running variables does thus not come with a loss of efficiency relative to the bias-aware delta method CI in the canonical case of a strongly identified RD design with a continuous running variable (as pointed out above, the bias-aware delta method CI is already nearly-efficient in this context). The theorem is analogous to the result that there is no loss of efficiency when using the Anderson-Rubin approach for inference in exactly identified moment condition models relative to one based on a conventional t-test. Note that $n^{-2/5}$ neighborhoods are the appropriate ones to consider here because the length of $\mathcal{C}^\alpha_\Delta$ is $O_P(n^{-2/5})$ uniformly over $\mathcal{F}^\delta$. If we were to consider drifting parameters of the form $\theta_{\text{frd}} + c \cdot n^{-\gamma}$, the asymptotic coverage of both $\mathcal{C}^\alpha_\Delta$ and $\mathcal{C}^\alpha_{\text{ar}}$ would simply be equal to zero for $\gamma < 2/5$, and equal to $1-\alpha$ for $\gamma > 2/5$, irrespective of the value of κ.

6.4. **Comparison with Feir et al. (2016).** In related work, Feir et al. (2016) also propose an Anderson-Rubin-type CSs for FRD inference. The main practical difference is that Feir et al. (2016) do not explicitly account for the bias from their local linear regression steps, and instead assume that the chosen bandwidth is sufficiently small for the bias to be negligible. Formally, in our notation the CS that they propose is

$$\mathcal{C}^\alpha_{\text{ar,fml}}(h) = \left\{c : |\widehat{\tau}_M(h,c)| < q_{1-\alpha/2} \cdot \widehat{s}_M(h,c)\right\}.$$

where q_α is the usual α quantile of the standard normal distribution, and $\widehat{s}_M(h,c)$ is a standard error that is slightly different from ours. We note that Feir et al. (2016) use the same bandwidth for estimating $\widehat{\tau}_M(h, c)$ irrespective of the value of $c \in \mathbb{R}$, which is clearly inefficient as it does not account for the dependence of the bias and variance of $\widehat{\tau}_M(h, c)$ on the value of $c \in \mathbb{R}$. Feir et al. (2016) also do not specify how this bandwidth should be chosen in practice. In their empirical application, they report $\mathcal{C}^{\alpha}_{\mathrm{ar,fml}}(h)$ for a range of bandwidth values. In our simulations below, we study a version of their procedure in which a different bandwidth is chosen for every $c \in \mathbb{R}$ as $\widehat{h}_{\mathrm{mse}}(c) \cdot n^{-1/20}$, where $\widehat{h}_{\mathrm{mse}}(c)$ is an estimate of the pointwise-MSE-optimal bandwidth of Imbens and Kalyanaraman (2012). With this common implementation of undersmoothing, the CS from Feir et al. (2016) exhibits moderate finite-sample coverage distortions, and is highly inefficient relative to our procedure.

6.5. **Optimized Linear Estimation.** We focus on methods based on local linear regression for inference in RD designs in this paper. An alternative approach, considered recently by Armstrong and Kolesár (2018) and Imbens and Wager (2019), is to directly compute the minimax linear estimator of the respective object of interest through numerical optimization, and then use this estimator as a basis for inference. Existing results suggest that proceeding like this in the context of an FRD Anderson-Rubin-type CS construction should yield more efficient inference in settings with a multi-dimensional running variable, or in ones where the support of the running variable is rather coarse. It would, however, be quite expensive from a computational point of view, as an involved numerical optimization would have to be repeated for every $c \in \mathbb{R}$ under consideration; and little is to be gained in terms of efficiency for the most common setup of a univariate running variable with relatively rich support.

7. SIMULATIONS

7.1. **Setup.** In this section, we report the results of a Monte Carlo study of the performance of our bias-aware Anderson-Rubin-type CS, and that of competing procedures. We consider a number of data generating processes that vary with respect to the degrees of nonlinearities of the conditional expectation functions, the richness of the running variable's support, and the strength of identification. Specifically, we generate data as

$$Y_i = (B_Y/2)\mathrm{sign}(X_i) \cdot f(X_i) + \mathbf{1}\{X_i \geq 0\}\tau_Y + 0.1 \cdot \varepsilon_{1i},$$
$$T_i = \mathbf{1}\{(B_T/2)\mathrm{sign}(X_i) \cdot f(X_i) + \mathbf{1}\{X < 0\}\tau_T + 0.3 \leq \Phi(\varepsilon_{2i})\}$$

where $(\varepsilon_{1i}, \varepsilon_{2i})$ are bivariate standard normal random variables with covariance 0.5; the running variable X_i either follows a continuous uniform distribution over $[-1, 1]$ or a discrete

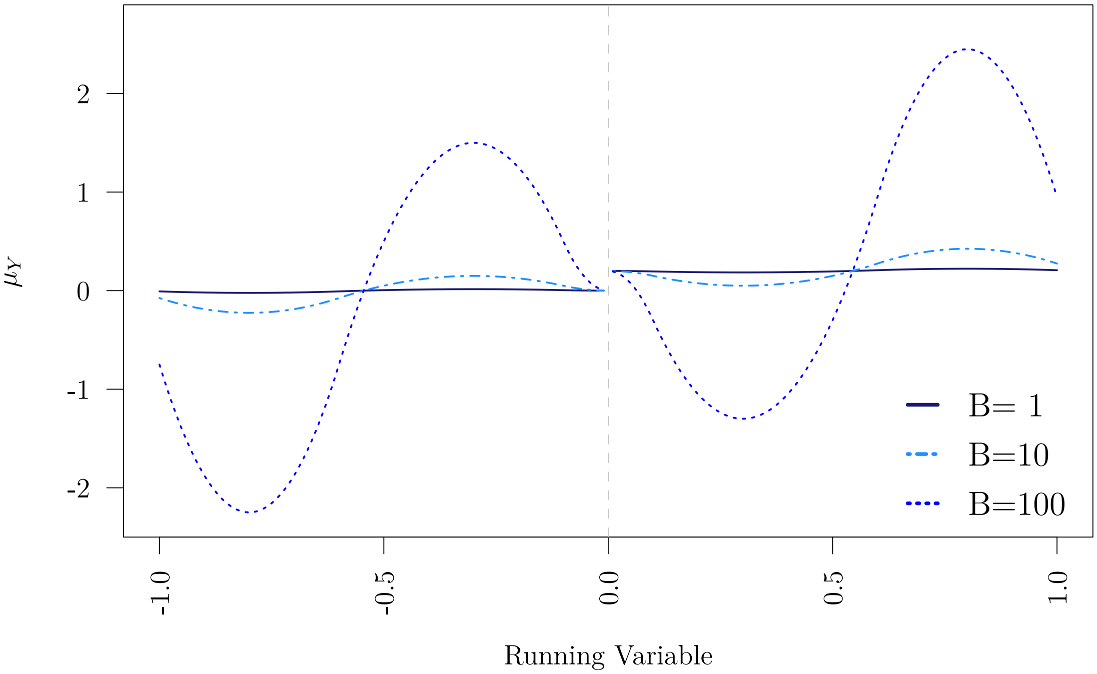


Figure 1: Shape of conditional expectation function μ_Y for different parameter values used in the simulation.

uniform distribution over $\{\pm 1/15, \pm 2/15, \ldots \pm 1\}$; and

$$f(x) = x^2 - 1.5 \cdot \max(0, |x| - 0.1)^2 + 1.25 \cdot \max(0, |x| - 0.6)^2.$$

The latter choice implies that the functions μ_Y and μ_T are second order splines whose maximal absolute second derivative over $[-1, 1]$ is B_Y and B_T, respectively. To illustrate the general shape of these functions, we plot μ_Y in Figure 2 for different values of B_Y and $\tau_Y = 1$. We then consider the parameter values $(\tau_Y, \tau_T) \in \{(1, 0.2), (0.5, 0.1)\}$, $B_T \in \{0.2, 1\}$, and $B_Y \in \{1, 10, 100\}$; and set the sample size to $n = 1,000$. Note that the values of (τ_Y, τ_T) are such that $\theta = 2$ in all of these settings. We refer to DGPs with $\tau_T = 0.1$ as weakly identified, and those with $\tau_T = 0.5$ as strongly identified.

We consider the performance of a number of different Anderson-Rubin type CSs in our simulations: (i) our bias-aware CSs, using the true values of the smoothness constants B_Y and B_T of the respective data generating process; (ii) our bias-aware CSs that uses estimates of the smoothness constants B_Y and B_T based on a rule-of-thumb, discussed in Armstrong and Kolesár (2019), that fits a global fourth order polynomial on each side of the cutoff, and then calculates its maximal second derivatives; (iii) CSs based on robust bias correction, using a local quadratic specification to estimate the bias, and estimates of the Imbens and Kalyanaraman (2012) "pointwise-MSE-optimal" bandwidth, henceforth IK bandwidth; (iv) naive CSs that ignore the bias, and also use the IK bandwidth; and (v) undersmoothing CS that use the estimated IK bandwidth multiplied by $n^{-1/20}$. In addition, we also consider delta-method-type CIs for θ as described in Section 6 with all of the five just-mentioned methods to handle the bias.

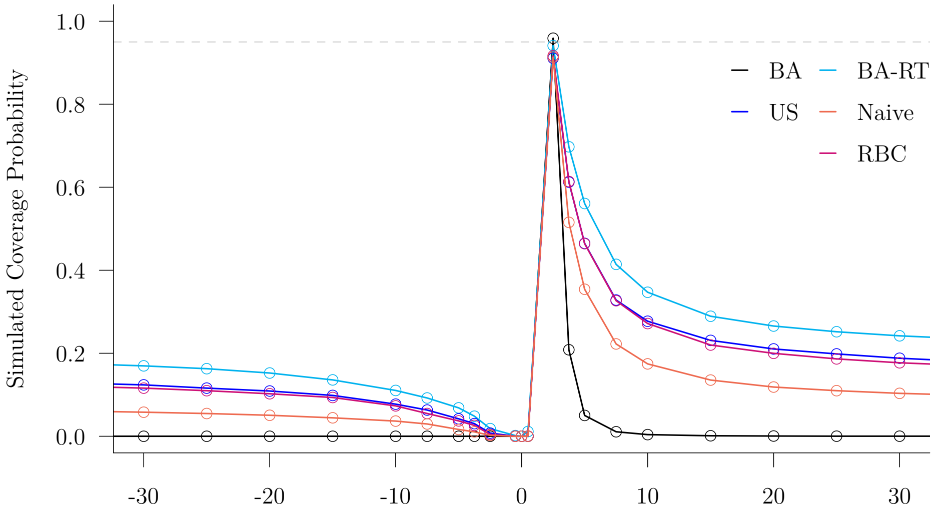
Following standard practice, we use a triangular kernel to compute the local linear (and local quadratic, in the case of methods based on robust bias correction) estimators involved in the construction of the CSs we consider. We remark that these estimators are only well-defined with a discrete running variable if the bandwidth is such that positive kernel weights are assigned to at least two (or three, in the case of methods based on robust bias correction) support points on either side of the cutoff. In our simulations, the estimated IK bandwidth is often very small and does not satisfy this criterion. Correspondingly, the standard software implementation of all methods for inference based on an estimated IK bandwidth (that is, all non-bias-aware methods) breaks down in such cases. Our results below include the rate at which the IK bandwidth fails in this sense across the different DGPs. If such a failure occurs, we manually set the bandwidth equal to 4/15, the value of the fourth largest support point, to compute the respective CS. With a triangular kernel, this means that positive weights are given to three support points on each side of the cutoff.

7.2. **Results.** Table 1 shows the simulated frequencies at which the various CSs that we consider cover the true parameter $\theta = 2$ across data generating processes. We first discuss results for Anderson-Rubin-type CSs, shown in the left panel. As predicted by our theoretical results, our bias-aware CSs have simulated coverage rates close to or greater than the nominal level irrespective of the distribution of the running variable, the degree of nonlinearity of the unknown functions, and the degree of identification strength. Using the rule-of-thumb choice for the smoothness bounds leads to considerable over-coverage, especially for setups with a discrete running variable. This is because the global quadratic approximation tends to over-estimate the smoothness bounds here. Combining a naive approach, undersmoothing, or robust bias correction with an Anderson-Rubin-type construction leads to CSs that undercover for all data generating processes we consider, with the distortions being more severe (up to about 20 percentage points) for larger values of the smoothness constants. When

Table 1: Simulated coverage rates (in %) of true parameter for various types of confidence sets

			Anderson-Rubin						Delta Method					
τ_T	B_Y	B_T	BA	BA-RT	Naive	US	RBC	IK Fail	BA	BA-RT	Naive	US	RBC	IK-Fail
Running Variable with Continuous Distribution														
0.5	1	0.2	95.9	96.1	91.1	91.1	91.7		94.7	91.1	89.4	88.2	90.1	
0.5	1	1.0	95.4	96.1	91.0	91.2	91.4		93.8	90.7	89.1	88.2	89.9	
0.5	10	0.2	94.7	96.1	90.2	90.3	90.4		95.0	92.6	90.4	88.3	91.3	
0.5	10	1.0	94.7	96.0	90.0	90.0	90.1		94.4	92.3	89.8	88.0	90.7	
0.5	100	0.2	94.2	97.4	77.6	84.5	74.3		89.6	94.0	93.8	89.1	94.3	
0.5	100	1.0	94.4	97.4	77.9	84.7	74.3		90.4	93.8	94.1	89.4	94.5	
0.1	1	0.2	96.2	97.0	91.7	91.7	92.0		89.3	77.5	73.8	71.3	77.0	
0.1	1	1.0	96.0	97.1	91.8	91.8	92.1		87.0	77.0	73.9	71.2	76.8	
0.1	10	0.2	95.7	97.1	91.1	91.2	91.5		88.5	81.0	75.2	70.1	78.2	
0.1	10	1.0	96.0	97.3	91.5	91.7	91.9		88.3	81.3	75.5	70.6	78.6	
0.1	100	0.2	95.8	98.0	83.8	89.0	80.5		88.6	88.4	83.5	73.7	85.1	
0.1	100	1.0	96.0	98.2	83.6	89.1	80.1		89.5	88.7	83.0	73.2	84.6	
Running Variable with Discrete Distribution														
0.5	1	0.2	96.7	99.1	92.5	92.5	94.8	58.5	95.5	95.1	88.7	88.5	93.2	59.3
0.5	1	1.0	96.5	99.0	92.4	92.4	94.9	59.1	94.8	94.6	88.3	87.9	93.1	59.7
0.5	10	0.2	96.7	99.3	91.7	91.8	95.1	66.5	97.1	97.7	90.8	90.8	94.6	98.8
0.5	10	1.0	96.8	99.3	91.7	91.9	95.0	67.8	97.2	97.4	90.1	90.1	94.1	98.8
0.5	100	0.2	96.4	100.0	65.7	65.7	84.5	100.0	89.9	96.3	99.4	99.4	94.5	100.0
0.5	100	1.0	96.4	100.0	64.8	64.8	83.9	100.0	92.2	96.6	99.3	99.3	94.3	100.0
0.1	1	0.2	96.9	99.2	93.5	93.6	95.0	57.8	89.5	77.9	67.6	66.5	80.5	59.4
0.1	1	1.0	97.0	99.2	93.3	93.4	95.0	58.5	87.3	77.1	66.8	65.7	79.6	58.8
0.1	10	0.2	97.7	99.4	92.9	93.0	95.2	64.7	93.5	87.6	72.1	72.1	87.7	99.2
0.1	10	1.0	97.8	99.5	93.2	93.3	95.3	66.5	93.1	87.7	72.5	72.5	87.7	99.3
0.1	100	0.2	97.0	100.0	72.1	72.1	86.3	100.0	95.0	96.4	93.6	93.6	96.8	100.0
0.1	100	1.0	97.0	100.0	71.8	71.8	85.5	100.0	95.5	96.1	94.2	94.2	96.7	100.0

Notes: Results based on 20,000 Monte Carlo draws for a nominal confidence level of 95%. BA: bias-aware approach with known smoothness bounds; BA-RT bias-aware approach with estimates smoothness bounds via rule of thumb; Naive: naive approach that ignores bias; US: undersmoothing approach; RBC: robust bias correction; IK-Fail: rate at which IK bandwidth selector fails to produce a bandwidth such that positive kernel weights are given to at least two support points on either side of the cutoff.



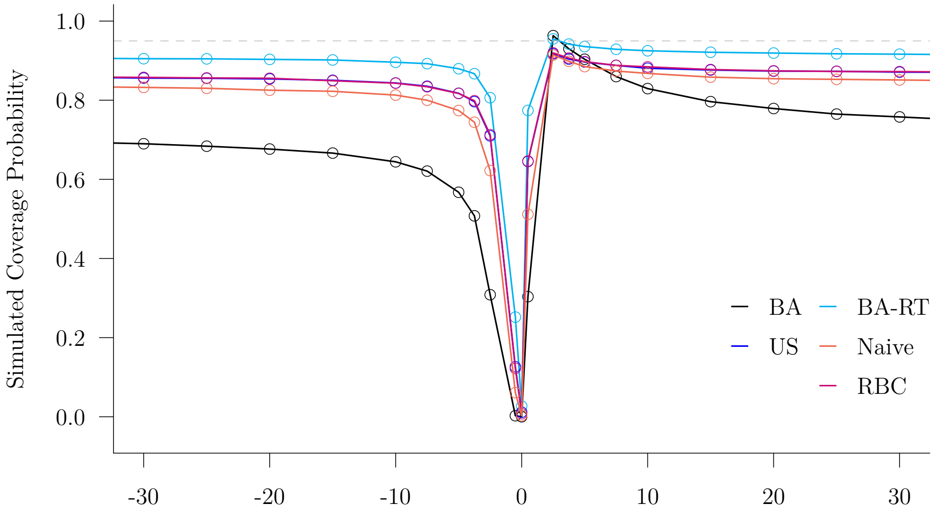


Figure 2: Simulated Rejection Probabilities. Results based on 20,000 Monte Carlo draws for a nominal confidence level of 95%. BA: bias-aware approach with known smoothness bounds; BA-RT bias-aware approach with estimates smoothness bounds via rule of thumb; Naive: naive approach that ignores bias; US: undersmoothing approach; RBC: robust bias correction

the running variable is discrete, the IK bandwidth fails in at least half of all simulation runs, and in every single instance if $B_Y = 100$.

Turning to result for delta method CIs shown in the right panel of Table 1, we see that bias-aware CIs do not necessarily provide correct coverage even under strong identification. This occurs because they only control the bias of a first-order approximation of the estimator on which the CI is based. Coverage distortions are further amplified by weak identification. Discreteness of the running variable, however, does not have a strong detrimental effect on bias-aware delta method CIs in our setups. Using the rule-of-thumb choice for the smoothness bounds leads to further distortions, but only for some setups. The coverage of delta method CIs that use the naive approach, undersmoothing, or robust bias correction is again severely distorted for most data generating processes, particularly those with weak identification.

In Figure 3, we focus on the special case $\tau_T = .5$, $B_Y = 1$, and $B_T = 0.2$, and plot the simulated rates at which the various Anderson-Rubin-type CSs cover a range of parameter values. We choose this particular data generating process because, as one can see from the first line of Table 1, the coverage of the true parameter is reasonably close to the nominal level for all procedures. This ensures that a comparison of coverage rates at "non-true" parameter values is meaningful across CSs. Generally speaking, a CS can be considered more "powerful" than a competing procedure if its coverage of non-true parameters is closer to zero over a wide range of the relevant parameter space. Figure 3 shows that the coverage rate of bias-aware Anderson-Rubin-type CSs drops very quickly to zero as we move away from the true parameter, and is clearly below that of all competing procedures over almost all of the parameter space. This confirms that the accurate coverage of our bias-aware CSs does not come at the expense of statistical power.

8. EMPIRICAL APPLICATION

In this section, we use data from Oreopoulos (2006) to illustrate the application of our CSs. Oreopoulos (2006) studied the effects of a 1947 education reform in the United Kingdom that raised the minimum school-leaving age from 14 to 15 years. The data are a sample of UK workers who turned 14 between 1935 and 1965, obtained by combining the 1984-2006 waves of the UK General Household Survey; see Oreopoulos (2006) for details. We focus on a single parameter of interest, the effect of attending school beyond age 14 on annual earnings measured in 1998 UK pounds. The running variable is the year in which the worker turned 14, and the threshold is 1947. For simplicity, we refer to data on workers who turned 14 in year x as "data for x" below. Figure 4 shows the average of log annual earnings and

Table 2: Confidence sets for the effect of one additional year of compulsory schooling

Results for full data set

		B_T				
		0.0025	0.025	0.05	0.15	0.2
	0.0025	(-0.156, 0.294)	(-0.156, 0.534)	(-0.156, 0.934)	(-0.756, 2.889)	$(-\infty, \infty)$
	0.025	(-0.276, 0.534)	(-0.316, 0.934)	(-0.356, 1.334)	(-1.156, 3.289)	$(-\infty, \infty)$
B_Y	0.05	(-0.390, 0.964)	(-0.437, 1.242)	(-0.496, 1.424)	(-1.185, 3.938)	$(-\infty, \infty)$
	0.15	(-0.840, 1.575)	(-0.936, 1.761)	(-1.075, 2.032)	(-2.906, 6.077)	$(-\infty, \infty)$
	0.2	(-1.070, 1.810)	(-1.195, 2.028)	(-1.376, 2.345)	(-3.908, 7.203)	$(-\infty, \infty)$

Results excluding data for 1947

		B_T				
		0.0025	0.025	0.05	0.15	0.2
	0.0025	(-0.132, 0.262)	(-0.136, 0.513)	(-0.177, 0.684)	(-1.095, 2.598)	$(-\infty, \infty)$
	0.025	(-0.294, 0.611)	(-0.356, 0.728)	(-0.454, 0.871)	(-1.708, 3.316)	$(-\infty, \infty)$
B_Y	0.05	(-0.478, 0.785)	(-0.556, 0.890)	(-0.653, 1.046)	(-2.462, 4.175)	$(-\infty, \infty)$
	0.15	(-1.020, 1.313)	(-1.160, 1.493)	(-1.371, 1.764)	(-6.053, 7.997)	$(-\infty, \infty)$
	0.2	(-1.289, 1.581)	(-1.468, 1.800)	(-1.739, 2.132)	(-8.037, 10.025)	$(-\infty, \infty)$

the empirical proportions of students who attended school beyond age 14 as a function of the running variable. The RD design is clearly seen to be fuzzy.

For reasons explained below, we conduct every analysis in this section separately for both the entire data and for the subset that excludes the data for 1947. Oreopoulos (2006) used global specifications in which the respective dependent variable is regressed on a dummy for turning 14 in or after 1947 and a quadratic polynomial in age to estimate FRD parameters. Here this approach yields the point estimate $\widehat{\theta} = 0.111$ with a heteroscedasticity-robust standard error of 0.033 and a 95% delta method CI of $(0.046, 0.176)$ for the full data; and a point estimate $\widehat{\theta} = .088$ with a standard error of 0.060 and a 95% delta method CI of $(-0.029, 0.205)$ if we exclude data for 1947. However, these results do not account for the potential misspecification of the global linear regression model.

In Table 2, we report our bias-aware Anderson-Rubin-type CSs with nominal level 95% for various values of the smoothness bounds, namely $(B_Y, B_T) \in \{0.0025, 0.025, 0.05, 0.15, 0.2\}^2$, separately for the entire data (top panel) and for the subsample that excludes data for 1947 (bottom panel). All CSs shown are formally valid given the respective choice of B_Y and B_T. To decide which particular CSs should be considered a reasonable description of sampling uncertainty, however, one needs to consider the empirical content of these smoothness bounds

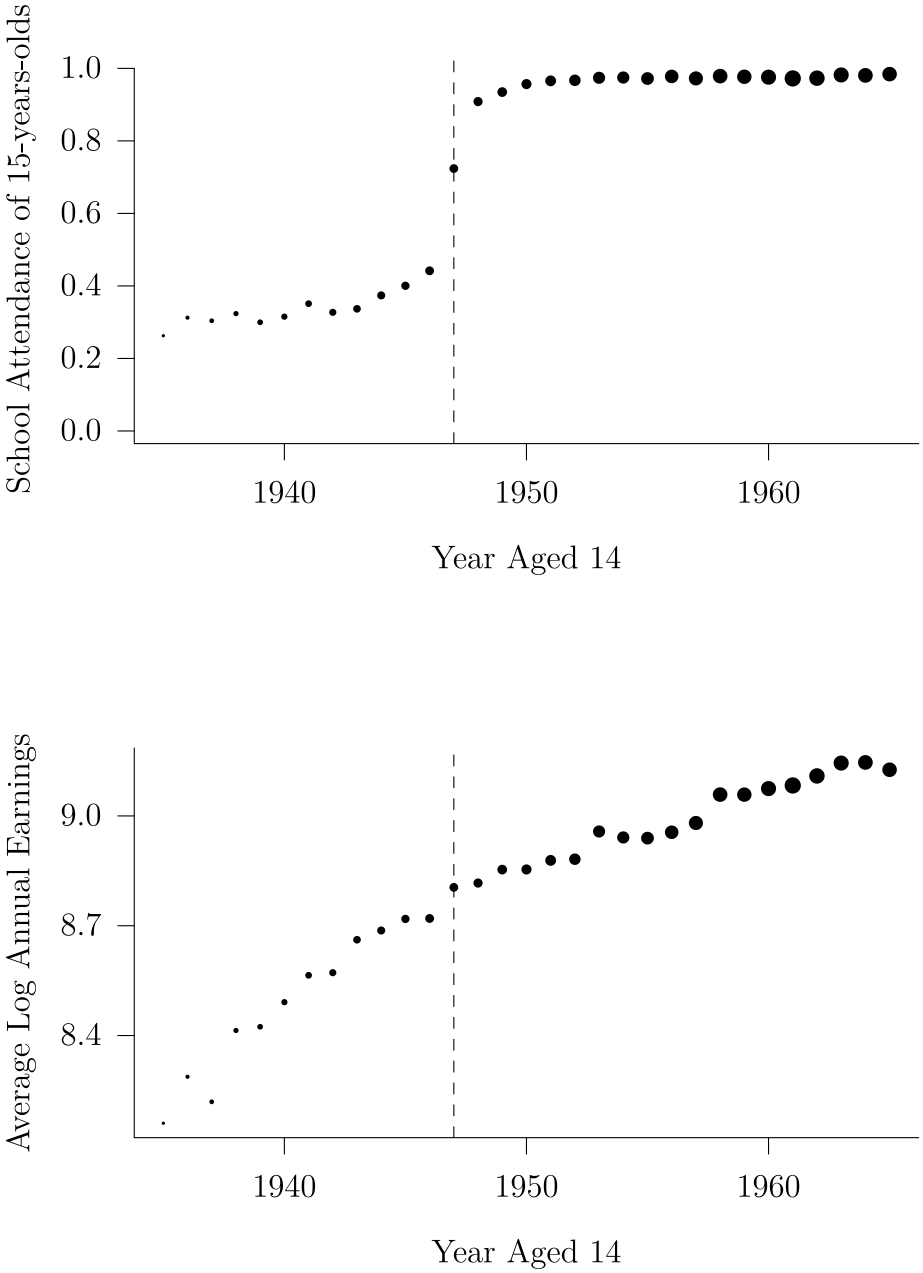


Figure 3: Fraction leaving full time education and average log annual earnings by cohort. The vertical lines indicate the year 1947, in which the minimum school leaving age changed from 14 to 15 years. Size of the dots is proportional to the cohort size. Only data from Great Britain are consider.

(one can always compute a CSs for any combination of B_Y and B_T, irrespective of whether these values make sense in a particular context).

The smallest value of B_Y that we consider corresponds to the assumption that μ_Y is very close to linear on either side of the cutoff, while larger values allow for increasing degrees of curvature of μ_Y. An analogous statement applies to B_T and μ_T. Following Kolesár and Rothe (2018), we can also interpret the value of a constant through the heuristic that a function $f \in \mathcal{F}_H(B)$ cannot deviate by more than $B/8$ from a straight line between two points that are one unit apart. When it comes to the choice of B_Y, we can use such reasoning together with the fact that a typical increase in log earnings per extra year in age is about 0.02 in the data to deduce that $B_Y = 0.025$ and $B_Y = 0.05$ should reasonable choices.

Regarding the choice of B_T, one has to be a bit more careful. From the top panel of Figure 4, we see that the empirical share of "treated" students increases very slowly after 1948, but jumps sharply from 0.724 for 1947 data to 0.909 for 1948 data. If we consider the latter change as a natural variation in treatment probabilities that was not directly associated with the reform, a minimum value for B_T of about 0.15 is necessary for μ_T to be compatible with the "hockey stick" shape of observed treatment probabilities on the right of the threshold.[8] Using such a B_T implies that a similar increase in treatment probabilities between 1946 and 1947 would have been plausible in the absence of the reform. For $B_T \geq 0.15$, the data are thus consistent with a value of τ_T that is very close to zero, which means that our parameter of interest is rather weakly identified. Indeed, for $B_Y \in \{0.025, 0.05\}$ and $B_T = 0.15$ the CSs in the top panel of Table 2 are extremely wide (in the sense that they contain values that are implausible candidates for the returns an additional year of compulsory schooling), and for $B_T = 0.2$ the CSs are in fact all equal to the entire real line.

If we take the arguably more realistic position that the change in treatment probabilities between 1947 and 1948 was still mostly caused by the introduction of the reform through delayed implementation, a more natural approach is to exclude the 1947 data for the analysis. With this sample selection the typical increase in treatment probability per year is about 0.015 on the left of the threshold and 0.005 right, which again suggests that $B_T = 0.025$ or $B_T = 0.0025$ should be reasonable choices. The parameter of interest is then rather strongly identified. The resulting CSs for $B_Y \in \{0.025, 0.05\}$ and $B_T \in \{0.0025, 0.025\}$ in the bottom panel of Table 2 are substantially more narrow than their counterparts discussed above. However, they still quite wide from an empirical point of view, which shows that the

[8] Kolesár and Rothe (2018) propose a method to estimate a lower bound on the value of B_T. Their procedure gives a point estimate of 0.158, with 95% CI of $[.126, \infty)$, for the data to the right of the threshold. For the data on the left the point estimate is zero, with 95% CI of $[0, \infty)$.

data are not very informative about the returns to schooling in principle.

9. CONCLUSIONS

Fuzzy regression discontinuity designs occur frequently in many areas of applied economics. Motivated by the various shortcomings of existing methods of inference, we propose new confidence sets for the causal effect in such designs, which are based on a bias-aware Anderson-Rubin-type construction. Our CSs are simple to compute, highly efficient, and have excellent coverage properties in finite samples because they explicitly take into account the exact smoothing bias from the local linear regression steps. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

REFERENCES

Abadie, A., G. W. Imbens, and F. Zheng (2014): “Inference for misspecified models with fixed regressors,” *Journal of the American Statistical Association*, 109, 1601–1614.

Anderson, T. and H. Rubin (1949): “Estimation of the parameters of a single equation in a complete system of stochastic equations,” *Annals of Mathematical Statistics*, 20, 46–63.

Armstrong, T. and M. Kolesár (2019): “Simple and honest confidence intervals in nonparametric regression,” *Working Paper.*

Armstrong, T. B. and M. Kolesár (2018): “Optimal inference in a class of regression models,” *Econometrica*, 86, 655–683.

Bertanha, M. and M. J. Moreira (2018): “Impossible Inference in Econometrics: Theory and Applications,” *Working Paper.*

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): “Robust nonparametric confidence intervals for regression-discontinuity designs,” *Econometrica*, 82, 2295–2326.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A Practical Introduction to Regression Discontinuity Designs: Foundations*, Elements in Quantitative and Computational Methods for the Social Sciences, Cambridge University Press.

Dong, Y. (2017): “An alternative assumption to identify LATE in regression discontinuity designs,” *Working Paper.*

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Feir, D., T. Lemieux, and V. Marmer (2016): “Weak identification in fuzzy regression discontinuity designs,” *Journal of Business & Economic Statistics*, 34, 185–196.

Fieller, E. C. (1954): “Some problems in interval estimation,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 16, 175–185.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design,” *Econometrica*, 69, 201–209.

Imbens, G. and K. Kalyanaraman (2012): “Optimal bandwidth choice for the regression discontinuity estimator,” *Review of Economic Studies*, 79, 933–959.

Imbens, G. and C. Manski (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72, 1845–1857.

Imbens, G. and S. Wager (2019): “Optimized regression discontinuity designs,” *Review of Economics and Statistics*, 101.

Imbens, G. W. and J. D. Angrist (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.

Imbens, G. W. and T. Lemieux (2008): “Regression discontinuity designs: A guide to practice,” *Journal of Econometrics*, 142, 615–635.

Kamat, V. (2018): “On nonparametric inference in the regression discontinuity design,” *Econometric Theory*, 34, 694–703.

Kolesár, M. and C. Rothe (2018): “Inference in Regression Discontinuity Designs with a Discrete Running Variable,” *American Economic Review*, 108, 2277–2304.

Lee, D. S. and D. Card (2008): “Regression discontinuity inference with specification error,” *Journal of Econometrics*, 142, 655–674.

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

Li, K.-C. (1989): “Honest confidence regions for nonparametric regression,” *Annals of Statistics*, 17, 1001–1008.

Oreopoulos, P. (2006): “Estimating average and local average treatment effects of education when compulsory schooling laws really matter,” *American Economic Review*, 96, 152–175.

Porter, J. (2003): “Estimation in the regression discontinuity model,” *Working Paper*, 2003.

Rothe, C. (2017): “Robust confidence intervals for average treatment effects under limited overlap,” *Econometrica*, 85, 645–660.

Staiger, D. and J. H. Stock (1997): “Instrumental Variables Regression with Weak Instruments,” *Econometrica*, 557–586.