BIAS-AWARE INFERENCE IN FUZZY REGRESSION DISCONTINUITY DESIGNS

Claudia Noack Christoph Rothe

Abstract

In this paper, we propose new confidence sets (CSs) for the regression discontinuity parameter in fuzzy designs. Our CSs are based on nonparametric local linear regression, and are bias-aware, in the sense that they take possible smoothing bias explicitly into account. Their construction shares similarities with that of Anderson-Rubin CSs in exactly identified instrumental variable models, and thereby avoids issues with "delta method" approximations that underlie most commonly used existing inference methods for fuzzy regression discontinuity analysis. Our CSs compare favorably in terms of both theoretical and practical performance to existing procedures in canonical settings with strong identification and a continuous running variable. However, due to their particular construction they are also valid under a wide range of empirically relevant conditions in which existing methods generally fail, such as setups with discrete running variables, donut designs, and weak identification.

First Version: June 11, 2019. This Version: February 15, 2021. We thank Tim Armstrong, Marinho Bertanha, Yingying Dong, Keisuke Hirano, Guido Imbens, Michal Kolesár, and numerous seminar participants for helpful comments and suggestions. The authors gratefully acknowledge financial support by the European Research Council (ERC) through grant SH1-77202. Contact information: Claudia Noack, Department of Economics, University of Mannheim, 68131 Mannheim, Germany, email: cnoack@mail.uni-mannheim.de. Christoph Rothe, Department of Economics, University of Mannheim, 68131 Mannheim, Germany, email: rothe@vwl.uni-mannheim.de, website: http://www.christophrothe.net.

1. INTRODUCTION

The regression discontinuity design is a popular empirical strategy for estimating causal treatment effects from observational data. In sharp (SRD) designs units receive a treatment if and only if a running variable falls above a known cutoff value, whereas in fuzzy (FRD) designs the treatment probability jumps at the threshold, but generally not from zero to one. Methods for estimation and inference based on local linear regression are widely used in empirical research for both kinds of designs, and their theoretical properties have been studied extensively; see Imbens and Lemieux (2008) or Lee and Lemieux (2010) for surveys, and Cattaneo et al. (2019) for a textbook treatment.

A key issue for SRD confidence intervals (CIs) is the handling of the estimator's smoothing bias, with undersmoothing (cf. Imbens and Lemieux, 2008) and robust bias correction (Calonico et al., 2014) being popular approaches in applications. However, Armstrong and Kolesár (2020b) show that common implementations of such CIs can have coverage issues in practice, mostly due to the way they select the bandwidth,[1] and that "bias-aware" CIs, which adjust the critical value to take possible bias into account, are more efficient than their counterparts based on either undersmoothing or robust bias correction, even at infeasible bandwidths. A further advantage of bias-aware SRD CIs relative to these alternatives is that they do not require a continuously distributed running variable.

In an FRD design, the usual point estimator is the ratio of two SRD estimators, and due to this nonlinearity one cannot directly use the same bias-handling techniques as in SRD setups. The CIs reported in empirical FRD papers therefore typically build on a delta method (DM) argument. This entails approximating the FRD estimator with a term that behaves like an SRD estimator, imposing conditions under which the corresponding error is negligible in large samples, and applying an SRD bias-handling approach to the leading term. Proceeding like this can exasperate the practical issues of undersmoothing and robust bias correction known from SRD contexts; and it can also create problems for the bias-aware approach, as bias-aware FRD DM CIs only account for an approximate bias. Moreover, any type of DM CI can only be asymptotically valid if the running variable is continuous with positive density around the cutoff, and the jump in treatment probabilities at the cutoff is "large". DM CIs

[1] Both methods typically take an estimate of a pointwise-MSE-optimal bandwidth (Imbens and Kalyanaraman, 2012) as an input. This bandwidth can be large even if the underlying function is highly nonlinear, which then leads to large smoothing biases in finite samples. Estimators of this bandwidth generally involve a regularization step to prevent extreme values, the result can depend critically on tuning parameters that are difficult to pick (Armstrong and Kolesár, 2020b).

generally break down in empirical settings that do not exhibit these properties, in the sense that their actual coverage can deviate substantially from the nominal level; and they cannot be salvaged by adjusting the method used to control the bias. This is important because empirical researchers often face running variables that take only a limited number of distinct values, like test scores or class sizes (Angrist and Lavy, 1999; Oreopoulos, 2006; Urquiola and Verhoogen, 2009; Fredriksson et al., 2013; Clark and Martorell, 2014; Hinnerich and Pettersson-Lidbom, 2014; Card and Giuliano, 2016; Jepsen et al., 2016); "donut designs" that exclude units close to the cutoff to increase the credibility of causal estimates (Almond and Doyle, 2011; Dahl et al., 2014; Dube et al., 2019; Le Barbanchon et al., 2019; Scott-Clayton and Zafar, 2019); or weakly identified setups with small jumps in treatment probabilities (Malenko and Shen, 2016; Coviello et al., 2018).

In this paper, we propose new confidence sets (CSs) for the FRD parameter that are not subject to such shortcomings. Our CSs avoid the use of the FRD point estimator, and are instead based on auxiliary statistics that can be computed directly via local linear regression. The construction avoids the approximation errors of the DM, and is somewhat analogous to that of an Anderson-Rubin (AR) statistic in an exactly identified linear instrumental variable model (Staiger and Stock, 1997). We then apply the bias-aware approach to these statistics, which allows us to account exactly for the possible smoothing bias. The resulting CSs are easy to compute; an `R` package is available on the authors' website.

We derive two main results under the common assumption that the second derivatives of the conditional expected outcome and the conditional treatment probability are bounded by some constant on either side of the cutoff. First, we show that our CSs are honest in the sense of Li (1989), meaning that they have correct asymptotic coverage uniformly over the class of functions satisfying our assumption, irrespective of the distribution of the running variable or the strength of identification. This property implies good CS performance across the entire range of plausible data generating processes, and is thus necessary for good finite-sample coverage. The novel insight here is not so much that AR CSs can accommodate weak identification, but that combining this construction with a bias-aware approach provides robustness to other deviations from the canonical setup, like discreteness of the running variable and "donut" designs.[2]

[2]Feir et al. (2016) already showed that undersmoothing AR CSs can have correct pointwise asymptotic coverage if the jump in treatment probabilities tends to zero with the sample size at an appropriate rate, while undersmoothing DM CIs generally do not have this property. Such CIs require a continuous running variable with positive density around the cutoff, and may, depending on the implementation of undersmoothing, not be honest. After circu-

Second, we show that bias-aware AR CSs are asymptotically equivalent to bias-aware DM CIs if the running variable is continuous and identification is strong, which are conditions needed for DM CIs to be honest in the first place. The robustness of bias-aware AR CSs does thus not come with a cost in terms of power relative to DM CIs in a canonical setup. Moreover, since Armstrong and Kolesár (2020b) show that bias-aware DM CIs outperform DM CIs based on undersmoothing and robust bias correction, the equivalence result implies that the same is true for our bias-aware AR CSs. These predictions are confirmed by simulation results reported in this paper.

We also make three contributions regarding the implementation of bias-aware inference that are not only important for our CSs, but can also be used more generally. First, we provide a new standard error for local linear regression estimates that is uniformly consistent over the class of functions with bounded second derivatives. It is a variation of the nearest-neighbor variance estimator (e.g. Abadie and Imbens, 2006) commonly used in the RD literature. Our proposal replaces the usual local average with a local linear projection among the nearest neighbors, which removes a bias term that is proportional to the underlying function's first derivative. Second, we propose a new empirical bandwidth that enforces an upper bound on Lindeberg weights to ensure that a normal approximation works well for our local linear estimates in finite samples. Third, we provide new graphical tools and an analysis of "rules of thumb" that can help guide the choice of the bounds on second derivatives, which are the main tuning parameters required for bias-aware inference.

As an extension, we also derive new bias-aware CSs for the fuzzy regression kink design (Card et al., 2015), and establish theoretical properties analogous to those we obtain for the FRD case. These results also apply more generally to settings in which the parameter of interest is the ratio of jumps in the vth-order derivatives of two conditional expectation functions at some threshold value.

Our paper contributes to a growing literature on "bias-aware" inference. Building on classical work (e.g. Sacks and Ylvisaker, 1978; Donoho, 1994), such methods, which take bias explicitly into account rather than trying to remove it, have recently been shown to yield powerful and practical CSs in a wide range of non- and semiparametric problems (Armstrong and Kolesár, 2018, 2020a,b; Kolesár and Rothe, 2018; Imbens and Wager, 2019; Ignatiadis and Wager, 2020; Schennach, 2020; Armstrong et al., 2020). A concern sometimes raised

lating the first draft of this paper, we were also made aware that Huang and Zhan (2020) have discussed combining a bias-aware approach with an AR-type statistic. Since they misinterpret the results from Armstrong and Kolesár (2020b), however, their proposed methods do not yield valid inference.

with these methods is that, in contrast to traditional approaches such as undersmoothing or robust bias correction, they require specifying explicit bounds on the smoothness of the underlying functions. However, this view neglects such bounds are implicitly required for traditional methods to work well in practice.[3] Following the literature on bias-aware inference, we recommend to vary the values of smoothness bounds in the construction of our CS in empirical practice as a form of sensitivity analysis. We also provide a number of tools to guide and communicate the choices.

The remainder of this paper is structured as follows. Section 2 describes our setup. Section 3 describes existing approaches to SRD and FRD inference, and discusses issues with DM CIs. Section 4 describes our bias-aware AR CSs, and Section 5 establishes their theoretical properties. Section 6 discusses implementation issues. Section 7 contains a simulation study, and Section 8 an empirical application. Section 9 concludes. The appendix contains the proofs of our main theorems. Further technical arguments, extensions and additional materials are given in the online appendix.

2. SETUP AND PRELIMINARIES

2.1. **Fuzzy RD Designs.** 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 a random sample of size n from a large population. Treatment is assigned if the running variable falls above a known cutoff. We normalize this threshold to zero, so that $Z_i = \mathbf{1}\{X_i \geq 0\}$. Because of limited compliance, it could be that $Z_i \neq T_i$. For a generic random variable W_i (which could be equal to Y_i or T_i, for example), we then write $\mu_W(x) = \mathbb{E}(W_i|X_i = x)$ for its conditional expectation function given the running variable; $\mu_{W+} = \lim_{x\downarrow 0} \mu_W(x)$ and $\mu_{W-} = \lim_{x\uparrow 0} \mu_W(x)$ for its right and left limit at zero; and $\tau_W = \mu_{W+} - \mu_{W-}$ for the jump in μ_W at the cutoff. The parameter of interest is

$$\theta = \frac{\tau_Y}{\tau_T},$$

which, in a potential outcomes framework with certain continuity and monotonicity conditions (e.g. Hahn et al., 2001; Dong, 2018), has a causal interpretation as the local average

[3]For example, in order for standard implementations of robust bias correction SRD CIs to have approximately correct coverage in finite samples, one must have a "sufficiently small" bound on the underlying function's third derivative (Kamat, 2018). Researchers that report such CIs and consider them reliable thus implicitly impose a smoothness bound. Moreover, if that bound was made explicit, a more efficient CI could be constructed through a bias-aware approach (Armstrong and Kolesár, 2020b).

treatment effect among "compliers" at the cutoff, where "compliers" are units whose treatment decision is affected by the assignment rule (Imbens and Angrist, 1994).

2.2. **Honest Confidence Sets.** Our goal is to construct confidence sets (CSs) that cover the parameter θ in large samples with at least some pre-specified probability, uniformly over (μ_Y, μ_T) in some function class $\mathcal{F}$ that embodies 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

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1-\alpha \tag{2.1}$$

for some $\alpha > 0$.[4] Following Li (1989), we refer to such CSs as *honest with respect to* $\mathcal{F}$. This is a much stronger requirement than correct pointwise asymptotic coverage:

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

In particular, under (2.1) we can always find a sample size n such that the coverage probability of $\mathcal{C}^\alpha$ is not below $1-\alpha$ by more than an arbitrarily small amount for every $(\mu_Y, \mu_T) \in \mathcal{F}$. Under (2.2) 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 (2.1) is necessary for good finite sample coverage of $\mathcal{C}^\alpha$ across data generating processes. Of course, we also want CSs that are efficient, in the sense that they are "small" while maintaining honesty.

2.3. **Smoothness Conditions.** Following Armstrong and Kolesár (2018, 2020b), 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 real functions 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 $B > 0$; and let

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

[4]Note that we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on μ_Y and μ_T implicit in our notation. Each function pair (μ_Y, μ_T) corresponds to a single distribution of $(Y, T, X, Z) = (\mu_Y(X) + \epsilon_M, \mathbf{1}\{\mu_T(X) \geq \epsilon_T\}, X, Z)$, where (ϵ_M, ϵ_T) is some fixed random vector.

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

$$(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^0(B_T) \equiv \mathcal{F}, \tag{2.3}$$

for some constants B_Y and B_T whose choice in empirical practice we discuss in Section 6.4. Note that in addition to imposing smoothness, condition (2.3) also rules out cross-restrictions between the shapes of μ_Y and μ_T, since $\mathcal{F}$ is a Cartesian product. This seems reasonable for applications in economics. Also note that we impose $\mu_T \in \mathcal{F}_H^0(B_T)$, and thus that $\tau_T \neq 0$, only to ensure that the parameter of interest $\theta = \tau_Y/\tau_T$ is well-defined. Our setup explicitly allows τ_T to be arbitrarily close to zero.

2.4. **Discrete Settings.** Conditional expectation functions are only well-defined over the support of the conditioning variable. One must therefore clarify the meaning of (2.3) if X_i is discrete, or more generally such that there are gaps in its support. Following Kolesár and Rothe (2018) and Imbens and Wager (2019), we understand this condition to mean that there exists a single "true" function pair $(\mu_Y, \mu_T) \in \mathcal{F}$ 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. This pair is then obviously point identified on the support of the running variable, and partially identified everywhere else through the shape restrictions implied by it being an element of $\mathcal{F}$. This reasoning further implies that θ must be contained in the identified set

$$\Theta_I = \left\{ \frac{m_{Y+} - m_{Y-}}{m_{T+} - m_{T-}} : (m_Y, m_T) \in \mathcal{F},\ (m_Y(X_i), m_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i)) \text{ w.p.1} \right\}.$$

This set is a singleton if X_i is supported on an open neighborhood around the cutoff, but generally it is either (i) a closed interval $[a_1, a_2]$; (ii) the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$; (iii) the entire real line; or, as a knife-edge case (iv) a half-line $[a_1, \infty)$ or $(-\infty, -a_1]$, with $a_1 > 0$. This holds because the range of $(m_{Y+} - m_{Y-}, m_{T+} - m_{T-})$ over $(m_Y, m_T) \in \mathcal{F}$ is a Cartesian product of two intervals $I_Y \times I_T$. The four cases then obtain depending on which of these two intervals contain zero, possibly as a boundary value.

Note that while it is not possible to consistently estimate either τ_Y, τ_T, or θ if Θ_I is not a singleton, inference is not futile in such cases. Indeed, our CSs described below are valid in the sense of Imbens and Manski (2004) irrespective of whether θ is point or partially identified, and without applied researchers having to decide which of the two notions of identification more accurately describes their particular setting.

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

$$\widehat{\tau}_W(h) = e_1^\top \underset{\beta \in \mathbb{R}^4}{\operatorname{argmin}} \sum_{i=1}^{n} K(X_i/h)(W_i - \beta'(Z_i, X_i, Z_i X_i, 1))^2, \tag{2.4}$$

where $K(\cdot)$ is a kernel function with support $[-1, 1]$, $h > 0$ is a bandwidth, and $e_1 = (1, 0, 0, 0)'$ is the first unit vector. The natural point estimator of θ is then given by $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$, for some value of h. A key feature of $\widehat{\tau}_W(h)$ is that it can be written as a weighted average of the W_i, with weights $w_i(h)$ that depend on the data through the realizations $\mathcal{X}_n = (X_1, \ldots, X_n)'$ of the running variable only:

$$\widehat{\tau}_W(h) = \sum_{i=1}^{n} w_i(h) W_i.$$

The exact form of the weights follows from standard least squares algebra, and is given explicitly in Appendix A. Estimators of the form (2.4) are the building blocks of our honest CSs described below, and we refer to $\widehat{\tau}_W(h)$ as an SRD-type estimator of τ_W in the following, as it is the conventional estimator in a hypothetical SRD design with outcome W_i.

3. EXISTING METHODS FOR RD INFERENCE

3.1. **SRD Inference.** We first review some techniques for inference based on SRD-type estimators, which are by now well-understood. To describe the bias-aware SRD CIs of Armstrong and Kolesár (2018, 2020b), let $b_W(h)$ and $s_W(h)$ denote the bias and standard deviation, respectively, of a generic SRD-type estimator $\widehat{\tau}_W(h)$ conditional on the realizations of the running variable; and let $\widehat{s}_W(h)$ be a standard error. Under mild conditions, the large sample distribution of the t-ratio $(\widehat{\tau}_W(h) - \tau_W)/\widehat{s}_W(h)$ is then that of the sum of a standard normal random variable and the ratio $b_W(h)/\widehat{s}_W(h)$. While the latter is unknown in practice, a bound $\widehat{r}_W(h) = (\sup_{\mu_W \in \mathcal{F}_H(B_W)} |b_W(h)|)/\widehat{s}_W(h)$ on $|b_W(h)/\widehat{s}_W(h)|$ can be calculated explicitly. One can then construct the bias-aware CI

$$\mathcal{C}_W^\alpha = [\widehat{\tau}_W(h) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}_W(h))\widehat{s}_W(h)],$$

where the critical value $\mathrm{cv}_{1-\alpha}(r)$ is the $(1-\alpha)$-quantile of $|N(r, 1)|$, the distribution of the absolute value of a normal random variable with mean r and unit variance. Armstrong and

Kolesár (2018, 2020b) show that this CI is honest with respect to $\mathcal{F}_H(B_W)$ irrespective of the distribution of the running variable, valid for any bandwidth (for which the quantities involved in its construction are well-defined), and highly efficient if the running variable is continuous and the bandwidth is chosen to minimize the length of $\mathcal{C}_W^\alpha$.

Other popular approaches to SRD inference include undersmoothing, or using a "small" bandwidth for which the "bias to standard error" ratio is asymptotically negligible (cf. Imbens and Lemieux, 2008); and robust bias correction, which involves subtracting a bias estimate from $\widehat{\tau}_W(h)$, and adjusting the standard error (Calonico et al., 2014). In either case, CIs are formed with the usual critical value $\mathrm{cv}_{1-\alpha}(0)$. Both approaches assume a continuously distributed running variable, but Armstrong and Kolesár (2020b) show that common implementations of undersmoothing and robust bias correction can still have finite-sample issues in such settings. One reason is that both methods typically take an estimate of a pointwise-MSE-optimal bandwidth (Imbens and Kalyanaraman, 2012) as an input. This bandwidth can be very large even if the underlying function is highly nonlinear, which then leads to large smoothing biases in finite samples. While estimators of the pointwise-MSE-optimal bandwidth generally involve a regularization step to prevent extreme bandwidth values, in practice the result is often still unstable and depends critically on the values of tuning parameters, which are difficult to pick. Armstrong and Kolesár (2020b) also show that undersmoothing and robust bias correction CIs are inefficient, in that they tend to be much longer than bias-aware counterparts, even with infeasible bandwidths.

3.2. **Delta Method FRD Inference.** The above mentioned methods for SRD inference critically rely on the "weighted average" representation of local linear regression estimators. Since the FRD estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$ is a nonlinear transformation of two SRD-type estimators, such methods cannot simply be applied directly. Instead, the CIs commonly reported in empirical FRD studies are based on a "delta method" (DM) argument. From a simple Taylor expansion, it follows that $\widehat{\theta}(h) - \theta$ can be written as the sum of an SRD-type estimator $\widehat{\tau}_U(h)$ as in (2.4), with an unobserved dependent variable U_i, and a remainder $\widehat{\rho}(h)$:

$$\widehat{\theta}(h) - \theta = \widehat{\tau}_U(h) + \widehat{\rho}(h), \quad \widehat{\tau}_U(h) = \sum_{i=1}^n w_i(h)U_i, \quad U_i = \frac{Y_i - \tau_Y}{\tau_T} - \frac{\tau_Y(T_i - \tau_T)}{\tau_T^2},$$
$$\widehat{\rho}(h) = \frac{\widehat{\tau}_Y(h)(\widehat{\tau}_T(h) - \tau_T)^2}{2\widehat{\tau}_T^*(h)^3} - \frac{(\widehat{\tau}_Y(h) - \tau_Y)(\widehat{\tau}_T(h) - \tau_T)}{\tau_T^2},$$

with $\widehat{\tau}_T^*(h)$ an intermediate value between τ_T and $\widehat{\tau}_T(h)$. With DM inference, one then imposes regularity and bandwidth conditions under which $\widehat{\rho}(h)$ is an asymptotically negligible

relative to $\widehat{\tau}_U(h)$, and forms a CI for θ by applying some method for SRD inference to $\widehat{\tau}_U(h)$. Since U_i is unobserved, any such method is must be made feasible by using an estimate $\widehat{U}_i$ in which τ_Y and τ_T are replaced by suitable preliminary estimators. Versions of such CIs are proposed, for example, by Calonico et al. (2014) and Armstrong and Kolesár (2020b) in combination with robust bias correction and a bias-aware approach, respectively.[5]

An obvious downside of such constructions, to which we refer as DM CIs, is that they only control the bias of a first-order approximation of $\widehat{\theta}(h)$, and not the bias of $\widehat{\theta}(h)$ itself. Moreover, replacing U_i with an estimate $\widehat{U}_i$ introduces additional uncertainties in finite samples. In practice, all DM FRD CIs are thus subject to additional distortions relative to conventional SRD CIs. A more principal, and more practically important issue with DM CIs is that the central condition for their validity, namely that $\widehat{\rho}(h)$ is asymptotically negligible relative to $\widehat{\tau}_U(h)$, is not innocuous. In particular, this condition is not compatible with a discrete running variable, or more generally one with support gaps around the cutoff.

To see this last point, recall from Section 2.4 that consistent estimation of τ_T and τ_Y is generally not possible with a discrete running variable. The terms $\widehat{\tau}_U(h)$ and $\widehat{\rho}(h)$ therefore have non-zero probability limits in this case, and $\widehat{\rho}(h)$ cannot be ignored for the purpose of inference on θ. This issue occurs irrespective of the method chosen to control the bias of $\widehat{\tau}_U(h)$, including bias-aware inference. Since running variables with discrete or irregular support are ubiquitous in practice, this is an important limitation.

Another issue for DM CIs is that the conditions for their validity rule out weakly identified settings with τ_T close to zero. This issue occurs even if the running variable is continuously distributed. To see this, note that for any DM CI to be honest with respect to $\mathcal{F}$, the term $\widehat{\rho}(h)$ must be of smaller order than $\widehat{\tau}_U(h)$ not only at the "true" function pair (μ_Y, μ_T), but uniformly over all $(\mu_Y, \mu_T) \in \mathcal{F}$. But since τ_T can be arbitrarily close to zero over $(\mu_Y, \mu_T) \in \mathcal{F}$, we have that $\sup_{\mu_Y, \mu_T} |\widehat{\rho}(h)| = \infty$, which means that DM CIs break down.[6]

[5] In empirical papers, FRD estimates are sometimes obtained through the two-stage least squares regression $Y_i = \theta T_i + \beta_+ X_i Z_i + \beta_- X_i (1 - Z_i) + \varepsilon_i$ with Z_i as an instrument for T_i, using only data in some window around the cutoff. This is numerically equivalent to a ratio of local linear regressions with a uniform kernel, and the resulting CI is thus of the DM type (Hahn et al., 2001; Imbens and Lemieux, 2008).

[6] Feir et al. (2016) also point out covarage issues of DM CIs under weak identification, although through a different technical argument. Specifically, they show that DM CIs based on infeasible "undersmoothing" bandwidths do not have correct asymptotic coverage under pointwise asymptotics when τ_T tends to zero with the sample size at an appropriate rate.

4. BIAS-AWARE FUZZY RD CONFIDENCE SETS

We propose an alternative approach to FRD inference that avoids the inherent shortcomings of DM CIs by directly considering an object that can be estimated by an SRD-type estimator. We define the "auxiliary" parameter $\tau_M(c) = \tau_Y - c\tau_T$, which can be written as

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

That is, $\tau_M(c)$ is the jump in the conditional expectation $\mu_M(x,c)$ of the constructed outcome $M_i(c)$ given the running variable X_i at the cutoff $x = 0$. We can form a bias-aware CI for $\tau_M(c)$ based on the SRD-type estimator $\widehat{\tau}_M(h,c)$, which is as in (2.4) but with $M_i(c)$ replacing W_i, and a bandwidth that might depend on c. Note that to keep the notation simple, the estimator $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_T(h)$ uses the same bandwidth on each side of the cutoff, and also the same bandwidth for estimating τ_Y and τ_T. It is straightforward to accommodate more general bandwidth choices; see Online Appendix B for details.

Our CS for the actual parameter of interest θ is then obtained by collecting all values of c for which the "auxiliary" CI contains zero:

$$\mathcal{C}^{\alpha}_{\text{ar}} = \{c \in \mathbb{R} : \text{a } (1-\alpha) \text{ bias-aware CI for } \tau_M(c) \text{ contains } 0\}. \tag{4.1}$$

This construction shares similarities with that of Anderson and Rubin (1949) for inference in exactly identified linear IV models, and Fieller (1954) for inference on ratios. Emphasizing the former connection, we refer to such CSs as bias-aware AR CSs for θ.

To describe the approach in more detail, recall the notation from Section 2.5 and denote the conditional bias and standard deviation of $\widehat{\tau}_M(h,c) = \sum_{i=1}^n w_i(h)M_i(c)$ given $\mathcal{X}_n = (X_1, \ldots, X_n)'$ by $b_M(h,c) = \mathbb{E}(\widehat{\tau}_M(h,c)|\mathcal{X}_n) - \tau_M(c)$ and $s_M(h,c) = \mathbb{V}(\widehat{\tau}_M(h,c)|\mathcal{X}_n)^{1/2}$, respectively. These quantities can be written more explicitly as

$$b_M(h,c) = \sum_{i=1}^n w_i(h)\mu_M(X_i,c) - (\mu_{M+}(c) - \mu_{M-}(c)), \quad s_M(h,c) = \left(\sum_{i=1}^n w_i(h)^2\sigma^2_{M,i}(c)\right)^{1/2},$$

with $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ the conditional variance of $M_i(c)$ given X_i. The bias depends on (μ_Y, μ_T) through the transformation $\mu_M = \mu_Y - c \cdot \mu_T$ only, and $\mu_Y - c \cdot \mu_T \in \mathcal{F}_H(B_Y + |c|B_T)$ by (2.3) and linearity of the second derivatives operator. Following Armstrong and Kolesár (2020b), we can bound $b_M(h,c)$ in absolute value over the functions contained in $\mathcal{F}$, for any

value of the bandwidth h, by

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

with the supremum being achieved by a pair of piecewise quadratic functions with second derivatives equal to $(B_Y \cdot \text{sign}(x), B_T \cdot \text{sign}(x))$ over $x \in [-h, h]$.[7] Under standard regularity conditions, the statistic

$$\frac{\widehat{\tau}_M(h,c) - \tau_M(c)}{s_M(h,c)} = \frac{\widehat{\tau}_M(h,c) - \tau_M(c) - b_M(h,c)}{s_M(h,c)} + \frac{b_M(h,c)}{s_M(h,c)}$$

is then the sum of a term that is approximately standard normal in large samples conditional on $\mathcal{X}_n$, and a term that is bounded in absolute value by $r_M(h,c) = \bar{b}_M(h,c)/s_M(h,c)$, the "worst case" bias to standard deviation ratio. For every $c \in \mathbb{R}$ we can thus construct an (infeasible) auxiliary bias-aware CI for the pseudo parameter $\tau_M(c)$ as $C_M^\alpha(h,c) = [\widehat{\tau}_M(h,c) \pm \text{cv}_{1-\alpha}(r_M(h,c))s_M(h,c)]$, where $\text{cv}_{1-\alpha}(r)$ is again the $(1-\alpha)$-quantile of the $|N(r,1)|$ distribution. Since the construction of this CI 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, it is also valid for any choice of bandwidth $h = h(c)$, including fixed ones that do not depend on the sample size. Its asymptotic length is minimized by

$$h_M(c) = \operatorname*{argmin}_h \text{cv}_{1-\alpha}(r_M(h,c))s_M(h,c).$$

Following the idea from (4.1), an efficient infeasible CS for θ is then given by the collection of all values of c for which the auxiliary CI $C_M^\alpha(h,c)$, evaluated at $h_M(c)$, contains zero:

$$\mathcal{C}_*^\alpha = \left\{c : |\widehat{\tau}_M(h_M(c),c)| \leq \text{cv}_{1-\alpha}(r_M(h_M(c),c))s_M(h_M(c),c))\right\}. \tag{4.2}$$

Our proposed class of CSs for θ are then feasible versions of (4.2) that replace $s_M(h,c)$ and $h_M(c)$ with suitable empirical analogues $\widehat{s}_M(h,c)$ and $\widehat{h}_M(c)$, respectively:

$$\mathcal{C}_{\text{ar}}^\alpha = \left\{c : |\widehat{\tau}_M(\widehat{h}_M(c),c)| \leq \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c),c))\widehat{s}_M(\widehat{h}_M(c),c))\right\}, \tag{4.3}$$

[7] Note that this bound may not be sharp if no such pair of piecewise quadratic functions is a feasible candidate for (μ_Y, μ_T). For example, 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 $x \in [-h,h]$ if $h > (2/B_T)^{1/2}$. Still, the bias bound is valid in such cases.

with $\widehat{r}_M(h,c) = \overline{b}_M(h,c)/\widehat{s}_M(h,c)$. Such CSs could in principle be implemented in a variety of ways, and our theoretical analysis below therefore only imposes some weak "consistency" conditions. However, we propose a specific standard error $\widehat{s}_M(h,c)$ that substitutes appropriate nearest-neighbor estimates $\widehat{\sigma}^2_{M,i}(c)$ into the above expression for $s_M(h,c)$ in Section 6.1; and a feasible bandwidth $\widehat{h}_M(c)$ that combines a plug-in construction with a safeguard against certain small sample distortions in Section 6.2.

In Online Appendix D, we present an extension of our approach that allows constructing a bias-aware AR CSs for the ratio of the jumps in the vth-order derivatives of two conditional expectation functions at some threshold value, using pth-order local polynomial regression. The most prominent example of such setup is the Fuzzy Regression Kink Design (Card et al., 2015), where the parameter of interest is the ratio of jumps in first derivatives, and the CSs are typically based on local quadratic regression.

5. THEORETICAL PROPERTIES

5.1. **Assumptions.** To study the theoretical properties of our proposed CSs, we introduce the following assumptions.

Assumption 1. *(i) The data* $\{(Y_i, T_i, X_i), i = 1, \dots, n\}$ *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 most kernel functions commonly used in applied RD analysis, such as the triangular or the Epanechnikov kernels.

Assumption 2. *The following holds uniformly over* $(\mu_Y, \mu_T) \in \mathcal{F}$*: (i)* $\widehat{h}_M(c) = h_M(c)(1 + o_P(1))$*; and (ii)* $\widehat{s}_M(\widehat{h}_M(c), c) = s_M(h_M(c), c)(1 + o_P(1))$*.*

Part (i) of Assumption 2 states that the empirical bandwidth is consistent for the infeasible optimal one, and part (ii) states that that the empirical standard error is consistent

for the true standard deviation at the infeasible optimal bandwidth. We discuss specific implementations in Sections 6.1 and 6.2.

Assumption LL1. The support of the running variable X_i is finite and symmetric, in the sense that it is of the form $\{\pm x_1, \ldots, \pm x_k\}$, for positive constants $(x_1, \ldots, x_k)$ over some open neighborhood of the cutoff.

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 $x \in \text{supp}(X_i)$ and $(\mu_Y, \mu_T) \in \mathcal{F}$ for every $c \in \mathbb{R}$; and (iii) $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^4|X_i = x)$ is uniformly bounded over $x \in \mathbb{R}$ and $(\mu_Y, \mu_T) \in \mathcal{F}$ for every $c \in \mathbb{R}$.

Assumptions LL1–LL2 are standard descriptions of setups with a discrete and a continuously distributed running variable, respectively.[8] In Lemma A.1 in the Appendix, we show that these assumptions have two main implications that we use in the proofs of the main results below: (i) using an estimate of the optimal bandwidth instead of its population version has a minor impact, in some appropriate sense, on the quantities involved in the construction of our CS; (ii) the magnitude of each of the weights $w_i(h_M(c))$ becomes arbitrarily small relative to the others' in large samples, in the sense that $w_{\text{ratio}}(h_M(c)) = o_P(1)$, where $w_{\text{ratio}}(h) = \max_{j=1,\ldots,n} w_j(h)^2 / \sum_{i=1}^n w_i(h)^2$, which means that a CLT applies to an appropriately standardized version of the estimator of $\tau_M(c)$.

5.2. **Honesty.** 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 (2.1) under the rather weak conditions introduced in the previous subsection. As mentioned above, such a property is necessary to guarantee that a CS has good finite sample coverage.

Theorem 1. *Suppose that Assumptions 1–2 and either LL1 or LL2 hold. Then $\mathcal{C}^\alpha_{\text{ar}}$ is honest with respect to $\mathcal{F}$ in the sense of* (2.1)*.*

5.3. **Shape.** Since $\mathcal{C}^\alpha_{\text{ar}}$ is defined through an inversion argument, it is interesting to study its general shape. Recall that $c \in \mathcal{C}^\alpha_{\text{ar}}$ if and only if

$$|\widehat{\tau}_M(\widehat{h}_M(c), c)| - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c), c))\widehat{s}_M(\widehat{h}_M(c), c) \leq 0.$$

[8] A discrete running variable with asymmetric support can easily be accommodated by using a different bandwidth on each side of the cutoff, as in described Appendix B.

A simple sufficient condition for $\mathcal{C}^{\alpha}_{\text{ar}}$ to be non-empty is that $h_M(c)$ is continuous in c, but beyond that it is difficult to make general statements. This is because the quantities involved in the above inequality depend on c directly, but also indirectly through the bandwidth $\widehat{h}_M(c)$. While the former dependence is rather simple in structure, the latter introduces complicated nonlinearities that make it impossible to give a simple analytical result regarding the shape of our CS. Such a characterization is possible, however, for a version that uses bandwidth that does not depend on c.

Theorem 2. *Let $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ be a version of $\mathcal{C}^{\alpha}_{\text{ar}}$ that uses a bandwidth h that does not depend on c. Then either $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, \infty)$, or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, \infty)$ or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, -a_1]$, for some constants $a_1 < a_2$.*

This result mirrors the identification analysis in Section 2.4, and suggests that our actual CS should also take one of these general shapes as long as $\widehat{h}_M(c)$ does not vary "too much" with c. We found this to be the case in every simulation run and every empirical analysis that we conducted in the context of this paper. The last two cases in Theorem 2, in which $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ is a half-line, are also "knife-edge" cases: they only occur if one of the boundaries of a bias-aware CI for τ_T is exactly equal to zero. Since this is a probability zero event under standard asymptotics, these two cases are largely irrelevant for empirical practice.

5.4. **Comparison with Bias-Aware Delta Method CIs.** Armstrong and Kolesár (2020b) study bias-aware DM CIs under conditions for which such DM CIs are asymptotically valid. These include Assumption LL2, which implies that is X_i continuously distributed, and that $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}^{\delta}_H(B_T) \equiv \mathcal{F}^{\delta}$ for some $\delta > 0$, which means that τ_T is well-separated from zero. Armstrong and Kolesár (2020b) show that in this case bias-aware DM CIs honest with respect to $\mathcal{F}^{\delta}$, and also near-optimal, in the sense that no other method can substantially improve upon its length in large samples. This construction thus dominates others commonly used in empirical practice, such as robust bias correction (Calonico et al., 2014).

The next theorem shows that our bias-aware AR CSs are as efficient as their DM counterparts in settings for which DM CIs are specifically designed. In order to avoid introducing additional high-level assumptions about the implementation details we consider an infeasible version of the bias-aware DM CI from Armstrong and Kolesár (2020b), and compare them to our infeasible counterpart $\mathcal{C}^{\alpha}_{*}$; see the proof for further discussion and the exact construction of $\mathcal{C}^{\alpha}_{\Delta}$. Equal efficiency is established in the sense that both CSs have the same local asymptotic coverage for a drifting parameter within a neighborhood of θ; the most interesting being of order $O(n^{-2/5})$, as the length of $\mathcal{C}^{\alpha}_{\Delta}$ is $O_P(n^{-2/5})$ uniformly over $\mathcal{F}^{\delta}$.

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

$$\limsup_{n\to\infty} \sup_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} \left|\mathbb{P}\left(\theta^{(n)} \in \mathcal{C}_*^\alpha\right) - \mathbb{P}\left(\theta^{(n)} \in \mathcal{C}_\Delta^\alpha\right)\right| = 0.$$

This result parallels the well-known finding that there is no loss of efficiency when using the AR approach in exactly identified IV models relative to one based on a conventional t-test (e.g. Andrews et al., 2019). It is not a simple corollary, however, as there are, for example, no analogues to the bandwidth and the smoothing bias in an IV model. Note that bias-aware DM CIs do not account for the actual bias of the estimator of interest, but only for the bias of the leading term in a stochastic approximation; and even that bound needs to be estimated. They are thus subject to additional higher-order distortions that could affect their finite sample performance relative to that of our AR CSs.

6. IMPLEMENTATION DETAILS

6.1. **Standard Errors.** Given the form of the conditional standard deviation $s_M(h,c)$, it is natural to use a standard error of the form $\widehat{s}_M(h,c) = (\sum_{i=1}^n w_i(h)^2\widehat{\sigma}^2_{M,i}(c))^{1/2}$, with $\widehat{\sigma}^2_{M,i}(c)$ some estimate of $\sigma^2_{M,i}(c)$. Nearest-neighbor estimators that defines $\widehat{\sigma}^2_{M,i}(c)$ as the squared difference between the outcome of unit i and the average outcome among its nearest neighbors in terms of the running variable (Abadie and Imbens, 2006; Abadie et al., 2014) are a common recommendation in the RD literature for this purpose (e.g. Calonico et al., 2014; Armstrong and Kolesár, 2018, 2020b). However, such a standard error is actually not uniformly consistent over $\mathcal{F}$ because the leading bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the first derivative of $\mu_M(\cdot,c)$ at X_i (Abadie and Imbens, 2006), which is unbounded over $\mathcal{F}$. We therefore propose a novel nearest-neighbor procedure in which the local sample average is replaced with a best linear predictor.

Specifically, let R 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 $r(j,i) \le Q_i$, where Q_i is the smallest integer such that $\mathcal{R}_i$ contains at least R elements, and let 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 the set of unit i's R nearest neighbors' indices; but with ties in the data R_i could be greater than R. We then define $\widehat{\sigma}^2_{M,i}(c)$ as the scaled squared difference

between $M_i(c)$ and its best linear predictor given its R_i nearest neighbors:

$$\widehat{\sigma}^2_{M,i}(c) = \frac{1}{1+H_i}\left(M_i(c) - \widehat{M}_i(c)\right)^2, \text{ with}$$
$$\widehat{M}_i(c) = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top M_j(c), \quad H_i = \widetilde{X}_i\left(\sum_{j\in\mathcal{R}_i}\widetilde{X}_j^\top\widetilde{X}_j\right)^{-1}\widetilde{X_i}^\top.$$

Here $\widetilde{X}_i = (1, X_i)^\top$ if the running variable takes at least two distinct values among the R_i nearest neighbors of unit i, and $\widetilde{X}_i = 1$ otherwise. The scaling term H_i, whose form follows from standard regression theory, ensures that $\widehat{\sigma}^2_{M,i}(c)$ is approximately unbiased in large samples. The next result shows that our new standard error is indeed uniformly consistent.

Theorem 4. *Suppose that Assumption 1, Assumption 2(i), and either Assumption LL1 or Assumption LL2 are satisfied. Then Assumption 2(ii) holds for the standard error described in this subsection.*

This result holds because the bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the second derivative of $\mu_M(\cdot, c)$ at X_i, which is bounded in absolute value over $\mathcal{F}$ by $B_Y + |c|B_T$. In contrast, the result would not hold for the conventional nearest-neighbor estimator, whose bias is proportional to the first derivative of $\mu_M(\cdot, c)$ at X_i and therefore unbounded. We therefore recommend using our standard error not just the construction of our CS, but more generally in bias-aware inference problems that work with bounds on second derivatives. We use $R = 5$ in the simulations and the empirical application in this paper.

6.2. **Bandwidth Choice.** An obvious candidate for a feasible bandwidth is the empirical analogue of $h_M(c)$, which minimizes the length of the auxiliary CI in Section 4:

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

While this choice is attractive in principle, in finite samples it could yield some coverage distortions if $B_Y + |c|B_T$ is very large relative to sampling uncertainty. To see why, recall from the discussion at the end of Section 5.1 that asymptotic normality of $\widehat{\tau}_M(h, c) = \sum_{i=1}^n w_i(h)M_i(c)$ follows from a CLT if $w_{\text{ratio}}(h) = o_P(1)$. Normality should thus be a "good" finite-sample approximation if $w_{\text{ratio}}(h)$ is "close" to zero. If $B_Y + |c|B_T$ is large, however, $\widehat{h}^*_M(c)$ is typically small in order to control the bias. The weights $w_i(\widehat{h}^*_M(c))$ then concentrate on few observations close to the cutoff, $w_{\text{ratio}}(\widehat{h}^*_M(c))$ is large, and CLT approximations could be inaccurate as $\widehat{\tau}_M(\widehat{h}^*_M(c), c)$ then effectively behaves like a sample average of a small number of observations.

To address this issue, we propose imposing a lower bound on the bandwidth, chosen such that the value of $w_{\text{ratio}}(h)$ remains below some reasonable threshold constant $\eta > 0$:

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

To motivate a choice for 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 $h = 1$. In this case $w_{\text{ratio}}(h) \approx .075$, and a CLT approximation should be reasonably accurate for $\widehat{\tau}_M(h, c)$, which is a weighted least squares estimator with 50 observations on each side of cutoff. Choosing $\eta \in [0.05, 0.1]$ therefore seems reasonable in practice; and we actually use $\eta = .075$ in our simulations.

As $\widehat{h}_M(c) \geq \widehat{h}^*_M(c)$, the constrained bandwidth could over-smooth the data relative to the one that would be asymptotically optimal for inference. If that happens, the resulting increase in finite-sample bias is the cost for normality being a better finite-sample approximation. This trade-off seems worthwhile since our CS construction explicitly accounts for the exact bias through, while deviations from normality cannot be captured. Under standard conditions like Assumption LL1 or LL2 the lower bound on the bandwidth clearly never binds asymptotically, but it can improve the finite-sample coverage of our CSs. The same idea can also be used for SRD inference, and more generally in settings where the finite-sample accuracy of inference faces a similar "bias vs. normality" trade-off. For example, Armstrong and Kolesár (2020a) use our approach for inference on average treatment effects under unconfoundedness with limited overlap.

6.3. **Computation.** Although our CS is defined through an inversion argument, it can be computed rather efficiently. We start by noting that our CS can be written as

$$\mathcal{C}^\alpha_{\text{ar}} = \{c : \widehat{p}(c) \geq 0\}, \text{ where } \widehat{p}(c) = 1 - \alpha - F\left(\left|\frac{\widehat{\tau}_M(\widehat{h}_M(c), c)}{\widehat{s}_M(\widehat{h}_M(c), c)}\right|, \widehat{r}_M(\widehat{h}_M(c), c)\right), \tag{6.1}$$

and $F(\cdot, r)$ is the CDF of the $|N(r, 1)|$ distribution. Computing $\mathcal{C}^\alpha_{\text{ar}}$ thus reduces to finding the roots of $\widehat{p}(c)$. Algorithm 1 describes how this is implemented in the `R` package that we provide with this paper. The main idea is to first evaluate $\widehat{p}(c)$ on a coarse grid over the plausible range of θ to get a "rough" picture of $\widehat{p}(c)$, and then search for a root between grid points where the sign of $\widehat{p}(c)$ changes. Following the discussion after Theorem 2, we assume that the boundaries of a bias-aware CI for τ_T are not exactly equal to zero, and exploit that $(-\infty, a_1] \cup [a_2, \infty) \subset \mathcal{C}^\alpha_{\text{ar}}$ for some $a_1 < a_2$ if zero is contained in such a CI (this holds because the t-ratios of $\widehat{\tau}_M(h, c)$ and $\widehat{\tau}_T(h)$ become equal for $|c| \to \infty$). In line with the conjecture after Theorem 2, $\widehat{p}(c)$ turned out to have either two or no roots in all of our

Algorithm 1. Computes the CS $\mathcal{C}^{\alpha}_{\text{ar}}$ for θ given bounds B_Y and B_T on the second derivatives of μ_Y and μ_T, respectively, and the number R of nearest neighbors to be used for the variance estimates that enter standard error.

1. Pick an interval $[c_L, c_U]$ that covers the plausible range of θ, and define grid points $c_j = c_L + j(c_U - c_L)/J$ for $j = 0, \ldots, J$ and some integer $J \geq 2$.
2. Compute $\widehat{p}(c_j)$ as in (6.1) for $j = 0, \ldots, J$. If $\widehat{p}(c_j)$ and $\widehat{p}(c_{j+1})$ have different sign, use the `uniroot` algorithm to find a root of $\widehat{p}(\cdot)$ over the interval (c_j, c_{j+1}). Denote the number of roots found by $S \geq 0$, and the roots themselves by $a_1, \ldots, a_S$.
3. Compute $\mathcal{C}^{\alpha}_{T}$, a bias-aware CI for τ_T, the jump in treatment probability.
4. Return the bias-aware AR CS $\mathcal{C}^{\alpha}_{\text{ar}}$ according to the following rules.
   (a) If $0 \in \mathcal{C}^{\alpha}_{T}$ and $S = 0$, then return $\mathcal{C}^{\alpha}_{\text{ar}} = (-\infty, \infty)$.
   (b) If $0 \in \mathcal{C}^{\alpha}_{T}$, S is positive and even, $\widehat{p}$ is decreasing at a_s if s is odd, and increasing if s is even, then return $\mathcal{C}^{\alpha}_{\text{ar}} = (-\infty, a_1] \cup [a_2, a_3] \cup \ldots \cup [a_S, \infty)$.
   (c) If $0 \notin \mathcal{C}^{\alpha}_{T}$, S is increasing at a_s if s is odd, and decreasing if s is even, then return $\mathcal{C}^{\alpha}_{\text{ar}} = [a_1, a_2] \cup [a_3, a_4] \cup \ldots \cup [a_{S-1}, a_S]$.

   If none of the four conditions is satisfied, restart the algorithm with a larger interval $[c_L, c_U]$ and/or a larger number of grid points J.

numerical examples, but our algorithm does not assume that this is the case.

The runtime of Algorithm 1 is mostly driven by the computational cost of evaluating the function $\widehat{p}(c)$. This cost is rather low with efficient programming: even with $n = 10^5$ data points, our algorithm computes $\mathcal{C}^{\alpha}_{\text{ar}}$ in about 20 seconds on a standard desktop computer. For comparison, it takes the widely used `rdrobust` package about 45 seconds to compute a robust bias correction DM CI on the same machine with the same number of data points (with smaller samples there is generally no practically relevant difference between the computation times of the two packages). Much computation time can be saved by noting that the nearest-neighbor variance estimates do not have to be computed from scratch for every value of c. This is because $\widehat{\sigma}^2_{M,i}(c) = \widehat{\sigma}^2_{Y,i} + c^2\widehat{\sigma}^2_{T,i} - 2c\widehat{\sigma}_{YT,i}$ is a quadratic function in c, with coefficients given by two variance terms and one covariance term that need to be computed only once. Also note that computing $\widehat{h}_M(c)$ is not too costly, as the corresponding optimization problem only involves a single linear regression for every candidate value of the bandwidth. This step is much less involved than, say, leave-one-out cross validation, which would require n linear regressions for every candidate bandwidth.

6.4. **Choosing Smoothness Bounds.** In order to compute $\mathcal{C}_{\text{ar}}^{\alpha}$, one needs to specify values for the smoothness bounds B_Y and B_T. Such bounds cannot be estimated consistently without imposing strong additional assumptions; and without specifying such bounds it is generally not possible to conduct inference on θ that is both valid and informative, even in large samples (Low, 1997; Armstrong and Kolesár, 2018; Bertanha and Moreira, 2018).

Roughly speaking, small values of B_Y and B_T amount to the assumption that the respective functions are "close" to linear on either side of the cutoff, whereas for larger values they are allowed to be increasingly "curved". This choice should be guided by subject knowledge, but in empirical applications there will generally be no single objectively right one. We hence recommend considering a range of plausible values as a form of sensitivity analysis. In the following subsections, we give some suggestions for how to determine such ranges, and for how to communicate their implications. For simplicity, we focus on the choice of B_Y, but the choice of B_T follows from analogous considerations.

We note that, as pointed out in the introduction, the need to specify smoothness bounds arises generally with bias-aware inference, but not with other popular methods like undersmoothing or robust bias correction. While at first glance this might seem like a disadvantage, in effect other methods also require such bounds to guarantee approximately correct CI coverage in practice.[9] Having to specify B_Y and B_T is thus not a meaningful impediment of our approach, but helps clarifying the assumptions on which inferential statements are based.

6.4.1. *Visualizing Smoothness Bounds.* Determining whether a particular value of B_Y is plausible in practice requires intuition for what functions are actually contained in $\mathcal{F}_H(B_Y)$. We suggest a procedure that visualizes some "extreme" elements of $\mathcal{F}_H(B_Y)$ to convey such intuition. Specifically, our proposal is to pick functions that match the scale of the data, and whose second derivative is equal to B_Y near the cutoff, through the following algorithm. Let $g(X_i)$ be a vector of basis transformations of X_i and its interaction with $\mathbf{1}\{X_i \geq 0\}$, with

[9]For example, an undersmoothing SRD CI can only be expected to have approximately correct coverage in finite samples if the bias is "small" relative to the standard error. With local linear estimation, this can only be the case if the underlying function is "close" to linear, which is equivalent to its maximum second derivative being "close" to zero. A similar point applies to robust bias correction, which in its standard implementation can only be expected to deliver CIs with approximately correct coverage in finite samples if the maximum third derivative of the underlying function is "close" to zero (Kamat, 2018). A researcher that reports such a CI and considers it reliable thus implicitly imposes a smoothness bound. If that bound was made explicit, however, a more efficient CI could be constructed through a bias-aware approach. See Armstrong and Kolesár (2020b) for more details on this point.

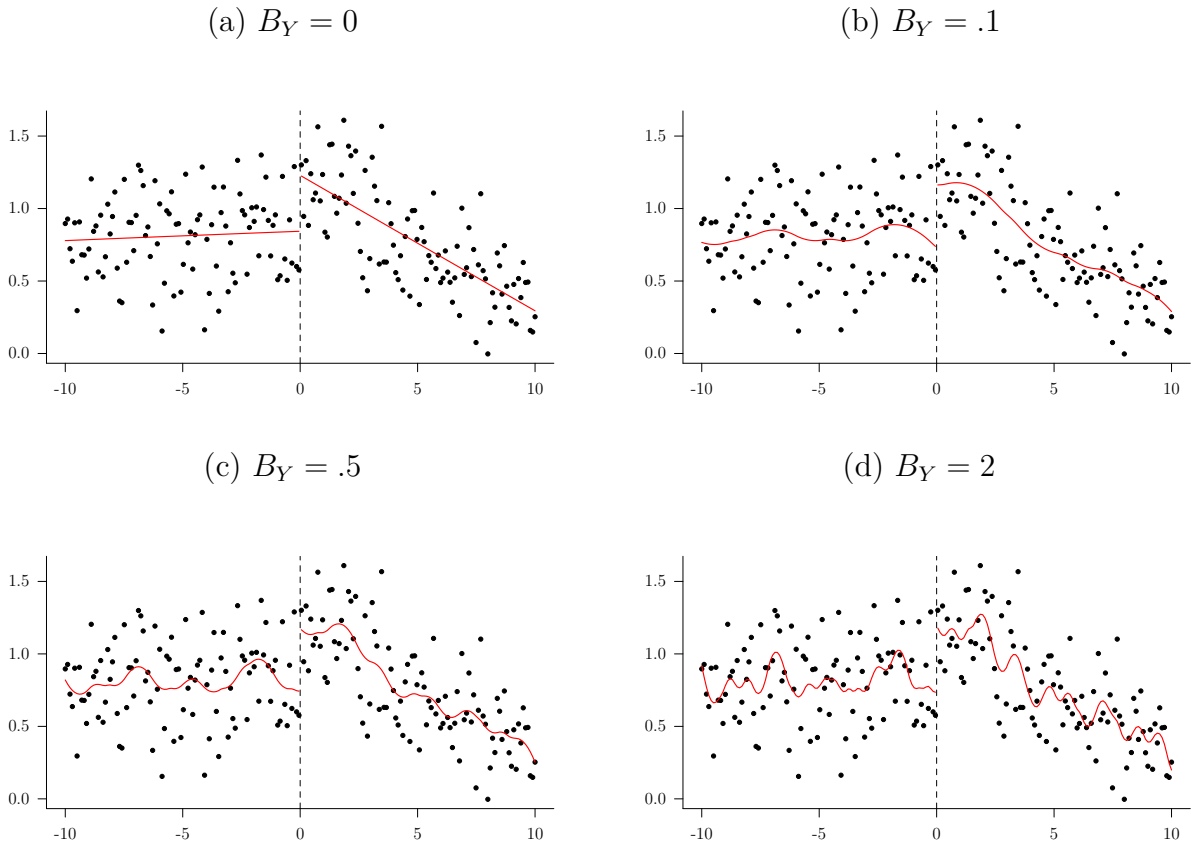


Figure 1: Examples of elements of $\mathcal{F}_H(B_Y)$ for various values of B_Y, each superimposed over the same hypothetical data set. Examples are constructed with $g(x)$ containing splines of order $k = 2$ and 50 knots on each side, and $\epsilon = .1$. Applied researchers can produce such graphs, and then pick the largest value of B_Y for which the resulting plot is empirically plausible. here the linear case in panel (a) is given for reference, panel (b) could be seen as adequate, panel (c) as a borderline case at best, and panel (d) would probably be considered implausible in economic applications.

sufficiently many entries for an OLS regression of Y_i on $g(X_i)$ to result in an erratic overfit of the data; and consider functions of the form $\widetilde{\mu}_Y(x) = g(x)^\top \widehat{\gamma}$, where $\widehat{\gamma}$ solves

$$\min_{\gamma} \sum_{i=1}^{n} (Y_i - g(X_i)^\top \gamma)^2 \text{ s.t. } \|g''(\cdot)^\top \gamma\|_\infty \leq B_Y, |g''(x_0)^\top \gamma| = |g''(-x_0)^\top \gamma| = B_Y,$$

for some $x_0 \geq 0$. The function $\widetilde{\mu}_Y$ is thus obtained by a constrained regression of Y_i on $g(X_i)$ in which the absolute second derivative is bounded by B_Y overall, and equal to B_Y near the cutoff. This optimization can easily be solved via quadratic programming.

We stress that $\widetilde{\mu}_Y$ is not supposed to be a good estimate of μ_Y, but simply an example of an "extreme" element of $\mathcal{F}_H(B_Y)$. The idea is to plot this function for various values

of B_Y (and possibly x_0) to obtain a better understanding for what kind of functions are contained in $\mathcal{F}_H(B_Y)$. For example, one could start with a very small B_Y, implying an almost linear function, and then increase the value in small steps until the resulting $\widetilde{\mu}_Y$ becomes implausibly erratic. Figure 1 illustrates this approach for a hypothetical data set.

6.4.2. *One-Sided CI for Smoothness Bound.* While it is not possible to obtain a valid data-driven upper bound on the curvature of μ_Y, it is possible to estimate a lower bound $\widehat{B}_{Y,\text{low}}$ for B_Y, and to compute a one-sided CI $[\widehat{B}^{\alpha}_{Y,\text{low}}, \infty)$ that covers B_Y with probability $1-\alpha$ in large samples (cf. Armstrong and Kolesár, 2018; Kolesár and Rothe, 2018). We recommend computing these quantities in empirical practice to guard against overly optimistic choices of the smoothness bounds.

6.4.3. *Rules of Thumb.* While it is not possible to consistently estimate the smoothness bounds from data, we are aware of two heuristic "rules of thumb" (ROT) that have been suggested as a way of determining plausible values in practice. Both rules are based on fitting global polynomial specifications $\widetilde{\mu}_{Y,k}$ of order k on either side of the cutoff by conventional least squares. Armstrong and Kolesár (2020b) consider fourth-oder polynomials, and propose the ROT bound $\widehat{B}_{Y,\text{ROT1}} = \sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,4}(x)|$, where $\mathcal{X}$ denotes the support of the running variable. Imbens and Wager (2019) consider a ROT in which the maximal curvature implied by a quadratic fit is multiplied by some moderate factor, say 2, to guard against overly optimistic values, yielding $\widehat{B}_{Y,\text{ROT2}} = 2\sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,2}(x)|$.

Such rules can provide a useful first guidance to choosing smoothness bounds, but they should be complemented with other approaches in a sensitivity analysis. We strongly recommend to always check the fit of the respective polynomial specification, and to dismiss the ROT value if the fit is obviously poor. In Online Appendix C, we compare the properties of ROT1 and ROT2 in a simple simulation study. We argue that in "roughly quadratic" settings the fourth-order polynomial specification that underlies ROT1 tends to produce quite erratic over-fits of the data. This leads to vast over-estimates of the true smoothness bounds, and corresponding CSs with poor statistical power. ROT2, on the other hand, tends to produce more reasonable values many such setups. See also our main Monte Carlo results in Section 7 for further details on this points.

7. SIMULATIONS

7.1. **Setup.** We now compare the practical performance of our bias-aware AR CS to that of alternative procedures though a Monte Carlo Study. We consider a number of data generating processes with varying curvature of the conditional expectation functions, richness

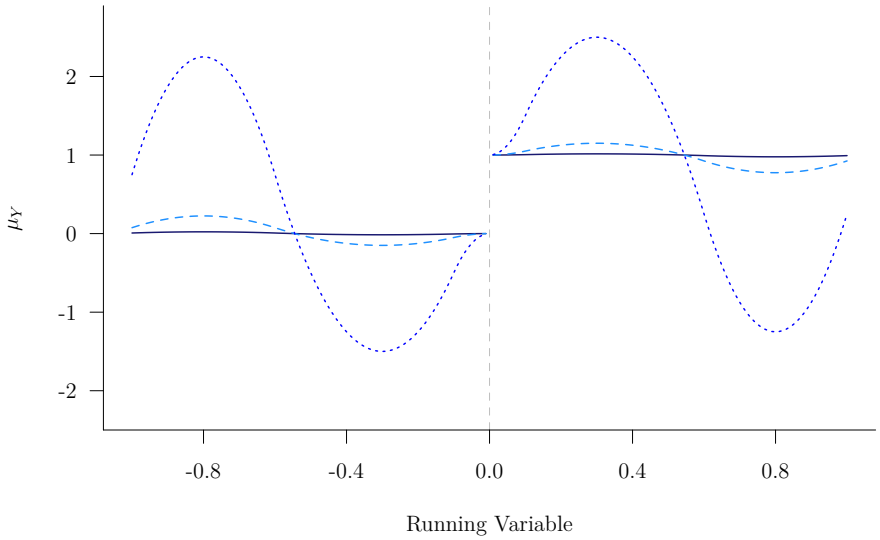


Figure 2: Conditional expectation function μ_Y for $\tau_Y = 1$ and various values of the smoothness bounds (solid line: $B_Y = 1$; dashed line: $B_Y = 10$; dotted line: $B_Y = 100$).

of the running variable's support, strength of identification. Specifically, we simulate X_i from either a continuous uniform distribution over $[-1, 1]$ or a discrete uniform distribution over $\{\pm 1/15, \pm 2/15, \ldots \pm 1\}$; and let

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

where $(\varepsilon_{1i}, \varepsilon_{2i})$ are bivariate standard normal with correlation 0.5, and $f(x) = x^2 - 1.5 \cdot \max(0, |x| - .1)^2 + 1.25 \cdot \max(0, |x| - .6)^2$. The functions μ_Y and μ_T are then second order splines with maximal absolute second derivative B_Y and B_T, respectively, over $[-1, 1]$. Figure 2 shows μ_Y for different values of B_Y. We consider the parameter values $(\tau_Y, \tau_T) \in \{(1, .2), (.5, .1)\}$, so that $\theta = 2$ in all settings, $B_T \in \{.2, 1\}$, and $B_Y \in \{1, 10, 100\}$; and set the sample size to $n = 1,000$. We refer to DGPs with $\tau_T = .1$ as weakly identified, and those with $\tau_T = .5$ as strongly identified.

We consider the performance of eight different AR CSs in our simulations: (i) our bias-aware CS, using the true B_Y and B_T; (ii) our bias-aware CS, using twice the true B_Y and B_T; (iii) our bias-aware CS, using half the true B_Y and B_T; (iv) our bias-aware CS, using ROT1

estimates of B_Y and B_T; (v) our bias-aware CS, using ROT2 estimates of B_Y and B_T; (vi) a naive CS that ignores bias, using an estimate of the "pointwise-MSE optimal" bandwidth (Imbens and Kalyanaraman, 2012, henceforth IK); (vii) an undersmoothing CS, using $n^{-1/20}$ times the estimated IK bandwidth;[10] (viii) a robust bias correction CS, using local quadratic regression to estimate the bias, and estimated IK bandwidths. In addition, we also consider the performance of eight different DM CIs using the just-mentioned approaches to handling bias. Note that DM CIs based on undersmoothing and robust bias correction are currently the most common CSs in empirical FRD studies.[11]

7.2. **Results.** Table 1 shows simulated coverage rates of the various CSs we consider for $\theta = 2$. We first discuss results for AR CSs, shown in the left panel. With the true smoothness bounds, coverage rates our bias-ware CSs are close to and mostly slightly above 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. The slight overcoverage occurs because the function $\mu_Y(x) - \theta\mu_T(x)$ is not exactly quadratic, and thus does not achieve the worst-case bias. Using twice or half the true bounds mostly leads to minor increases and decreases in simulated coverage, respectively. Using ROT1 for the smoothness bounds leads to over-coverage, especially for setups with a discrete running variable. This is because the underlying global quadratic approximation tends to severely over-estimate the smoothness bounds in our DGPs. ROT2 bounds generally lead to good coverage except for DGPs with $B_Y = 100$, where the underlying quadratic approximation leads to severe under-estimates of the smoothness bounds. Combining a naive approach, undersmoothing, or robust bias

[10]This CS corresponds to the one proposed by Feir et al. (2016) with a particular implementation of undersmoothing. Undersmoothing could in principle be implemented in a variety of ways, and hence the performance of the resulting CS must be interpreted accordingly.

[11]All computations are carried out with the statistical software package `R`. All bias-aware CSs are computed using our own software, which builds on the package `RDHonest`. All other CSs are computed using functions from the package `rdrobust`. A triangular kernel is used in all cases. Note that all CSs are only well-defined if the respective bandwidths are such that positive kernel weights are assigned to at least two (or three, in case of robust bias correction) distinct points on either side of the cutoff. In our simulations, the IK bandwidth estimates computed by `rdrobust` often do not satisfy this criterion if the running variable is discrete. We then manually set the bandwidth to 4/15, so that positive weights are given to three support points on each side of the cutoff. We also carried out a variant of our simulations in which we replace the IK bandwidth with the "coverage error optimal" bandwidth proposed by Calonico et al. (2018), using again the implementation in `rdrobust`. The results, which are qualitatively very similar to the ones reported in this section, are reported in Appendix F.

correction with an AR construction leads to CSs that undercover in all DGPs we consider, with the distortions being more severe (up to about 30 percentage points) for those with larger values of B_Y and B_T.

Turning to result for DM CIs in the right panel of Table 1, we see that combining a bias-aware approach with this construction does not necessarily lead to a CI with correct coverage even under strong identification. This is because bias-aware DM CIs only control the bias of a first-order approximation of the estimator on which they are based. Such coverage distortions are further amplified by weak identification in our simulations. Discreteness of the running variable does not have a strong detrimental effect on bias-aware DM CIs in this particular setup though. Using the ROT choices for the smoothness bounds leads to further distortions in some cases. The coverage of DM CIs that use the naive approach, undersmoothing, or robust bias correction is again distorted for most DGPs, with particularly severe deviations for weak identification and large values of the smoothness constants.

To show that our bias-aware AR CSs not only have good coverage properties, but also yield comparatively powerful inference, we simulate the rates at which the various CSs we consider cover parameter values other than the true one. We report the results for the DGP with $(B_Y, B_T) = (1, .2)$ and strong identification in Figure 3.[12] To avoid having all 16 coverage curves in one plot, we split the results into four panels: the five bias-aware AR CSs in (a), the three other AR CSs in (b), the five bias-aware DM CIs in (c), and the three other DM CIs in (d). Panels (b)–(d) also show the curve for our bias-aware AR CS with the true constants to have a common point of reference.

Panel (a) then shows that the coverage rate of bias-aware AR CSs drops very quickly to zero away from the true parameter, except for the CS based on ROT1 (which, as mentioned above, severely overestimates the smoothness bounds). Panels (b)–(d) show that the coverage of bias-aware AR CSs is also below that of all competing procedures over almost all the parameter space. This confirms that the accurate coverage of our CSs in settings with discrete running variables and weak identification does not come at the expense of statistical power in a canonical setup, for which most competing CS are specifically constructed.

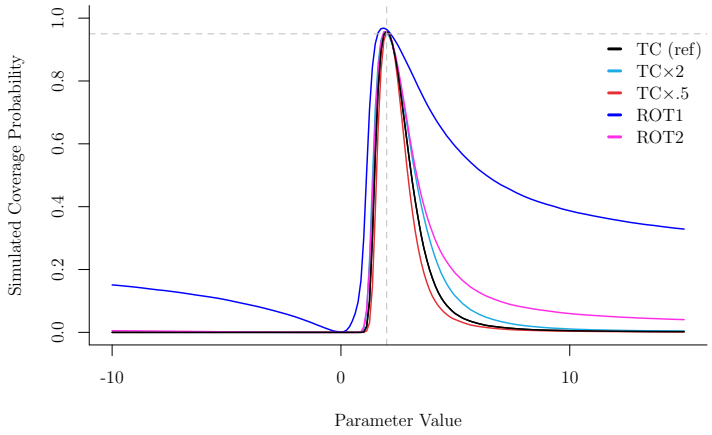
[12]We focus on these results because the coverage of the true parameter is reasonably close to the nominal level for all procedures, and thus comparison of coverage rates at "non-true" parameter values is meaningful across CSs. Analogous plots for other DGPs are available from the authors.

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

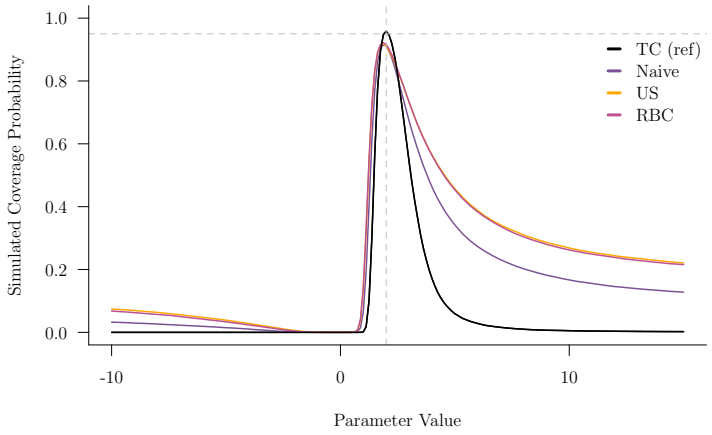
			Anderson-Rubin								Delta Method							
			Bias-Aware								Bias-Aware							
τ_T	B_Y	B_T	TC	TC×2	TC×.5	ROT1	ROT2	Naive	US	RBC	TC	TC×2	TC×.5	ROT1	ROT2	Naive	US	RBC
Running Variable with Continuous Distribution																		
0.5	1	0.2	97.2	97.1	96.9	96.4	96.9	93.1	93.4	93.4	97.3	96.6	97.6	92.6	95.3	90.8	90.4	91.3
0.5	1	1.0	96.7	96.5	96.7	96.4	96.5	93.0	93.3	93.3	95.8	94.5	97.5	92.6	95.3	90.4	89.9	91.0
0.5	10	0.2	95.8	95.6	96.3	96.8	96.4	92.4	93.0	92.5	95.0	94.5	95.3	93.2	94.3	88.3	88.3	88.7
0.5	10	1.0	95.5	95.4	96.2	96.6	96.1	92.2	92.8	92.2	94.5	93.9	95.7	93.0	94.0	87.9	88.2	88.4
0.5	100	0.2	95.1	98.9	88.8	99.5	86.1	78.5	87.9	74.7	94.5	98.0	86.0	98.1	79.1	72.8	80.8	72.2
0.5	100	1.0	95.1	99.0	88.6	99.5	86.0	78.2	88.0	74.4	93.3	97.8	87.1	98.1	79.9	72.6	80.9	72.1
0.1	1	0.2	97.2	97.3	96.8	97.1	97.3	93.7	94.0	94.0	92.3	90.0	92.0	79.0	87.0	76.6	74.0	79.2
0.1	1	1.0	97.3	97.1	96.8	97.1	96.9	93.4	93.8	93.8	89.6	86.5	93.2	78.7	87.6	76.5	73.8	79.0
0.1	10	0.2	96.9	96.6	97.1	97.4	97.1	93.4	94.0	93.7	84.1	85.8	82.8	79.3	82.6	71.0	69.6	73.3
0.1	10	1.0	96.9	96.7	97.1	97.5	97.1	93.2	93.9	93.5	85.0	83.0	87.2	78.8	83.5	70.3	69.2	72.7
0.1	100	0.2	96.3	99.2	91.5	99.6	89.7	83.2	91.0	79.0	83.6	96.0	65.4	92.3	54.4	36.7	47.4	37.1
0.1	100	1.0	96.4	99.2	91.7	99.6	89.8	83.2	91.0	79.1	82.4	94.4	72.6	92.3	58.1	36.5	47.4	37.0
Running Variable with Discrete Distribution																		
0.5	1	0.2	97.4	97.6	96.9	99.3	97.9	94.3	94.6	94.6	97.6	97.2	97.8	95.4	96.0	89.9	88.9	91.0
0.5	1	1.0	97.5	97.7	96.9	99.2	97.5	94.0	94.2	94.4	96.5	95.5	98.0	95.2	96.2	89.6	88.5	90.5
0.5	10	0.2	97.7	98.2	97.6	99.5	95.8	93.6	93.9	93.7	95.9	96.0	95.7	96.0	95.2	85.3	84.8	85.4
0.5	10	1.0	97.7	98.2	97.7	99.5	94.6	93.6	93.7	93.6	96.5	96.6	96.9	95.8	95.1	84.6	84.4	84.6
0.5	100	0.2	96.9	100.0	91.2	100.0	86.2	67.9	60.4	57.8	95.0	97.5	48.1	98.8	25.3	26.8	27.9	17.1
0.5	100	1.0	96.8	100.0	91.0	100.0	85.8	67.2	59.5	57.2	93.1	98.0	54.0	98.7	26.8	26.5	27.6	16.7
0.1	1	0.2	97.5	97.9	96.6	99.5	98.1	94.7	94.9	95.1	92.7	89.4	92.1	79.2	87.5	71.2	64.8	75.6
0.1	1	1.0	97.8	98.1	96.9	99.4	97.9	94.5	94.7	94.7	90.0	86.8	93.5	78.6	88.3	70.5	64.5	74.8
0.1	10	0.2	98.5	99.0	98.1	99.6	96.2	94.5	94.5	94.6	82.3	88.9	78.9	75.0	86.3	55.9	52.6	59.9
0.1	10	1.0	98.5	99.0	98.2	99.6	95.3	94.5	94.5	94.6	82.7	84.7	85.0	74.9	86.8	55.8	51.8	59.6
0.1	100	0.2	97.2	100.0	93.6	100.0	91.5	73.7	66.9	63.9	94.9	96.3	94.3	96.6	89.9	68.6	69.5	65.3
0.1	100	1.0	97.3	100.0	93.8	100.0	91.5	73.2	66.4	63.0	95.3	96.1	96.3	96.8	91.1	68.1	69.0	64.8

Notes: Results based on 50,000 Monte Carlo draws for a nominal confidence level of 95%. Columns show results for bias aware approach with true constants (TC), two times true constants (TC×2), half true constants (TC×.5), and with rule of thumb estimates (ROT1) and (ROT2); naive approach that ignores bias (Naive); undersmoothing (US); and robust bias correction (RBC).

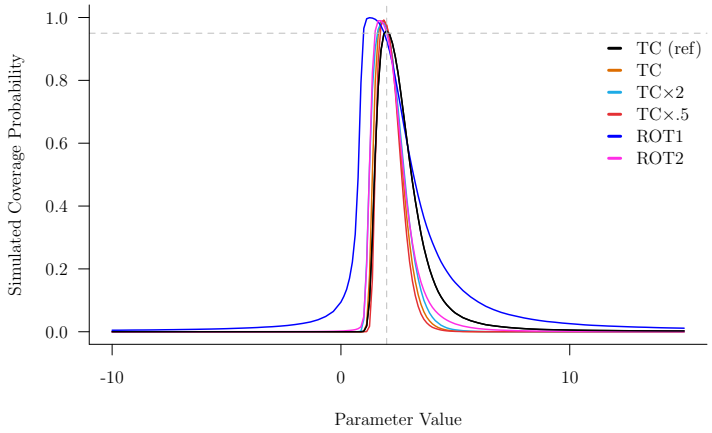
(a) Bias-aware Anderson-Rubin CSs



(b) Other Anderson-Rubin CSs



(c) Bias-Aware Delta Method CI



(d) Other Delta Method CIs

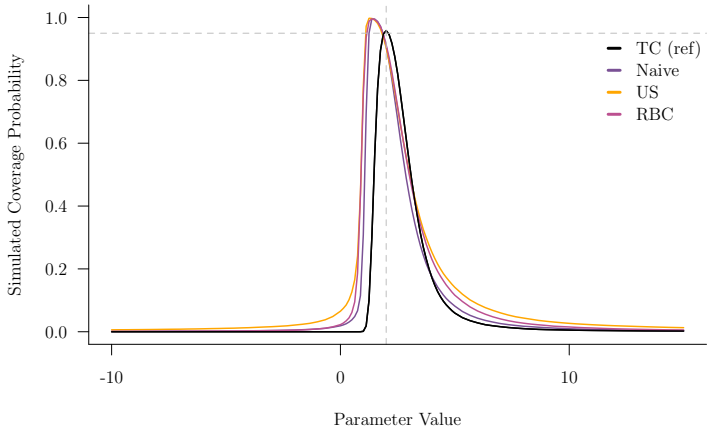


Figure 3: Simulated coverage rates of various values of parameter values and for different types of confidence sets. Based on the DGP described in the main text with $\tau_T = .5$, $B_T = .2$, and $B_Y = 1$. Bias aware approach with true constants (TC (ref); as reference function in all graphs), two times true constants (TC×2), 0.5 times true constants (TC×.5), and with rule of thumb smoothness bounds (ROT1) and (ROT2); naive approach that ignores bias (Naive); undersmoothing (US); and robust bias correction (RBC).

8. EMPIRICAL APPLICATION

In this section, we apply our methods to data from Oreopoulos (2006, 2008), who studies the effects of a 1947 education reform in Great Britain that raised the minimum school-leaving age from 14 to 15 years. The data are a sample of $n = 73,954$ workers who turned 14 between 1935 and 1965, obtained by combining the 1984-2006 waves of the UK General Household Survey. We take the effect of attending school beyond age 14 on annual earnings measured in 1998 UK pounds as the parameter of interest. The running variable is the year in which the worker turned 14, and the threshold is 1947. Figure 4 shows the average of log annual earnings and 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 the analysis for both the entire data and the subset that excludes the 1947 cohort. Oreopoulos (2006) uses a parametric approach in which the respective dependent variable is regressed on a dummy for turning 14 in or after 1947 and a 4th order polynomial in age. This yields the estimate $\widehat{\theta} = .146$ with a 95% DM CI $[-.009; .300]$ based on a heteroscedasticity-robust standard error for the entire data, and $\widehat{\theta} = .111$ with a 95% DM CI $[-.032; .255]$ if the 1947 cohort is excluded.[13] These CIs, however, do not account for the model misspecification bias one should expect here.

To compute our bias-aware AR CSs, we first have to determine plausible values for the smoothness constants B_Y and B_T. To do that, we compute the ROT values, the lower bound estimates and one-sided CIs, and various graphs of candidate functions, all as described in Section 6.4. All graphs are shown in Appendix E. Regarding the value of B_Y, inspection of the top panel of Figure 4 suggests that the function μ_Y should not be too erratic. Indeed, we estimate a lower bound of $\widehat{B}_{Y,\text{low}} = 0$ for B_Y, meaning that the data cannot rule out that μ_Y is linear. We also have $\widehat{B}_{Y,\text{ROT1}} = .023$ and $\widehat{B}_{Y,\text{ROT2}} = .012$, with the fit of the underlying polynomials seeming adequate in both cases. Including also some conservative values, we then consider $[0; .04]$ as a plausible range for B_Y.

Regarding the choice of B_T, one has to be more careful. From the bottom panel of Figure 4, we see that the empirical share of "treated" students increases very slowly after

[13]The numerical result here differ from those in Oreopoulos (2006) because (i) we use the data set from its online corrigendum (Oreopoulos, 2008), which includes additional waves of the UK General Household Survey; (ii) Oreopoulos (2006) considers a slightly different parameter of interest; and (iii) Oreopoulos (2006) uses Lee and Card (2008) standard errors that are clustered by the running variable. Kolesár and Rothe (2018) show that such clustering does not alleviate the issues caused by a discrete running variable, but tends to produce CIs with poor coverage properties, and hence such standard errors should not be used.

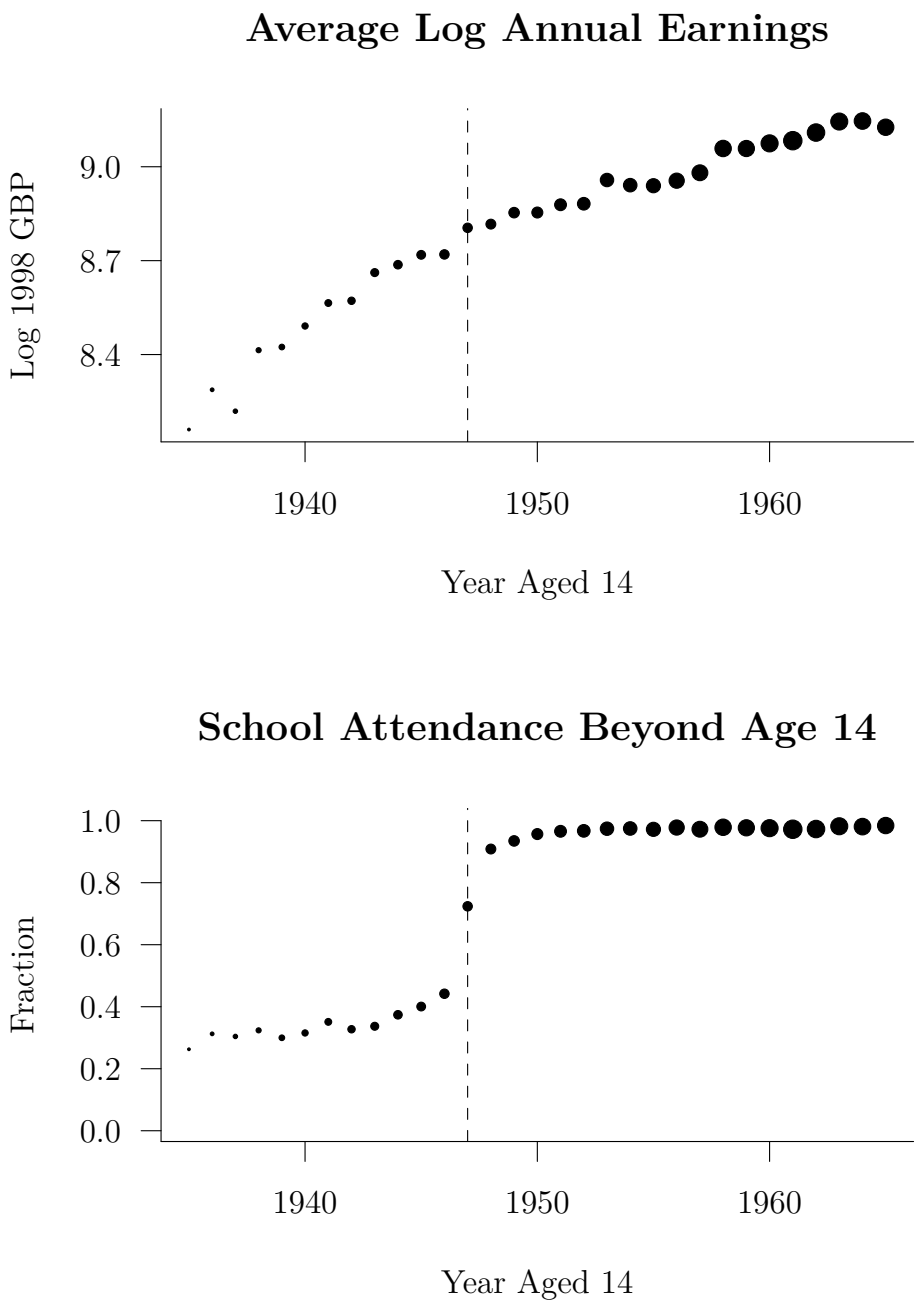


Figure 4: Average log annual earnings (top panel) and fraction of individuals in full time education beyond age 14 by birth year cohort. Dashed vertical lines indicate the year 1947, in which the minimum school leaving age changed from 14 to 15 years. Size of dots is proportional to the cohort size in the data.

Table 2: Bias-aware Anderson-Rubin confidence sets for the effect of one additional year of compulsory schooling for various values of the smoothness bounds

	B_Y				
B_T	0	.01	.02	.03	.04
	Panel A: Results for full data set				
.12	[-.239; 1.841]	[-.366; 1.953]	[-.458; 2.068]	[-.555; 2.183]	[-.655; 2.301]
.14	[-.343; 2.395]	[-.448; 2.554]	[-.569; 2.716]	[-.694; 2.881]	[-.824; 3.049]
.16	[-.432; 3.608]	[-.591; 3.887]	[-.762; 4.174]	[-.941; 4.467]	[-1.131; 4.767]
.18	[-.637; 10.049]	[-.907; 11.152]	[-1.217; 12.279]	[-1.575; 13.427]	[-1.995; 14.590]
.20	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$	$(-\infty; \infty)$
	Panel B: Results excluding data for 1947				
0	[-.108; .080]	[-.152; .441]	[-.237; .546]	[-.313; .619]	[-.386; .687]
.01	[-.100; .224]	[-.168; .496]	[-.257; .589]	[-.338; .665]	[-.415; .733]
.02	[-.117; .406]	[-.187; .554]	[-.280; .638]	[-.367; .714]	[-.459; .778]
.03	[-.125; .495]	[-.208; .606]	[-.307; .692]	[-.400; .765]	[-.489; .825]
.04	[-.126; .566]	[-.232; .664]	[-.340; .749]	[-.439; .814]	[-.522; .879]

Notes: All CSs have 95% nominal level. Results based on 73,954 data points for Panel A and 73,954 data points for Panel B. See main text for a justification of the smoothness bounds value considered.

1948, but jumps sharply from 0.724 for 1947 to 0.909 for 1948. If we consider the latter change to be natural variation in treatment probabilities, then only rather large values of B_T are consistent with the data. Indeed, we estimate a lower bound $\widehat{B}_{T,\text{low}} = .158$, with a 95% one-sided CI of $[0.126; \infty)$. The two ROTs yield much smaller values, namely $\widehat{B}_{T,\text{ROT1}} = .031$ and $\widehat{B}_{T,\text{ROT2}} = .011$. But since the fit of both underlying polynomial specifications is poor we choose to disregard these values, and consider $[.12; .2]$ as a plausible range for B_T. The upper end was chosen because it turns out that for $B_T \geq .2$ our CS is always equal to the real line, and thus considering larger values would not affect the results.

If we take the arguably more realistic position that the change in treatment probabilities between 1947 and 1948 was largely caused by delayed implementation of the reform, a more natural approach is to exclude the 1947 cohort and conduct a "donut" analysis. We then estimate a lower bound $\widehat{B}_{T,\text{low}} = 0$ for B_T, meaning that linearity of μ_T cannot be ruled out, and the ROTs yield $\widehat{B}_{T,\text{ROT1}} = .013$ and $\widehat{B}_{T,\text{ROT2}} = .009$, with the fitted polynomial being adequate in both cases. To also include some conservative values, we then consider $[0; .04]$ as a plausible range for B_T in this donut setup.

In Table 2, then we report bias-aware AR CSs with nominal level 95%, separately for the

entire data (top panel) and for the subsample that excludes the 1947 cohort (bottom panel), and for values of B_Y and B_T in regular grids over the ranges motivated above. All CSs in panel (a) are extremely wide, in the sense that even the shortest one is much larger than all plausible values for the return to increased compulsory schooling. This is because treating the sharp increase in treatment probability from 1947 to 1948 as natural variation implies that the parameter of interest is only weakly identified. In panel (b), which excludes 1947 cohort data, and considers an appropriate range for B_T, the CSs become much shorter, but they still all cover zero and many contain the full plausible parameter space.

Our overall preferred specification is the one that excludes the 1947 cohort, and uses $B_Y = .02$ and $B_T = .01$ (the grid values in between the respective ROT estimates), which yields the bias-aware AR CS $[-.257, .589]$. This CS is almost three times as large as the reference CS $[-.032; .255]$ based on the parametric specification. Overall, the data are not very informative about the returns to schooling.

9. CONCLUSIONS

FRD 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 AR 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.

A. PROOFS OF MAIN RESULTS

In this Appendix, we prove the main results from Section 5. We use repeatedly that, using basic least squares algebra, the statistic $\widehat{\tau}_M(h,c)$ can be written as

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

$$w_{i,+}(h) = e_1^\top Q_+^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_+ = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i' \mathbf{1}\{X_i \geq 0\}$$

$$w_{i,-}(h) = e_1^\top Q_-^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_- = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i' \mathbf{1}\{X_i < 0\},$$

with $\widetilde{X}_i = (1, X_i)'$. To simplify the notation, throughout the proofs we write $A_n(\mu) = o_{P,\mathcal{F}}(1)$ if $\sup_{\mu\in\mathcal{F}} P(|A_n(\mu)| > \epsilon) = o(1)$ for all $\epsilon > 0$ and a generic sequence $A_n(\mu)$ of random variables indexed by $\mu \in \mathcal{F}$. We also drop the dependency on c from the notation for the optimal bandwidth in most instances, writing h_M instead of $h_M(c)$.

A.1. **Proof of Theorem 1.** We first establish the following lemma.

Lemma A.1. *Suppose that Assumption 1–2 and either Assumption LL1 or Assumption LL2 are satisfied. Then the following holds uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$: (i) $w_{\text{ratio}}(h_M(c)) = o_P(1)$; (ii) $(\widehat{\tau}_M(\widehat{h}_M(c), c) - \widehat{\tau}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$; and (iii) $(\bar{b}_M(\widehat{h}_M(c), c) - \bar{b}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$.*

Proof. We first show part (i). Suppose that Assumption LL1 is satisfied. With probability approaching 1, we have that

$$\max_{i\in\{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j=1}^n w_j(h_M)^2} \leq \max_{i\in\{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j:X_j=X_i} w_j(h_M)^2} = \max_{i\in\{1,\ldots,n\}} \frac{1}{\sum_{j:X_j=X_i} \mathbf{1}\{X_i = X_j\}}.$$

As $n \to \infty$, the number of units whose realization of the running variable is equal to any particular value in its support tends to infinity, and we obtain the statement of the lemma.

Now suppose that Assumption LL2 is satisfied. First, it is easy to see that the minimizer of $\text{cv}_{1-\alpha}(r_M(h, c)) \cdot s_M(h, c)$ must satisfy $h_M \to 0$ and $nh_M \to \infty$ as $n \to \infty$. Under these conditions, the bias and variance of the local linear regression estimator scale as h_M^2 and $1/(nh_M)$, respectively. From the properties of the function $\text{cv}_{1-\alpha}(\cdot)$, it then follows that $h_M \propto n^{-1/5}(1 + o_P(1))$. It also holds that

$$\max_{i\in\{1,\ldots,n\}} \frac{w_i(h_M)^2}{\sum_{j=1}^n w_j(h_M)^2} \leq \max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} + \max_{i:Z_i=0} \frac{w_i(h_M)^2}{\sum_{j:Z_j=0} w_j(h_M)^2}.$$

It then suffices to show that the first term on the right-hand side of the last inequality tends to zero in probability uniformly over $\mathcal{F}$, as the same arguments can be used to prove an analogous result for the second term. Note that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} \\
&\quad = \max_{i:Z_i=1} \frac{K(X_i/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_i \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}{\sum_{j:Z_j=1} K(X_j/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}.
\end{aligned}$$

Treating the numerator of the right-hand side of the second line as a function of X_i, it follows from the fact that the kernel is bounded from above by Assumption 1 that this

function is bounded from above by a quadratic function in $X_i \in [0, h]$. The maximum of this quadratic function is bounded by a constant multiplied by $[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2]^2 + h_M^2[\sum_{l:Z_l=1} X_l K(X_l/h_M)]^2$. Taken together, this means that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h_M)^2}{\sum_{j:Z_j=1} w_j(h_M)^2} \\
&\quad \leq C \frac{(\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2)^2 + h_M^2(\sum_{l:Z_l=1} X_l K(X_l/h_M))^2}{\sum_{j:Z_j=1} K(X_j/h_M)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h_M)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h_M)]^2}.
\end{aligned}$$

for some finite constant C, and for n sufficiently large. Standard kernel calculations than yield that the numerator on the right-hand side of the last inequality is an $O_P(n^2 h_M^2)$ term, while the denominator an $O_P(n^3 h_M^3)$ term. As $nh_M \to \infty$ as $n \to \infty$, this completes part (i).

Now consider part (ii)–(iii). Suppose Assumption LL1 holds. With a discrete running variable, it is clear that the optimal bandwidth h_M shrinks with the sample size, but it cannot tend to zero as it has to be greater than the support point second closest to the cutoff in order for the local linear regression estimator to be well-defined. Furthermore, any bandwidth h between the second and third support point closest to the cutoff implies the same local linear regression weights $w_i(h)$ for all i. Hence any bandwidth between the second and third support point closest to the cutoff is asymptotically optimal. Part (ii)–(iii) then follow trivially, as each expression under consideration depends on h only through $w_i(h)$.

Now suppose that Assumption LL2 holds. Statements (ii)–(iii) of Lemma A.1 then follow as in the proof of Theorem E.1 in Armstrong and Kolesár (2020b). □

We now proceed with the proof of the core statement of Theorem 1. Since $\theta \in \mathcal{C}^\alpha_{\text{ar}}$ if and only if $\tau_M(\theta) \in \mathcal{C}^\alpha(\theta)$, it suffices to show that for any $c \in \mathbb{R}$

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

Note that it follows from Lemma A.1 (ii)–(iii) and uniform continuity of $\text{cv}_{1-\alpha}(\cdot)$ that

$$\begin{aligned}
&\frac{|\widehat{\tau}_M(\widehat{h}_M, c) - \tau_M(c)|}{\widehat{s}_M(\widehat{h}_M, c)} - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M, c)) \\
&\quad = \left| \frac{\widehat{\tau}_M(h_M, c) - \mathbb{E}\left[\widehat{\tau}_M(h_M, c)|\mathcal{X}_n\right]}{s_M(h_M, c)} + \frac{b_M(h_M, c)}{s_M(h_M, c)} \right| - \text{cv}_{1-\alpha}(r_M(h_M, c)) + o_{P,\mathcal{F}}(1).
\end{aligned}$$

We now apply Lyapunov's CLT to show that $(\widehat{\tau}_M(h_M, c) - \mathbb{E}\left[\widehat{\tau}_M(h_M, c)|\mathcal{X}_n\right])/s_M(h_M, c)$ converges in distribution to a standard normally distributed random variable, uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$. Specifically, let C be a positive constant, let $\delta > 2$, and recall that

$\widehat{\tau}_M(h_M, c) = \sum_{i=1}^n w_i(h_M) M_i(c)$. Lyapunov's CLT can be applied conditional on $\mathcal{X}_n$ since

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}\left[|w_i(h_M)(M_i(c) - \mathbb{E}[M_i(c)|\mathcal{X}_n])|^\delta \,|\mathcal{X}_n\right]}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2 \sigma^2_{M,i}}\right)^\delta} \leq \lim_{n\to\infty} C \sum_{i=1}^n \frac{|w_i(h_M)|^\delta}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2}\right)^\delta}$$
$$\leq \lim_{n\to\infty} C \max_{i=1,\dots n} \left(\frac{|w_i(h_M)|}{\sqrt{\sum_{i=1}^n w_i(h_M)^2}}\right)^{\delta-2} = o_{P,\mathcal{F}}(1)$$

by Assumption 1(i)–(iii) and Lemma A.1(i). Standard arguments then yield that

$$\liminf_{n\to\infty} \left(\inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\tau_M(c) \in \mathcal{C}^\alpha(c)) - \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}\left(\left|S + \frac{b_M(h_M,c)}{s_M(h_M,c)}\right| \leq \mathrm{cv}_{1-\alpha}(r_M(h_M,c))\right)\right) = 0,$$

with S a generic standard normal random variable. The statement of the theorem now follows from the definition of the critical value function $\mathrm{cv}_{1-\alpha}(\cdot)$ if

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h_M,c)/s_M(h_M,c)| \leq r_M(h_M,c).$$

Note that Armstrong and Kolesár (2020b, Theorem B.3) show that the last statement holds with equality if μ_Y and μ_T have unbounded domain. In our setup, we only have a potentially weak inequality because μ_T is naturally constrained to take values in $[0,1]$, and the supremum is thus taken over a smaller set of functions. This completes our proof. □

A.2. **Proof of Theorem 2.** To simplify the exposition, we emphasize the dependence of various estimators on c in our notation, but suppress their dependency on the bandwidth h (which does not depend on c under the conditions of this theorem). The CS $\mathcal{C}^\alpha_{\mathrm{ar}}(h)$ is given by the set of all values of c satisfying

$$\vartheta(c) \leq 0, \quad \text{where} \quad \vartheta(c) \equiv |\widehat{\tau}_Y - c\widehat{\tau}_T| - \mathrm{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c).$$

The function $\vartheta(c)$ is continuous in c, as $\mathrm{cv}_{1-\alpha}(\cdot)$ is a uniformly continuous function, and both the standard error $\widehat{s}_M(c) = (\widehat{s}^2_Y - 2c\widehat{s}_{TY} + c^2\widehat{s}^2_T)^{1/2}$ and the worst case bias $\overline{b}_M(h,c) = -(B_Y + |c|B_T)/2 \cdot \sum_{i=1}^n w_i(h) X_i^2 \cdot \mathrm{sign}(X_i)$ are continuous in c. Moreover, the term $\mathrm{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is also strictly convex in c, because both the standard error and the worst-case bias are convex in c and $\mathrm{cv}_{1-\alpha}(\cdot)$ is strictly convex and increasing. The shape of $\mathcal{C}^\alpha_{\mathrm{ar}}(h)$ is then determined by the roots of $\vartheta(c)$. While one can in principle solve analytically for the roots of $\vartheta(c)$, doing so is very tedious.

To prove the theorem, it suffices to show that the function $\vartheta(c)$ always fits into one of

the following four categories: (i) $\vartheta(c) \leq 0$ for all c; (ii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) < 0$ for all $|c| > c^*$; (iii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) > 0$ for all $|c| > c^*$, and (iv) $\vartheta(c)$ has one root. Then $\mathcal{C}^{\alpha}_{\text{ar}}(h) = \mathbb{R}$ in case (i), $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$ for some $a_1 < a_2$ in case (ii); and by $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$ for some $a_1 < a_2$ in case (iii), and $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_2]$ or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, \infty)$ in case (iv). We now go through a number of case distinctions.

If $\widehat{\tau}_T = 0$, then $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a constant function in c. As $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex in c and unbounded, $\vartheta(c)$ must be either of form (i) or (ii). We threfore suppose that $\widehat{\tau}_T \neq 0$ from now on , and write $\widehat{\theta} = \widehat{\tau}_Y/\widehat{\tau}_T$. Since $\vartheta(\widehat{\theta}) < 0$ by construction, the function $\vartheta(c)$ cannot be strictly positive. As $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a piecewise linear function and $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex, the function $\vartheta(c)$ can also have at most two roots for $c \leq \widehat{\theta}$, and at most two roots for $c > \widehat{\theta}$. If it does not have any root, $\vartheta(c)$ is of the form (i).

Let us first assume that $\lim_{c\to\pm\infty} \vartheta(c) \neq 0$. It follows from basic algebra that there exists some c^* sufficiently large such that $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = 1$ or $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = -1$ and $\vartheta(c) \neq 0$ for all $c > c^*$. The function $\vartheta(c)$ therefore cannot have one or three roots; so it must have either four roots or two roots or none. If $\text{sign}(\vartheta(c)) = -1$ for all $|c| > c^*$, which means that $|\widehat{\tau}_Y - c\widehat{\tau}_T| > \text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$. The function $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ intersects once with the function $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ for $c < \widehat{\theta}$, and once for $c > \widehat{\theta}$. Therefore $\vartheta(c)$ must be of form (iii) in this case. If $\text{sign}(\vartheta(c)) = -1$ for all $|c| > c^*$, the above reasoning only yields that $\vartheta(c)$ has at most four roots. However, note that for $|c| \to \infty$ the absolute value of the first derivative of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ with respect to c converges to some constant ϖ, and that for any value of $\varsigma \in \mathbb{R}$ the expression $\text{sign}(c)\cdot(\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M - |\varsigma + \varpi \cdot c|)$ converges to a constant. Choose ς such that the latter constant is zero, and set $\varrho(c) = |\varsigma + \varpi c|$. By construction, $\varrho(c)$ intersects with $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ twice either for $c \leq \widehat{\theta}$ or $c \geq \widehat{\theta}$. It also holds that $\varrho(c) \leq \text{cv}_{1-\alpha}(\widehat{r}(c)) \cdot \widehat{s}_M(c)$ for all c by strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}(c)) \cdot \widehat{s}_M(c)$. This reasoning implies that $\vartheta(c)$ can have at most two roots, and must be of form (ii) in this case.

Now suppose that $\lim_{c\to\pm\infty} \vartheta(c) = 0$, which is an event that only occurs if $\widehat{\tau}_T = \pm\text{cv}_{1-\alpha}(\widehat{r}_T(c))\cdot \widehat{s}_T(c)$. It then follows from strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ that $\vartheta(c)$ cannot have three roots. $\vartheta(c)$ is therefore of form (i) if it does not have any root, and otherwise of form (iv). This completes the proof. □

A.3. **Proof of Theorem 3.** We begin by giving a formal description of a bias-aware DM CI. Recall the definition of U_i from Section 3.2, 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 conditional bias and standard deviation, respectively, of the SRD-type

estimator $\widehat{\tau}_U(h)$. Exploiting linearity, one can write

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

where $\mu_U(x) = (\mu_Y(x) - \tau_Y)/\tau_T - \tau_Y(\mu_T(x) - \tau_T)/\tau_T^2$ 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 $\mathcal{F}^\delta$ is

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

An infeasible bias-aware DM CI is then given by

$$\mathcal{C}_\Delta^\alpha = \left[\widehat{\theta}(h_U) \pm \text{cv}_{1-\alpha}\left(\overline{b}_U(h_U)/s_U(h_U) \right) s_U(h_U) \right],$$

where $h_U = \text{argmin}_h \, \text{cv}_{1-\alpha}\left(\overline{b}_U(h)/s_U(h) \right) s_U(h)$ is the bandwidth that minimizes its length.

Making this CI feasible would require three main modifications. First, replacing the unknown bias bound with an estimate $\widehat{\overline{b}}_U(h)$ which replaces τ_Y and τ_T with feasible estimates (obvious candidates would be local linear estimates $\widehat{\tau}_Y = \widehat{\tau}_Y(g_Y)$ and $\widehat{\tau}_T = \widehat{\tau}_T(g_T)$ based on preliminary bandwidths g_Y and g_T). Second, replacing the standard deviation $s_U(h)$ with a valid standard error (this could be achieved as in Section 6.1, using 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 U_i). Third, replacing the bandwidth h_U with a suitable empirical analogue (such as an adaptation of the restricted plug-in procedure described in Section 6.2). Since such modifications can be shown not to affect the asymptotic coverage properties of the CI under standard additional regularity conditions, we simply base our result on a comparison of $\mathcal{C}_*^\alpha$ and $\mathcal{C}_\Delta^\alpha$.

To prove Theorem 3, we now make the dependence of quantities like $h_M(c)$ on c again explicit in our notation. We begin by noting that the events $\theta^{(n)} \in \mathcal{C}_\Delta^\alpha$ and $\theta^{(n)} \in \mathcal{C}_*^\alpha$ occur if and only if

$$\frac{|\widehat{\theta}(h_U) - \theta^{(n)}|}{s_U(h_U)} - \text{cv}_{1-\alpha}\left(\frac{\overline{b}_U(h_U)}{s_U(h_U)} \right) \leq 0 \tag{A.1}$$

$$\text{and} \quad \frac{|\widehat{\tau}_M(h_M(\theta^{(n)}), \theta^{(n)})|}{s_M(h_M(\theta^{(n)}), \theta^{(n)})} - \text{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta^{(n)}), \theta^{(n)})}{s_M(h_M(\theta^{(n)}), \theta^{(n)})} \right) \leq 0, \tag{A.2}$$

respectively. Since the left-hand sides of the last two displays are both approximated by

a constant plus the absolute value of a normal random variable with variance 1 in large samples, it suffices to show that the difference between the respective left-hand sides of the last two displays converges to zero in probability, uniformly over $\mathcal{F}^\delta$. To show this, note first that standard delta method arguments yield that the left-hand side of (A.1) is equal to

$$\frac{|\widehat{\tau}_U(h_U) - \kappa n^{-2/5}|}{s_U(h_U)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_U(h_U)}{s_U(h_U)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Next, note that $U_i = M_i(\theta)/\tau_T$, and that we thus have that

$$\widehat{\tau}_U(h) = \frac{\widehat{\tau}_M(h,\theta)}{\tau_T}, \qquad s_U(h) = \frac{s_M(h,\theta)}{|\tau_T|}, \qquad \overline{b}_U(h) = \frac{\overline{b}_M(h,\theta)}{|\tau_T|},$$

for any $h > 0$. Substituting these identities into the definition of h_U, we also find that

$$h_U = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h,\theta)}{s_M(h,\theta)}\right) \cdot \frac{s_M(h,\theta)}{|\tau_T|} = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h,\theta)}{s_M(h,\theta)}\right) s_M(h,\theta) = h_M(\theta).$$

The left-hand side of (A.1) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T\kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Now consider the term on the left-hand side of (A.2). By simple algebra, we have that

$$\begin{aligned}
\overline{b}_M(h,\theta^{(n)}) &= \overline{b}_M(h,\theta) + n^{-2/5}|\kappa|\overline{b}_T(h), \\
s_M(h,\theta^{(n)}) &= s_M(h,\theta) + n^{-2/5}|\kappa|(s_T(h) - 2\tilde{s}_{M(\theta),T}(h)),
\end{aligned}$$

with $\tilde{s}_{M(\theta),T}(h) = (\sum_{i=1}^n w_i(h)^2\sigma_{M(\theta),T,i})^{1/2}$ a conditional covariance term of the same order as $s_T(h)$. These identities imply that evaluation at $\theta^{(n)}$ does not change the leading terms of the (conditional) bias and the standard deviation (which are of order h^2 and $1/\sqrt{nh}$, respectively) relative to evaluation at θ. Since the leading term of $h_M(\theta)$ is a smooth transformation of the leading terms of the bias and standard deviation, this means that $h_M(\theta^{(n)}) = h_M(\theta)(1 + o_{P,\mathcal{F}^\delta}(1))$. Arguing as in the proof of Lemma A.1, the left-hand side of (A.2) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T\kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1),$$

which completes the proof. □

REFERENCES

Abadie, A. and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

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.

Almond, D. and J. Doyle (2011): "After midnight: A regression discontinuity design in length of postpartum hospital stays," *American Economic Journal: Economic Policy*, 3, 1–34.

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.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Angrist, J. D. and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics*, 114, 533–575.

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

——— (2020a): "Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness," *Working Paper*.

——— (2020b): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*.

Armstrong, T. B., M. Kolesár, and S. Kwon (2020): "Bias-Aware Inference in Regularized Regression Models," *Working Paper*.

Bertanha, M. and M. J. Moreira (2018): "Impossible Inference in Econometrics: Theory and Applications," *Journal of Econometrics*.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, 113, 767–779.

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

Card, D. and L. Giuliano (2016): "Can tracking raise the test scores of high-ability minority students?" *American Economic Review*, 106, 2783–2816.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on causal effects in a generalized regression kink design," *Econometrica*, 83, 2453–2483.

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.

Clark, D. and P. Martorell (2014): "The signaling value of a high school diploma," *Journal of Political Economy*, 122, 282–318.

Coviello, D., A. Guglielmo, and G. Spagnolo (2018): "The effect of discretion on procurement performance," *Management Science*, 64, 715–738.

Dahl, G. B., K. V. Løken, and M. Mogstad (2014): "Peer effects in program participation," *American Economic Review*, 104, 2049–74.

Dong, Y. (2018): "Alternative assumptions to identify LATE in fuzzy regression discontinuity designs," *Oxford Bulletin of Economics and Statistics*, 80, 1020–1027.

Donoho, D. L. (1994): "Statistical estimation and optimal recovery," *Annals of Statistics*, 22, 238–270.

Dube, A., L. Giuliano, and J. Leonard (2019): "Fairness and frictions: The impact of unequal raises on quit behavior," *American Economic Review*, 109, 620–63.

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.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013): "Long-term effects of class size," *The Quarterly Journal of Economics*, 128, 249–285.

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.

Hinnerich, B. T. and P. Pettersson-Lidbom (2014): "Democracy, redistribution, and political participation: Evidence from Sweden 1919–1938," *Econometrica*, 82, 961–993.

Huang, X. and Z. Zhan (2020): "Does health behavior change after diagnosis? Evidence from a reliable fuzzy regression discontinuity approach," *Working Paper*.

Ignatiadis, N. and S. Wager (2020): "Bias-aware confidence intervals for empirical Bayes analysis," *Working Paper*.

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.

Jepsen, C., P. Mueser, and K. Troske (2016): "Labor market returns to the GED using regression discontinuity analysis," *Journal of Political Economy*, 124, 621–649.

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.

Le Barbanchon, T., R. Rathelot, and A. Roulet (2019): "Unemployment insurance and reservation wages: Evidence from administrative data," *Journal of Public Economics*, 171, 1–17.

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.

Low, M. (1997): "On nonparametric confidence intervals," *Annals of Statistics*, 25, 2547–2554.

Malenko, N. and Y. Shen (2016): "The role of proxy advisory firms: Evidence from a regression-discontinuity design," *The Review of Financial Studies*, 29, 3394–3427.

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

——— (2008): "Corrigendum: Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *Internet-Only Corrigendum; American Economic Review*.

Sacks, J. and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *Annals of Statistics*, 6, 1122–1137.

Schennach, S. M. (2020): "A bias bound approach to nonparametric inference," *Review of Economic Studies*, to appear.

Scott-Clayton, J. and B. Zafar (2019): "Financial aid, debt management, and socioeconomic outcomes: Post-college effects of merit-based aid," *Journal of Public Economics*, 170, 68–82.

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

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

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.