Inference in Regression Discontinuity Designs with a Discrete Running Variable*

Michal Kolesár† Christoph Rothe‡

June 21, 2016

Abstract

We consider inference in regression discontinuity designs when the running variable only takes a moderate number of distinct values. In particular, we study the common practice of using confidence intervals (CIs) based on standard errors that are clustered by the running variable. We derive theoretical results and present simulation and empirical evidence showing that these CIs have poor coverage properties and therefore recommend that they not be used in practice. We also suggest alternative CIs with guaranteed coverage properties under easily interpretable restrictions on the conditional expectation function.

*We thank Tim Armstrong, Joshua Angrist, Guido Imbens, Philip Oreopoulos, Miguel Urquiola, and seminar participants at Columbia University for helpful comments.

†Woodrow Wilson School and Department of Economics, Princeton University. Electronic correspondence: mkolesar@princeton.edu

‡Department of Economics, Columbia University. Electronic correspondence: cr2690@columbia.edu

1. INTRODUCTION

The regression discontinuity design (RDD) is a popular quasi-experimental empirical strategy that exploits fixed cutoff rules present in many institutional settings to estimate treatment effects. In its most basic version, the sharp RDD, observational units receive treatment if and only if an observed running variable falls above a known threshold value. For example, students may be awarded a scholarship if their test score is above some pre-specified level. If unobserved confounders vary smoothly around the assignment threshold, the jump in the conditional expectation function of the outcome given the running variable at the threshold identifies the average treatment effect (ATE) for units at the margin for being treated (Hahn et al., 2001).

In order to obtain a reliable estimate of the ATE, the recent theoretical literature on RDDs has emphasized the importance of flexible specifications of the functional form of the conditional expectation function. A standard approach in empirical practice is to use local polynomial regression. In its simplest form, this method approximates the conditional expectation function within a window around the threshold by a low-order polynomial on either side of the threshold, which can be fitted using ordinary least squares (OLS) after discarding observations outside of the window. If the window is sufficiently small for the bias from the polynomial approximation to be negligible relative to the standard error of the estimator,[1] constructing a valid confidence interval (CI) for the ATE is straightforward. For instance, one can use the "usual" CI based on the Eicker-Huber-White (EHW) heteroskedasticity-robust standard error.

In an important paper, Lee and Card (2008, LC from hereon) point out that this approach to constructing CIs may not be feasible if the running variable only takes on a moderate number of distinct values on at least one side of the threshold. In particular, if the gaps between the values closest to the threshold are sufficiently large, researchers may be forced to choose a window around the threshold that is too large for the bias from the polynomial approximation to be negligible in order to ensure a reasonably low level of sampling noise. This means that the EHW CI undercovers the ATE, as it is not adequately centered. This concern applies to many empirical settings: a wide range of treatments are triggered when quantities that inherently take on a limited number of values, like the test score of a student, the enrollment number of a school, the number of employees of a company, or the year of birth of an individual, exceed some threshold.[2]

[1] This approach is known as "undersmoothing" in the nonparametric regression literature. See Calonico et al. (2014) for an alternative approach.

[2] This setting is conceptually different from cases in which the running variable is continuous in principle,

To address this problem, LC suggest using a CI based on a "cluster-robust" standard error (e.g. Liang and Zeger, 1986) that treats observational units with the same realization of the running variable as members of distinct groups. We refer to this type of standard error and the corresponding CI as being *clustered by the running variable* (CRV) in the following. LC's approach has been widely adopted in applied economics, It is routinely used in empirical studies[3], and recommended in survey papers (e.g. Lee and Lemieux, 2010) and government agency guidelines for carrying out RDD studies (e.g. Schochet et al., 2010).

In this paper, we examine the properties of CRV CIs. LC motivate these CIs by modeling the error in the (local) polynomial approximation to the conditional expectation function as random, with mean zero conditional on the running variable, and independent across "clusters". However, as we discuss in detail in Section 3.1, this heuristic is not compatible with the usual sampling framework that treats the data as generated by independent sampling from some fixed distribution. Indeed, in the usual framework, the approximation error is fixed across repeated samples conditional on the running variable.

Since the rationale for using CRV standard errors relies on non-standard modeling assumptions, we derive expressions for the asymptotic coverage of CRV CIs in the usual RDD setup and show that, in general, their coverage properties are poor. There are two effects at play. First, CRV standard errors tend to increase relative to EHW standard errors with the degree of misspecification. Second, coverage of CRV CIs decreases with the number of support points of the running variable, or the number of "clusters". The latter effect is due to the usual downward bias of the cluster-robust variance estimator in cases with a small number of clusters; see Cameron and Miller (2014) for a recent survey.

When the number of support points of the running variable (within the window around the threshold that is used to fit the polynomial approximation) is large, or the degree of misspecification is large, the first effect dominates, and the asymptotic coverage is better than that of EHW CIs. This formalizes the intuition that "the use of clustered standard errors will generally lead to wider confidence intervals" (LC, p. 656). However, we show, using both theoretical arguments and simulation evidence, that the coverage of CRV CIs may still be arbitrarily far below their nominal level, as the improvement in coverage over EHW CIs may not be sufficient when the coverage of EHW CIs is very poor to begin with.

but only a discretized or rounded version of its realization is recorded in the data. See Dong (2015) for an analysis of RDDs with this type of measurement error.

[3]Recent papers published in leading economics journals that conduct inference in RDDs by clustering standard errors by the running variable include Oreopoulos (2006), Card et al. (2008), Urquiola and Verhoogen (2009), Martorell and McFarlin (2011), Chetty et al. (2013) and Clark and Royer (2013), among many others.

Moreover, when the running variable has many support points, a more effective way to control misspecification bias would be to work with a narrower window around the threshold.[4]

When the number of support points of the running variable is small and the degree of misspecification is mild, the second effect dominates, and coverage of CRV CIs is typically *worse* than that of EHW CIs. To show that the amount of under-coverage can be substantial, we re-analyze data from Oreopoulos (2006), and find that EHW standard errors are larger than CRV standard errors by several orders of magnitude in this context, with the exact amount depending on the precise specification. The exercise shows that clustering by the running variable can lead to incorrect claims about the statistical significance of the estimated effect, even if we assume that the CIs are correctly centered.[5]

These results caution against clustering standard errors by the running variable in empirical applications, in spite of its great popularity.[6] We therefore propose two alternative methods for constructing CIs for the ATE in discrete RDDs that have guaranteed coverage properties under interpretable restrictions on the conditional expectation function. The first method makes the assumption that the magnitude of the approximation bias is no larger at the left limit of the threshold than at any point in the support of the running variable below the threshold, and similarly for the right limit. The second method relies on the assumption recently considered in Armstrong and Kolesár (2016b) that the second derivative of the conditional expectation function is bounded by a constant. Both CIs are "honest" in the sense of Li (1989) in that they achieve asymptotically correct coverage for all possible model parameters, that is, they are valid uniformly in the value of the conditional expectation function.

The rest of the paper is organized as follows. Section 2 reviews the sharp regression discontinuity model and issues that arise then the running variable is discrete. Section 3 derives theoretical coverage properties of CRV CIs. Section 4 studies their properties in

[4]LC, in their Section 4, caution that in some circumstances the CRV standard error may understate the variability of the point estimate, although their reasoning is very different from ours. They suggest a further modification of the standard error formula. However, this modification can be shown to suffer from similar coverage issues as CRV CIs. Since this modification does not seem to be used much in the empirical literature we focus on LC's main proposal in this paper.

[5]Using alternative methods for constructing CIs suggested recently in the literature on inference with a small number of clusters (e.g. Cameron et al., 2008; Canay et al., 2015; Imbens and Kolesár, 2016) does not lead to reliable inference in this case: as we argued above, even if standard errors were free from bias, the improvement in coverage over EHW CIs may not be sufficient.

[6]Of course, clustering standard errors at some level other than the running variable may be appropriate in applications where the data are not generated by independent sampling. Suppose, for example, that the observational units are students, and that the data are obtained by first collecting a sample of schools and then sampling students from within those schools. In this case, clustering on schools would be justified.

a series of simulation experiments and in an empirical application. Section 5 discusses alternative methods of inference that lead to honest CIs. Section 6 concludes. Proofs for Section 3 are given in Appendix A. Proofs for Section 5 are given in Appendix B.

2. SHARP REGRESSION DISCONTINUITY DESIGNS

In this section, we first introduce the sharp RDD and review standard methods for inference. We then discuss the implications of discreteness of the running variable.

2.1. **Basic model and inference**

The aim of a RDD is to infer the causal effect of a binary treatment on some outcome of interest. We observe a random sample of N units, indexed by $i = 1, \ldots, N$, from some large population. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcome with and without receiving treatment, respectively, and let $D_i \in \{0, 1\}$ be an indicator variable for the event that the unit receives treatment. The actual outcome is given by $Y_i = Y_i(D_i)$. While the issues that we study in this paper also arise in fuzzy and kink RDDs, for ease of exposition we focus on the sharp case, in which a unit is treated if and only if a running variable X_i crosses a known threshold, which we normalize to zero without loss of generality:

$$D_i = \mathbb{I}\{X_i \geq 0\}.$$

The parameter of interest is the average treatment effect (ATE) at the threshold,

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = 0).$$

Let $\mu_d(X_i) = \mathbb{E}(Y_i(d) \mid X_i)$, $d \in \{0, 1\}$ denote the conditional expectation functions for the potential outcomes, and let $\mu(X_i) = \mathbb{E}[Y_i \mid X_i] = \mu_1(X_i)\mathbb{I}\{X_i \geq 0\} + \mu_0(X_i)\mathbb{I}\{X_i < 0\}$ denote the conditional expectation of the observed outcome given the running variable. If $\mu_0(x)$ and $\mu_1(x)$ are continuous at the threshold, then the ATE is equal to the discontinuity in $\mu(x)$ at the threshold:

$$\tau = \lim_{x \downarrow 0} \mu(x) - \lim_{x \uparrow 0} \mu(x).$$

To estimate τ, researchers therefore have to estimate the right and left limits of the conditional expectation function $\mu(x)$ at the threshold. The recent literature on RDDs has emphasized the importance of using flexible specifications for $\mu(x)$, with local polynomial regression having become the standard choice for estimation. In its simplest form, this estimation strategy involves fixing a polynomial order $p \in \mathbb{N}$, with $p = 1$ and $p = 2$ being the most common choices, and a bandwidth $h > 0$, and then estimating the ATE by $\widehat{\tau}$, which is

given by

$$\widehat{\tau} = e_1'\widehat{\theta}, \qquad \widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - m(X_i)'\theta)^2 \mathbb{I}\left\{|X_i| \leq h\right\}, \tag{2.1}$$

where $e_1 = (1, 0, \ldots, 0)'$ denotes the first unit vector and

$$m(x) = (\mathbb{I}\left\{x \geq 0\right\}, 1, x, \ldots, x^p, \mathbb{I}\left\{x \geq 0\right\} x, \ldots, \mathbb{I}\left\{x \geq 0\right\} x^p)'.$$

The approach thus amounts to discarding observations that are further than h away from the threshold, and fitting a pth order polynomial approximation on either side of the threshold by ordinary least squares (OLS).[7,8] Basic properties of OLS estimators imply that for any fixed p and h, in finite samples $\widehat{\tau}$ is approximately unbiased for the pseudo-parameter τ_h, which is given by

$$\tau_h = e_1'\theta_h, \qquad \theta_h = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}((Y_i - m(X_i)'\theta)^2 \mathbb{I}\left\{|X_i| \leq h\right\}).$$

The magnitude of the bias $\tau_h - \tau$ is determined by how well $\mu(x)$ is approximated by a pth order polynomial over $(-h, h)$. If $\mu(x)$ is a smooth function, smaller values of h generally lead to a decrease in the bias, at the cost of an increase in the variance of $\widehat{\tau}$.

For any fixed p and h, $\widehat{\theta}$ is simply an OLS estimator based on the $N_h = \sum_{i=1}^{N} \mathbb{I}\left\{|X_i| \leq h\right\}$ observations within an h-window around the threshold. Therefore, a natural candidate for a CI for τ with nomial level $(1-\alpha)$ is

$$C_{\text{EHW}}^{1-\alpha} = (\widehat{\tau} \pm z_\alpha \times \widehat{\sigma}_{\text{EHW}}/\sqrt{N_h}),$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, and $\widehat{\sigma}^2_{\text{EHW}}$ is the top-left element of the usual Eicker-Huber-White (EHW) heteroskedasticity-robust estimator of the asymptotic variance of $\widehat{\theta}$. That is,

$$\widehat{\sigma}^2_{\text{EHW}} = e_1'\widehat{Q}^{-1}\widehat{\Omega}_{\text{EHW}}\widehat{Q}^{-1}e_1, \qquad \widehat{\Omega}_{\text{EHW}} = \frac{1}{N_h}\sum_{i=1}^{N} \widehat{u}_i M_i M_i', \qquad \widehat{Q} = \frac{1}{N_h}\sum_{i=1}^{N} M_i M_i',$$

[7] We use a uniform kernel in equation (2.1) to simplify the exposition. Analogous results could be obtained if the indicator function $\mathbb{I}\left\{|X_i| \leq h\right\}$ was replaced with another standard kernel function. Our exposition follows Imbens and Lemieux (2008) in this regard.

[8] Our notation also formally covers the global polynomial approach to estimating τ by choosing $h = \infty$ and a rather large value of p. While this estimation approach is used in a number of empirical studies, theoretical results suggest that it typically performs poorly relative to local linear or local quadratic regression (Gelman and Imbens, 2014). This is because the method often implicitly assigns very large weights to observations far away from the threshold.

with $M_i = m(X_i)\mathbb{I}\{|X_i| \leq h\}$ and $\widehat{u}_i = Y_i - m(X_i)'\widehat{\theta}$ the ith regression residual.

Under standard regularity conditions, for any fixed p, the quantity $\sqrt{N_h}(\widehat{\tau} - \tau_h)/\widehat{\sigma}_{\text{EHW}}$ is normally distributed in large samples irrespective of the magnitude of the bias $\tau_h - \tau$. Therefore $C^{1-\alpha}_{\text{EHW}}$ is an asymptotically valid $(1-\alpha)$ CI for the pseudo-parameter τ_h. If the bias $\tau_h - \tau$ is asymptotically negligible relative to the standard error $\widehat{\sigma}_{\text{EHW}}/\sqrt{N_h}$, then $C^{1-\alpha}_{\text{EHW}}$ is also an asymptotically valid $(1-\alpha)$ CI for τ, the parameter of interest. More formally, it holds that $P(\tau \in C^{1-\alpha}_{\text{EHW}}) \to 1-\alpha$ as $N \to \infty$ if

$$\sqrt{N_h}(\widehat{\tau} - \tau_h)/\widehat{\sigma}_{\text{EHW}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{and} \quad \frac{\tau_h - \tau}{\widehat{\sigma}_{\text{EHW}}/\sqrt{N_h}} = o_P(1). \tag{2.2}$$

The condition (2.2) is satisfied for example if the sampling distribution of the running variable X_i is well-approximated by a continuous distribution with positive, continuous density in the neighborhood of zero, $\mu(x)$ is at least $p+1$ times continuously differentiable, and the bandwidth h "undersmooths" relative to the mean squared error optimal bandwidth such that, as $N \to \infty$, $h^{2p+2}N \to 0$ and $Nh \to \infty$ (see, for example, Hahn et al., 2001, Theorem 4). Such arguments justify the use of $C^{1-\alpha}_{\text{EHW}}$ as a CI for τ in practice when X_i has rich support and h is chosen sufficiently small.

2.2. **Discrete running variables**

Now suppose that the running variable is discrete and only takes on G^+ and G^- distinct values above and below the threshold, respectively. The support of X_i can then be written as

$$\mathcal{X} = \{x_1, \ldots, x_{G^-}, x_{G^-+1}, \ldots, x_G\}$$

for constants $x_1 < \cdots < x_{G^-} < 0 \leq x_{G^-+1} < \cdots < x_G$, where $G = G^+ + G^-$. In many respects it is not necessary to distinguish sharply between the case of a discrete and a continuously distributed running variable for the purpose of estimation and inference; and the basic arguments laid out in the previous subsection remain valid in principle when X_i is discrete.[9] However, as aptly pointed out by LC, if either G^+ or G^- are small, and the gaps between the support points closest to the threshold are sufficiently wide, justifying the small-bias condition in (2.2) is problematic. This is because in such a case the researcher might be forced to choose a bandwidth that she considers to be "too large" in terms of bias in order to guarantee that a reasonable number of data points fall within the corresponding

[9]While most results in the recent literature on local polynomial regression are formulated for settings with continuous covariates, an advantage of some early methods for inference, such as that of Sacks and Ylvisaker (1978), is that is is not necessary to distinguish between the discrete and the continuous case.

window on either side of the threshold. This in turn affects inference, as it implies that $C^{1-\alpha}_{\text{EHW}}$ may not have good coverage properties due to improper centering.

LC therefore suggest an alternative method for conducting inference on τ in RDDs with a discrete running variable, which has subsequently been widely adopted in the empirical literature. Their proposal is to replace the variance estimator $\widehat{\sigma}^2_{\text{EHW}}$ with the estimator

$$\widehat{\sigma}^2_{\text{CRV}} = e_1'\widehat{Q}^{-1}\widehat{\Omega}_{\text{CRV}}\widehat{Q}^{-1}e_1, \qquad \widehat{\Omega}_{\text{CRV}} = \frac{1}{N_h}\sum_{g=1}^{G}\mathbf{M}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{M}_g,$$

where $\mathbf{M}_g$ is the matrix that stacks $\sum_{i=1}^{N}\mathbb{I}\{X_i = x_g\}$ copies of $m_g = m(x_g)\mathbb{I}\{|x_g| \leq h\}$, and $\widehat{\mathbf{u}}_g$ the vector that stacks those of the regression residuals $\{\widehat{u}_i\}_{i=1}^{N}$ for which $X_i = x_g$. The estimator $\widehat{\sigma}^2_{\text{CRV}}$ is the top-left element of the usual "cluster-robust" estimator of the asymptotic variance of $\widehat{\theta}$ that "clusters by the running variable", i.e. it treats units with the same realization of the running variable as belonging to the same cluster (Liang and Zeger, 1986). The resulting CI for τ with nominal level $1-\alpha$ is given by

$$C^{1-\alpha}_{\text{CRV}} = (\widehat{\tau} \pm z_\alpha \times \widehat{\sigma}_{\text{CRV}}/\sqrt{N_h}).$$

Here the subscript CRV stands for "clustered by the running variable"; we refer to $C^{1-\alpha}_{\text{CRV}}$ as the CRV CI in the following, and to $\widehat{\sigma}_{\text{CRV}}/\sqrt{N_h}$ as the CRV standard error.

3. PROPERTIES OF CRV CONFIDENCE INTERVALS

In this section, we review LC's motivation for clustering standard errors by the running variable, and derive formal expressions for the asymptotic coverage of CRV CIs.

3.1. **Motivation for clustering by the running variable**

As pointed out in the introduction, the CI proposed by LC has been used in numerous empirical studies in various fields of economics. The rationale they provide for this CI is as follows. Suppose that h is held fixed as the sample size increases,[10] denote the misspecification bias of the (local) polynomial approximation to the true conditional expectation function by $\delta(x) = \mu(x) - m(x)'\theta_h$, put $\delta_i = \delta(X_i)$, let $\varepsilon_i = Y_i - \mu(X_i)$ be the deviation of Y_i from its true conditional expectation given X_i, and write the outcome of the ith unit as

$$Y_i = m(X_i)'\theta_h + \eta_i, \qquad \eta_i = \delta_i + \varepsilon_i. \tag{3.1}$$

[10]Formally, there is no bandwidth in the setup considered by LC. However, their setup is equivalent to local polynomial regression when a fixed bandwidth and a uniform kernel function are used.

LC then treat the misspecification error δ_i as a random effect. That is, they assume that conditional on X_i the term δ_i is random with conditional mean zero, and that it is independent for units with different realizations of the running variable, but potentially correlated for units with the same value of the running variable. Under this assumption, equation (3.1) is a correctly specified regression model in which the term $\eta_i = \delta_i + \varepsilon_i$ exhibits within-group correlation at the level of the running variable. That is, η_i and $\eta_{i'}$ are allowed to be correlated if $X_i = X_{i'}$, but the correlation is assumed to be zero if $X_i \neq X_{i'}$. Due to this group structure, the variance estimator $\widehat{\sigma}^2_{\text{CRV}}$ is appropriate under LC's assumption.

The rationale put forward by LC is unusual, however, in that their setup is not compatible with the standard assumption that the data $\{Y_i, X_i\}_{i=1}^N$ are drawn independently from the distribution of some random vector (Y, X). Under random sampling, $\delta_i = \delta(X_i)$ is not random conditional on X_i, as the function $\delta(x)$ depends only on the distribution of the vector (Y, X), which does not change across repeated samples. Furthermore, if $\delta(x)$ was random, then the conditional expectation function $\mu(x) = m(x)'\theta_h + \delta(x)$ would have to be random as well; and since τ is a functional of $\mu(x)$, this would imply that the ATE is not a population quantity but a random variable whose value changes across repeated samples. Finally, while by construction observations in the same cluster (i.e. units with the same value of the running variable) have the same misspecification error $\delta(X_i)$, if the true regression function $\mu(x)$ is smooth, observations with similar values of the running variable also have a similar value of the misspecification error. This violates LC's assumption that the term η_i is independent across the "clusters".

3.2. **Asymptotic properties**

Since the rationale for using CRV standard errors relies on non-standard modeling assumptions, it is interesting to investigate the properties of $C^{1-\alpha}_{\text{CRV}}$ in the usual RDD setup as outlined in Section 2.1. To understand whether $C^{1-\alpha}_{\text{CRV}}$ is indeed robust to misspecification, first note that, as outlined above, it is clear that $C^{1-\alpha}_{\text{EHW}}$ is an appropriate CI for the pseudo-parameter τ_h irrespective of how rich the support of the running variable is. If the bias $\tau - \tau_h$ is non-negligible relative to the EHW standard error $\widehat{\sigma}_{\text{EHW}}/\sqrt{N_h}$, then $C^{1-\alpha}_{\text{EHW}}$ undercovers as a CI for τ as it is not correctly centered. Since $C^{1-\alpha}_{\text{CRV}}$ is centered at the same point estimate as $C^{1-\alpha}_{\text{EHW}}$, any argument for its validity needs to establish that the CRV standard error $\widehat{\sigma}_{\text{CRV}}/\sqrt{N_h}$ enlarges the EHW standard error by an appropriate amount so that $C^{1-\alpha}_{\text{CRV}}$ achieves proper coverage of τ even if the bias $\tau - \tau_h$ is non-negligible.

In the following subsections, we show that this is *not* the case, and that $C^{1-\alpha}_{\text{CRV}}$ (i) undercovers the ATE under correct specification and (ii) can either over- or undercover the ATE

under misspecification by essentially arbitrary amounts. We derive large sample approximations to the stochastic properties of $\widehat{\sigma}^2_{\text{CRV}}$, which then directly translate into statements about the coverage properties of $C^{1-\alpha}_{\text{CRV}}$. Our results show that there are two effects at play in general. First, a small number of support points of the running variable on either side of the threshold tends to bias the CRV standard error downward, which is a consequence of the usual downward bias of the cluster-robust variance estimator in cases with a small number of clusters; see Cameron and Miller (2014) for a recent survey. Second, the CRV standard errors tend to increase relative to the EHW standard errors with the degree of misspecification. Thus, with a small number of support points and mild degree of misspecification, the CRV standard errors are typically *smaller* than the EHW standard errors. The CRV CI may therefore amplify, rather than solve, the problems for inference caused by misspecification bias. If the number of support points or the degree of misspecification are large the CRV standard errors are indeed larger relative to the EHW standard errors, and therefore lead to an improvement in coverage. However, the coverage of the resulting CI is still not guaranteed to be close to the nominal coverage, as the enlargement of the EHW standard errors may not be sufficient.

To allow us to better capture the finite-sample properties of $\widehat{\sigma}^2_{\text{CRV}}$ and $C^{1-\alpha}_{\text{CRV}}$, we consider different types of triangular array asymptotics in which the distribution of (Y, X) may potentially change with the sample size. To simplify the exposition, we leave the dependence of quantities such as $\mu(x)$, $\delta(x)$, G^+, G^-, or $\mathcal{X}$ on the sample size implicit in our notation. We denote the number of support points of the running variable that are at most h away from the threshold by G_h, and the number of those points that are above and below the threshold by G^+_h and G^-_h, respectively, and let $\mathcal{G}_h = \{g \colon |x_g| \leq h, x_g \in \mathcal{X}\}$ be the set of indices corresponding to support points within the estimation window. We let $\pi_g = P(X_i = x_g)$, $\pi = P(|X_i| \leq h)$, and write $\sigma^2_g = \mathbb{V}(Y_i \mid X_i = x_g)$ for the conditional variance of Y_i. Finally, we denote the population version of the matrix $\widehat{Q}$ by $Q = \mathbb{E}[M_i M_i' \mid |X_i| \leq h]$, and write $Q_g = (\pi_g/\pi) m_g m_g'$ and $\Omega = \mathbb{E}[u_i^2 M_i M_i \mid |X_i| \leq h]$. Throughout the section, we also maintain the regularity assumptions that $\sup_{N \in \mathbb{N}} \max_{g \in \mathcal{G}_h} \mathbb{E}[\varepsilon_i^4 | X_i = x_g] < \infty$, and $\sup_{N \in \mathbb{N}} \max_{g \in \mathcal{G}_h} \delta(x_g) < \infty$.

For simplicity, we assume that the bandwidth h is held fixed. In Appendix A, we present formal results regarding the properties of $\widehat{\sigma}^2_{\text{CRV}}$ that cover a more general framework that allows the bandwidth to change with the sample size. The propositions in the remainder of this section follow from those general results.

Fixed data generating process

For our first result, we consider the case that the joint distribution of (Y, X) does not change with the sample size N, and that the pth order polynomial model for $\mu(x)$ is misspecified over $x \in (-h, h)$, so that $\delta(x_g) \neq 0$ for at least one $g \in \mathcal{G}_h$. We also define

$$r = \sum_{g=1}^{G} \frac{\pi_g}{\pi} \delta(x_g)^2 e_1' Q^{-1} Q_g Q^{-1} e_1,$$

and note that r is a strictly positive constant in this case.

Proposition 1. *If the conditions above hold, then*

$$\frac{\widehat{\sigma}^2_{\text{CRV}}}{N_h} = r + o_P(1) \quad \text{and} \quad P(\tau \in C^{1-\alpha}_{\text{CRV}}) = \mathbb{I}\left\{|\tau_h - \tau| \leq z_\alpha \sqrt{r}\right\} + o(1).$$

The proposition shows that under its conditions the CRV standard error does not converge to zero as the sample size increases. This means that $C^{1-\alpha}_{\text{CRV}}$ does not shrink towards a singleton asymptotically, but instead is equal to some interval with positive length in large samples. Consequently, its asymptotic coverage probability is either zero or one, depending on the magnitude of the bias of $\widehat{\tau}$.

Local misspecification

The asymptotic approximation implied by Proposition 1 might not be very useful in finite samples if the degree to which $\mu(x)$ differs from a pth order polynomial is moderate or small relative to the overall sampling uncertainty. In order to obtain a better approximation, consider the case that the function $\mu(x)$ is within a $N^{-1/2}$ neighborhood of a pth order polynomial specification over $x \in (-h, h)$. This implies that

$$\sqrt{N_h}(\tau_h - \tau) \to b \quad \text{and} \quad \sqrt{N\pi_g}\delta(x_g) \to d_g$$

for all $g \in \mathcal{G}_h$ and some constants b and $d_1, \ldots d_G$. In this case, the weighted misspecification errors $\sqrt{\pi_g}\delta(x_g)$ are of the same order of magnitude as the standard deviation of $\widehat{\tau}$. Now let $B_1, \ldots, B_G$ be a collection of independent random variables with $B_g \sim \mathcal{N}(0, \pi_g \sigma_g^2/\pi)$, and note that

$$\sqrt{N_h}(\widehat{\tau} - \tau_h) \stackrel{d}{=} e_1' S + o_P(1), \quad \text{with} \quad S = Q^{-1} \sum_{g=1}^{G} m_g B_g.$$

For all $g \in \mathcal{G}$, we also define the quantity

$$W_g = e_1' Q^{-1} m_g \left(B_g - \frac{\pi_g}{\pi} m_g' Q^{-1} \sum_{j=1}^{G} m_j B_j + \sqrt{\frac{\pi_g}{\pi}} d(x_g) \right).$$

With this notation, we obtain the following result:

Proposition 2. *If the conditions above hold, then*

$$\widehat{\sigma}^2_{\text{CRV}} \stackrel{d}{=} (1 + o_P(1)) \sum_{g=1}^{G} W_g^2 \quad \text{and} \quad P(\tau \in C^{1-\alpha}_{\text{CRV}}) = P\left(\frac{|e_1' S + b|}{\sqrt{\sum_{g=1}^{G} W_g^2}} \leq z_\alpha \right) + o(1).$$

The proposition states that under its conditions the distribution of $\widehat{\sigma}^2_{\text{CRV}}$ converges to a non-degenerate limit as the sample size increases. To better understand the location of this limiting distribution, let $\sigma_\tau^2 = e_1' Q^{-1} \Omega Q^{-1} e_1$ denote the asymptotic variance of $\sqrt{N_h}(\widehat{\tau} - \tau_h)$. Treating moments of $\sum_{g=1}^{G} W_g^2$ as an approximation to the moments of $\widehat{\sigma}^2_{\text{CRV}}$, the asymptotic bias of $\widehat{\sigma}^2_{\text{CRV}}$ as an estimate of σ_τ^2 is given by

$$\sum_{g=1}^{G} \mathbb{E}(W_g^2) - \sigma_\tau^2 = \sum_{g=1}^{G} d_g^2 e_1' \lambda_g + \sum_{g=1}^{G} \left(\lambda_g' \Omega \lambda_g - 2\sigma_g^2 \lambda_g' Q \lambda_g \right), \tag{3.2}$$

where $\lambda_g = Q^{-1} Q_g Q^{-1} e_1$. The first term on the right-hand side of (3.2) is positive, and increasing in the degree of misspecification as measured by the terms $d_1, \ldots d_G$. The second term on the right-hand side of (3.2) does not depend on the degree of misspecification, and its sign is difficult to determine in general. Under homoskedasticity, that is $\sigma_g^2 = \sigma^2$ for all g, it holds that

$$\sum_{g=1}^{G} \left(\lambda_g' \Omega \lambda_g - 2\sigma_g^2 \lambda_g' Q \lambda_g \right) = \sigma^2 \sum_{g=1}^{G} \lambda_g' Q \lambda_g < 0.$$

We thus expect the second term on the right-hand side of (3.2) to be negative for small and moderate levels of heteroskedasticity. In consequence, the limiting distribution of $\widehat{\sigma}^2_{\text{CRV}}$ is centered around arbitrarily large values if the terms $d_1, \ldots d_G$ that measure the degree of misspecification are sufficiently large in absolute value; and it is centered below the correct asymptotic variance σ_τ^2 under correct specification of $\mu(x)$.

The CRV CI is thus generally *narrower* than the EHW CI under moderate misspecification of the conditional expectation function. Since the values of $d_1, \ldots d_G$ do not restrict the value of b, $C^{1-\alpha}_{\text{CRV}}$ can also have asymptotic coverage probability anywhere between zero or one, depending the magnitude of the bias $\tau_h - \tau$ and the extent to which our model for $\mu(x)$ is misspecified.

Local misspecification with increasing number of support points

One can show that the second term on the right-hand side of (3.2) is decreasing in the number of support points of the running variable within the estimation window. This component of the bias of $\widehat{\sigma}^2_{\text{CRV}}$ is thus analogous to the well-known distortion of clustered variance estimates in settings in which the data have an actual group structure and the number of groups is small. Our last result shows that this component of the bias of $\widehat{\sigma}^2_{\text{CRV}}$ vanishes when the number of support points of the running variable that fall within the interval $(-h, h)$ increases with the sample size. Continuing the use of notation introduced in the previous subsection, suppose that the function $\mu(x)$ is within a $N^{-1/2}$ neighborhood of a pth order polynomial specification over $x \in (-h, h)$, and that as $N \to \infty$ the number of support points increases in such a way that $G_h \to \infty$, $N\pi \to \infty$, $\max_{g \in \mathcal{G}_h} \pi_g / \pi \to 0$ and $\max_{g \in \mathcal{G}_h} |d_g| = O(1)$. We also define the quantity

$$t = \sigma^2_\tau + \sum_{g=1}^{G} d_g^2 e_1' \lambda_g.$$

With this notation, we obtain the following result.

Proposition 3. *If the conditions above hold, then*

$$\widehat{\sigma}^2_{\text{CRV}} = t + o_P(1) \quad \text{and} \quad P(\tau \in C^{1-\alpha}_{\text{CRV}}) = P\left(\frac{|e_1' S + b|}{\sqrt{t}} \leq z_\alpha \right) + o(1).$$

With a large number of support points, the variance estimator used to construct the CRV CI is thus an upward-biased estimate of the asymptotic variance of $\widehat{\tau}$ if $\mu(x)$ is misspecified, and inference on τ_h therefore becomes conservative. Consequently, there is a class $\mathcal{M}_{\text{CRV}}$ of functions that is strictly larger than the class of all pth order linear functions over $(-h, h)$, and such that

$$\lim_{N \to \infty} P_\mu(\tau \in C^{1-\alpha}_{\text{CRV}}) \geq 1 - \alpha \text{ if } \mu \in \mathcal{M}_{\text{CRV}}.$$

This result captures the argument in LC that clustering at the level of the running variable provide robust inference against deviations from pth order polynomial specification, in the sense that they achieve correct asymptotic coverage if $\mu \in \mathcal{M}_{\text{CRV}}$. However, this class $\mathcal{M}_{\text{CRV}}$ is difficult to characterize as it depends on the entire distribution of the running variable in non-obvious ways. As we show in the next section using simulation evidence, even in settings with many support points clustering may therefore not provide a meaningful degree of robustness with respect to the presence of misspecification. Moreover, if a researcher is worried about bias when the number of support points within the estimation window

is "large", a simpler way to address the problem would be to choose a narrower bandwidth. Alternatively, one can use one of the honest inference procedures that we outline in Section 5, which guarantee proper coverage in a precise sense.

4. NUMERICAL EVIDENCE

In this section, we first present the results of a Monte Carlo study that illustrates practical relevance of the theoretical findings presented above. We then re-analyze data from Oreopoulos (2006), who studies the effect of a change in the minimum school-leaving age in the United Kingdom on educational attainment and earnings.

4.1. **Simulations**

To study the accuracy of our asymptotic approximations in finite samples, we conduct a series of simulation experiments. We consider several data generating processes (DGPs) with different conditional expectation functions and different numbers of support points, and also several sample sizes. Each DGP is such that the support of the running variable is the union of an equally spaced grid of G^- points on $[-1, 0)$ and an equally spaced grid of G^+ points on $(0, 1]$, and we consider values $(G^-, G^+) \in \{5, 25, 50\}^2$. The distribution of X_i is always such that probability mass $1/2$ is spread equally across the support points on either side of the threshold.[11] The outcome variable is generated as $Y_i = \mu(X_i) + \varepsilon_i$, where ε_i and X_i are independent, $\varepsilon_i \sim \mathcal{N}(0, 1)$, and

$$\mu(x) = x + \lambda_1 \cdot \sin(\pi \cdot x) + \lambda_2 \cdot \cos(\pi \cdot x).$$

Since $\mu(x)$ is continuous at $x = 0$ for every (λ_1, λ_2), we have $\tau = 0$ in our all our DGPs. We consider $(\lambda_1, \lambda_2) \in \{(0, 0), (0.05, 0), (0, 0.05)\}$ and the sample sizes $N \in \{10^3, \ldots, 10^6\}$, estimate the ATE by fitting a linear specification on each side of the threshold (which corresponds to choosing $p = 1$ and $h = 1$), and set the number of replications to 100,000. We plot the versions of $\mu(x)$ that we consider together with the corresponding linear fit in Figure 1 for the case that $G^- = G^+ = 10$. As one can see, the departure from linearity is rather modest for $(\lambda_1, \lambda_2) \in \{(0.05, 0), (0, 0.05)\}$. In Tables 1–3 we then report the empirical standard deviation of $\widehat{\tau}$, the average values of the standard errors $\widehat{\sigma}_{\text{EHW}}/\sqrt{N}$ and $\widehat{\sigma}_{\text{CRV}}/\sqrt{N}$, and the empirical coverage probabilities of $C^{1-\alpha}_{\text{EHW}}$ and $C^{1-\alpha}_{\text{CRV}}$ for $\alpha = 0.05$.

[11] That is, in each DGP we put put $\mathcal{X} = \{x_1, \ldots, x_{G^-+G^+}\}$ with $x_g = g/(G^- + 1) - 1$ if $1 \leq g \leq G^-$ and $x_g = (g - G^-)/(G^+ + 1) + 1$ if $G^- < g \leq G$; and $P(X_i = x_g) = 1/(2G^-)$ if $1 \leq g \leq G^-$ and $P(X_i = x_g) = 1/(2G^+)$ if $G^- < g \leq G$

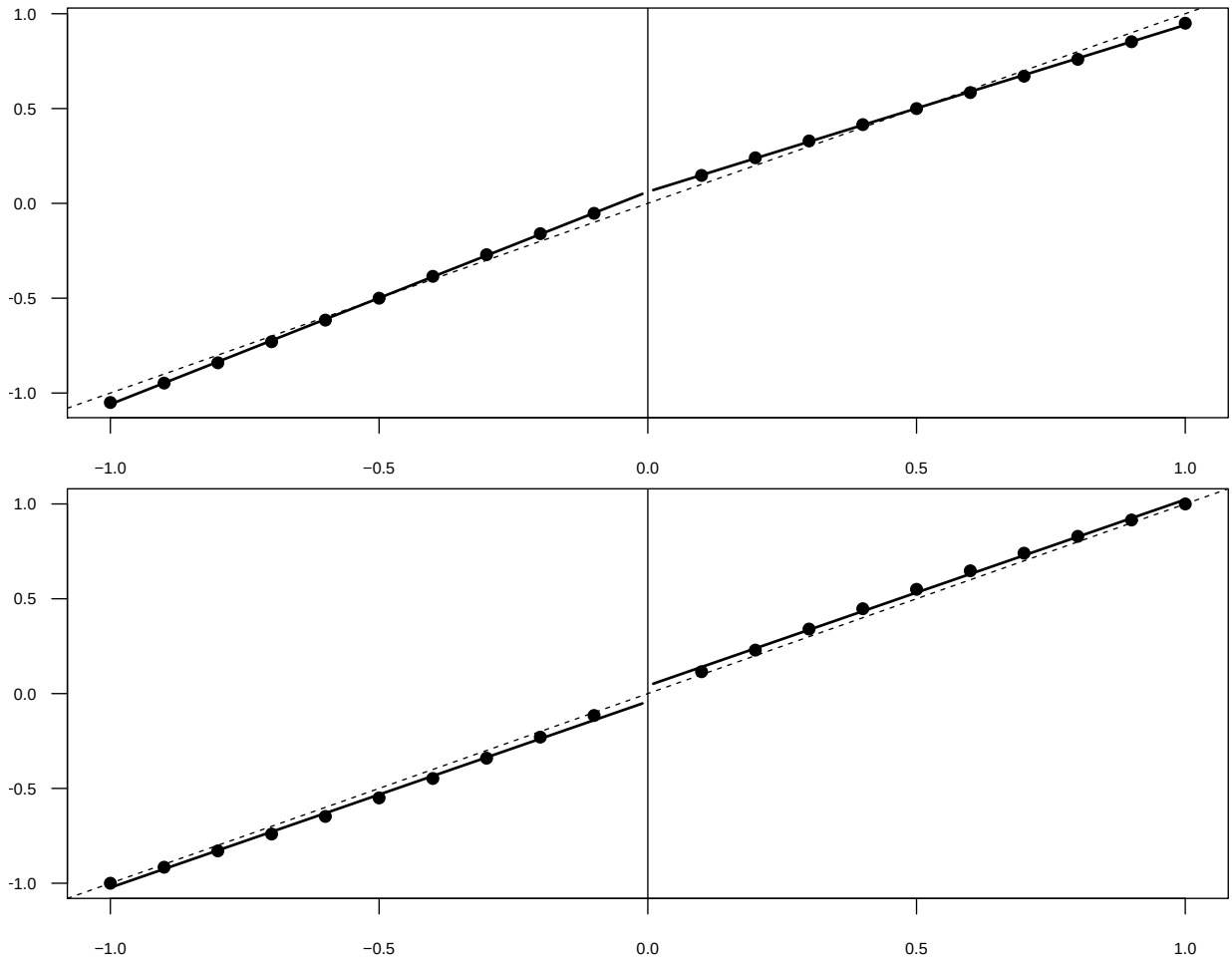

Figure 1: Plot of $\mu(x) = x + .05 \cdot \cos(\pi \cdot x)$ (top panel) and $\mu(x) = x + .05 \cdot \sin(\pi \cdot x)$ (bottom panel) for $G^- = G^+ = 10$. Dots indicate the value of the function at the respective support points of the running variable; solid lines correspond to linear fit above and below the threshold; dashed line indicates a line with intercept zero and slope one; and vertical line indicates the threshold.

Table 1 reports results for the case $(\lambda_1, \lambda_2) = (0, 0)$, in which the true conditional expectation function is linear and thus our fitted model is correctly specified. We see that the CRV standard error is a downward-biased estimate of the standard deviation of $\widehat{\tau}$, and therefore $C_{\text{CRV}}^{1-\alpha}$ under-covers the ATE. The distortion is most severe for the case with the least number of points on either side of the threshold ($G^- = G^+ = 5$), where it amounts to a deviation of 20 percentage points from the nominal level. With more support points the distortion becomes less pronounced, but it is still noticeable even for $G^- = G^+ = 50$. Note that these findings are the same for any of the sample sizes we consider.

Table 1: Simulation results for $\mu(x) = x$.

G^+	G^-	τ_h	N	$\text{sd}(\widehat{\tau})$	$\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$	$\mathbb{E}(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$	$P(\tau \in C^{0.95}_{\text{EHW}})$	$P(\tau \in C^{0.95}_{\text{CRV}})$
5	5	0	10^3	0.1485	0.1484	0.0970	0.9500	0.7552
			10^4	0.0469	0.0470	0.0307	0.9493	0.7530
			10^5	0.0149	0.0148	0.0097	0.9492	0.7524
			10^6	0.0047	0.0047	0.0031	0.9489	0.7559
5	25	0	10^3	0.1396	0.1398	0.1106	0.9497	0.8564
			10^4	0.0443	0.0442	0.0350	0.9491	0.8554
			10^5	0.0140	0.0140	0.0111	0.9489	0.8552
			10^6	0.0044	0.0044	0.0035	0.9498	0.8555
5	50	0	10^3	0.1390	0.1388	0.1120	0.9493	0.8666
			10^4	0.0438	0.0439	0.0354	0.9501	0.8677
			10^5	0.0139	0.0139	0.0112	0.9495	0.8666
			10^6	0.0044	0.0044	0.0035	0.9496	0.8678
25	25	0	10^3	0.1306	0.1305	0.1220	0.9495	0.9212
			10^4	0.0413	0.0413	0.0386	0.9490	0.9219
			10^5	0.0131	0.0130	0.0122	0.9496	0.9214
			10^6	0.0041	0.0041	0.0039	0.9500	0.9236
25	50	0	10^3	0.1298	0.1295	0.1230	0.9496	0.9292
			10^4	0.0410	0.0409	0.0389	0.9489	0.9289
			10^5	0.0130	0.0130	0.0123	0.9504	0.9313
			10^6	0.0041	0.0041	0.0039	0.9488	0.9284
50	50	0	10^3	0.1289	0.1285	0.1242	0.9481	0.9351
			10^4	0.0407	0.0406	0.0393	0.9493	0.9360
			10^5	0.0129	0.0129	0.0124	0.9492	0.9365
			10^6	0.0041	0.0041	0.0039	0.9499	0.9364

Note: Table reports standard deviation of $\widehat{\tau}$ ($\text{sd}(\widehat{\tau})$), average value of the robust and cluster-robust standard error estimators ($\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$ and $\mathbb{E}(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$), and empirical coverage of associated confidence intervals ($P(\tau \in C^{0.95}_{\text{EHW}})$ and $P(\tau \in C^{0.95}_{\text{CRV}})$).

Table 2 reports results for the case $(\lambda_1, \lambda_2) = (0, .05)$. Here $\mu(x)$ is nonlinear, but due to the symmetry properties of the cosine function $\tau_h = 0$. This setup mimics applications in which the bias of $\widehat{\tau}$ is small even though the functional form of $\mu(x)$ is misspecified. In line with our asymptotic approximations, the CRV standard error is downward biased for smaller values of N, and upward biased for larger sample sizes. Simulation results for the case that $N = 10^7$, which are not reported here, also confirm that the CRV standard error does not converge to zero. Correspondingly, $C^{0.95}_{\text{CRV}}$ under-covers the ATE for smaller values of N, and

Table 2: Simulation results for $\mu(x) = x + .05 \cdot \cos(\pi \cdot x)$.

G^+	G^-	τ_h	N	$\text{sd}(\widehat{\tau})$	$\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$	$E(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$	$P(\tau \in C^{0.95}_{\text{EHW}})$	$P(\tau \in C^{0.95}_{\text{CRV}})$
5	5	0	10^3	0.1486	0.1484	0.0969	0.9500	0.7529
			10^4	0.0471	0.0470	0.0312	0.9495	0.7602
			10^5	0.0149	0.0149	0.0113	0.9488	0.8184
			10^6	0.0047	0.0047	0.0066	0.9498	0.9854
5	25	0	10^3	0.1397	0.1398	0.1106	0.9501	0.8562
			10^4	0.0443	0.0442	0.0353	0.9492	0.8575
			10^5	0.0140	0.0140	0.0119	0.9506	0.8821
			10^6	0.0045	0.0045	0.0057	0.9488	0.9768
5	50	0	10^3	0.1386	0.1388	0.1121	0.9492	0.8680
			10^4	0.0440	0.0439	0.0357	0.9486	0.8679
			10^5	0.0139	0.0139	0.0120	0.9501	0.8903
			10^6	0.0044	0.0044	0.0056	0.9496	0.9756
25	25	0	10^3	0.1307	0.1305	0.1220	0.9492	0.9222
			10^4	0.0414	0.0413	0.0386	0.9488	0.9216
			10^5	0.0131	0.0131	0.0125	0.9493	0.9270
			10^6	0.0042	0.0042	0.0046	0.9478	0.9620
25	50	0	10^3	0.1293	0.1295	0.1231	0.9498	0.9299
			10^4	0.0410	0.0410	0.0390	0.9497	0.9302
			10^5	0.0129	0.0130	0.0125	0.9496	0.9338
			10^6	0.0041	0.0041	0.0045	0.9494	0.9609
50	50	0	10^3	0.1286	0.1285	0.1243	0.9488	0.9351
			10^4	0.0407	0.0406	0.0393	0.9493	0.9362
			10^5	0.0129	0.0129	0.0126	0.9478	0.9376
			10^6	0.0041	0.0041	0.0044	0.9493	0.9588

Note: Table reports standard deviation of $\widehat{\tau}$ ($\text{sd}(\widehat{\tau})$), average value of the robust and cluster-robust standard error estimators ($\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$ and $\mathbb{E}(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$), and empirical coverage of associated confidence intervals ($P(\tau \in C^{0.95}_{\text{EHW}})$ and $P(\tau \in C^{0.95}_{\text{CRV}})$).

over-covers for larger values. The distortions are again more pronounced for smaller values of G^+ and G^-.

Table 3 reports results for the case $(\lambda_1, \lambda_2) = (.05, 0)$. Here the linear model is misspecified as well, but in such a way that τ_h is substantially different from zero; with its exact value depending on G^+ and G^-. As with the previous sets of results, the CRV standard error is downward biased for smaller values of N, and upward biased for larger sample sizes. However, since $\tau_h \neq 0$ here, the coverage probability of $C^{0.95}_{\text{CRV}}$ is below the nominal level for

Table 3: Simulation results for $\mu(x) = x + .05 \cdot \sin(\pi \cdot x)$.

G^+	G^-	τ_h	N	$\text{sd}(\widehat{\tau})$	$\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$	$\mathbb{E}(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$	$P(\tau \in C^{0.95}_{\text{EHW}})$	$P(\tau \in C^{0.95}_{\text{CRV}})$
5	5	0.07	10^3	0.1488	0.1485	0.0990	0.8867	0.6641
			10^4	0.0472	0.0470	0.0371	0.3703	0.2566
			10^5	0.0150	0.0150	0.0239	0.0000	0.0005
			10^6	0.0052	0.0052	0.0225	0.0000	0.0000
5	25	0.03	10^3	0.1397	0.1398	0.1118	0.9014	0.7885
			10^4	0.0442	0.0443	0.0387	0.4714	0.3815
			10^5	0.0142	0.0142	0.0204	0.0000	0.0020
			10^6	0.0050	0.0050	0.0178	0.0000	0.0000
5	50	0.03	10^3	0.1387	0.1388	0.1132	0.9038	0.8033
			10^4	0.0440	0.0440	0.0388	0.4851	0.3977
			10^5	0.0141	0.0141	0.0197	0.0000	0.0019
			10^6	0.0050	0.0050	0.0170	0.0000	0.0000
25	25	0.07	10^3	0.1307	0.1305	0.1224	0.9130	0.8807
			10^4	0.0415	0.0413	0.0399	0.5901	0.5617
			10^5	0.0133	0.0133	0.0161	0.0002	0.0020
			10^6	0.0048	0.0048	0.0113	0.0000	0.0000
25	50	0.04	10^3	0.1293	0.1295	0.1235	0.9162	0.8915
			10^4	0.0411	0.0410	0.0400	0.6025	0.5815
			10^5	0.0131	0.0132	0.0154	0.0004	0.0019
			10^6	0.0047	0.0047	0.0100	0.0000	0.0000
50	50	0.06	10^3	0.1286	0.1285	0.1245	0.9176	0.9017
			10^4	0.0408	0.0407	0.0400	0.6166	0.6008
			10^5	0.0131	0.0131	0.0145	0.0006	0.0016
			10^6	0.0047	0.0047	0.0086	0.0000	0.0000

Note: Table reports standard deviation of $\widehat{\tau}$ ($\text{sd}(\widehat{\tau})$), average value of the robust and cluster-robust standard error estimators ($\mathbb{E}(\widehat{\sigma}_{\text{EHW}}/\sqrt{N})$ and $\mathbb{E}(\widehat{\sigma}_{\text{CRV}}/\sqrt{N})$), and empirical coverage of associated confidence intervals ($P(\tau \in C^{0.95}_{\text{EHW}})$ and $P(\tau \in C^{0.95}_{\text{CRV}})$).

all N, and tends to zero as the sample size increases. For smaller, somewhat more realistic values of N the coverage properties of $C^{0.95}_{\text{CRV}}$ are also worse than those of the standard CI $C^{0.95}_{\text{EHW}}$. The CRV CI only performs better in a relative sense than $C^{0.95}_{\text{EHW}}$ when N is very large, but for these cases both CIs are heavily distorted and have coverage probability very close to zero. So in absolute terms the performance of $C^{0.95}_{\text{CRV}}$ is still poor.

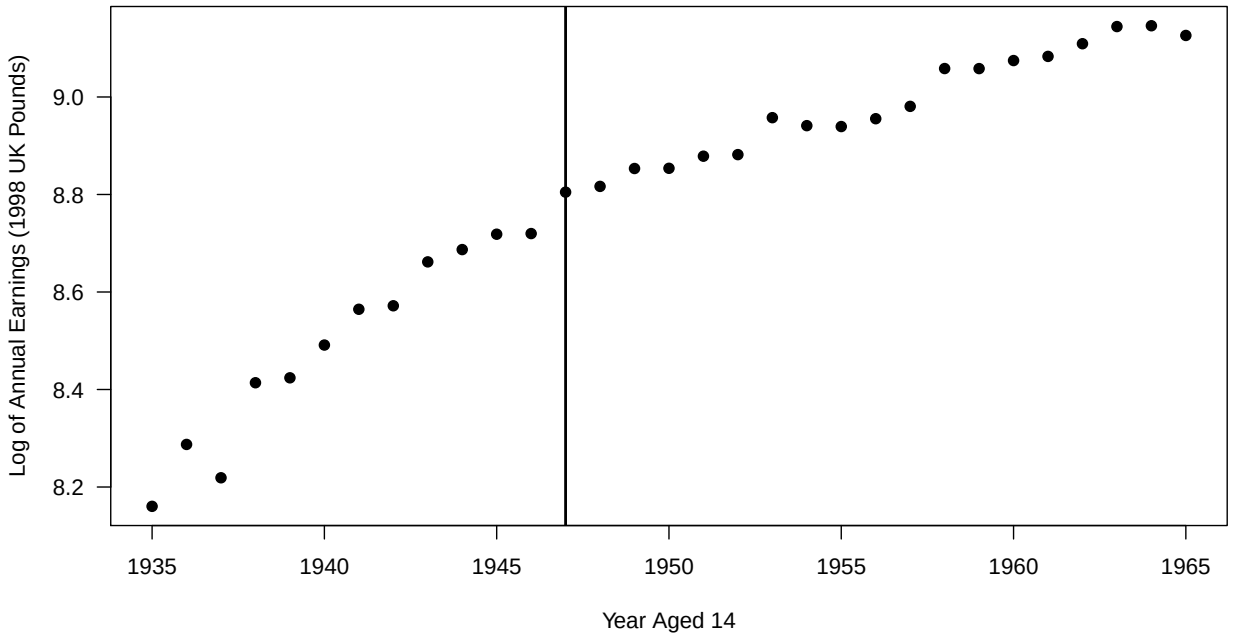

Figure 2: Average of natural logarithm of annual earnings by year aged 14. Vertical line indicates the year 1947, in which the minimum school-leaving age changed from 14 to 15.

4.2. **Empirical illustration**

Here we illustrate the impact of clustering on the running variable using data from Oreopoulos (2006), who studies the effect of a change in the minimum school-leaving age in the United Kingdom from 14 to 15 on schooling attainment and annual earnings in 1998. The change occurred in 1947 in Great Britain (England, Scotland and Wales), and in 1957 in Northern Ireland. The data are a random sample of UK workers who turned 14 between 1935 and 1965; see Oreopoulos (2006) for details.

For simplicity, we focus on the sub-sample of British workers, and restrict attention to the effect of being "treated" with a higher minimum school-leaving age on (the natural logarithm of) wages in 1998.[12] Oreopoulos (2006) uses a discrete RDD to estimate this parameter. The running variable is the year in which the worker turned 14, and the treatment threshold is 1947. The running variable thus has $G = 31$ support points, of which $G^+ = 19$ are above

[12]Our aim is not to provide a full replication of every result in Oreopoulos (2006), or to single out this particular study in any way. Instead, the findings in this section are meant as a simple illustration of the implications of our theoretical results in a setting with real data.

the threshold, and $G^- = 12$ ones are below. Oreopoulos (2006) uses the global specification

$$\log(EARN_i) = \gamma_0 + \sum_{k=1}^{4} \gamma_k \cdot YEAR14_i^k + \tau \cdot \mathbb{I}\{YEAR14_i \geq 1947\} + \varepsilon_i. \tag{4.1}$$

He obtains an ATE point estimate of 0.055 with a CRV standard error of 0.015, which corresponds to the 95% CRV CI (0.026, 0.084)[13].

In Table 4 we report point estimates along with CRV and conventional EHW standard errors and CIs for the original specification (column (1)), and for linear and quadratic specifications fitted separately on each side of the threshold using either the full data set (columns (2)–(3)), all data points within a bandwidth of $h = 6$ years around the threshold (columns (4)–(5)), or all data points within a bandwidth of $h = 3$ years (columns (6)–(7)).[14]

For Oreopoulos' original specification in column (1), the EHW standard error is twice as large as CRV standard error; and the corresponding 95% EHW CI covers zero, whereas the 95% CRV CI does not. For the linear specification using the full data in column (2), the point estimate is negative, and EHW standard errors are slightly smaller than CRV ones. Given Figure 2, the former finding seems to be due to substantial misspecification of a global linear model below the threshold, while the latter finding seems to be due to random variation in standard errors.

For the remaining specifications in Table 4, all EHW standard errors are larger than the CRV standard errors, by a factor between 1.7 and 31.7. For both linear and quadratic specifications, the factor generally increases as the bandwidth (and thus the number of support points that are being used for estimation) decreases. Moreover, the factor is larger for quadratic specifications than it is for the linear specifications. None of the EHW CIs in Table 4 contain zero, which implies that for the specifications in columns (1), (3) and (5)–(7), LC's approach leads to incorrect claims about the statistical significance of the estimated treatment effect.[15]

[13] See the revised version of Table 1 in the online data archive of Oreopoulos (2006), available at `http://www.aeaweb.org/articles.php?doi=10.1257/000282806776157641`. Note that he uses STATA's formula for computing cluster-robust variance estimates, $\widehat{\Omega}_{\text{STATA}} = G/(G-1) \times (N-1)/(N-k) \cdot \widehat{\Omega}_{\text{CRV}}$, where k is the number of regressors. To maintain comparability, we also use this formula for our calculations in this subsection.

[14] We consider different specifications in order to illustrate how these affect the relative magnitude of CRV and EHW standard errors. The question whether or not the estimated effect is significantly different from zero is secondary for our purposes. See e.g. Devereux and Hart (2010) for a discussion of the sensitivity of the point estimates in Oreopoulos (2006) with respect to model specification.

[15] While our analysis does not imply that EHW CIs have correct coverage in this setting, it does imply that any CI with correct coverage must be at least as wide as the EHW CI.

Table 4: Estimated effect of being subject to increases minimum school-leaving age on natural logarithm of annual earnings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimate	.055	-.011	.042	.021	.085	.065	.110
CRV SE	.015	.026	.019	.019	.016	.008	.004
CRV CI	(.026, .084)	(-.062, .041)	(.005, .078)	(-.016, .059)	(.054, .117)	(.049, .081)	(.102, .119)
EHW SE	.030	.023	.038	.033	.058	.049	.127
EHW CI	(-.003, .113)	(-.056, .035)	(-.032, .115)	(-.043, .085)	(-.029, .199)	(-.031, .161)	(-.138, .359)
Polyn. order	4	1	2	1	2	1	2
Separate fit	No	Yes	Yes	Yes	Yes	Yes	Yes
Localization	No	No	No	$h = 6$	$h = 6$	$h = 3$	$h = 3$
Eff. sample size	73,954	73,954	73,954	20,883	20,883	10,533	10,533

Note: Estimates use data for Great Britain only. See Oreopoulos (2006) for further details.

5. HONEST CONFIDENCE INTERVALS

The results in the previous sections caution against clustering standard errors by the running variable in empirical applications, in spite of the great popularity of the approach. In this section, we present alternative approaches to conduct inference in RDDs with a discrete running variable.

It is important to realize that if the regression function $\mu(x)$ is allowed to vary arbitrarily between the two support points of the running variable closest to the threshold, no method for inference on the ATE can be both valid and informative, because even in large samples any point on the real line remains a feasible candidate for the value of τ. To make progress, one therefore needs to place restrictions on $\mu(x)$. We formalize these restrictions by requiring that $\mu \in \mathcal{M}$ for some class of functions $\mathcal{M}$, and seek to construct CIs $C^{1-\alpha}$ that satisfy

$$\lim_{N\to\infty} \inf_{\mu^* \in \mathcal{M}} P_{\mu^*}(\tau \in C^{1-\alpha}) \geq 1-\alpha, \tag{5.1}$$

where P_{μ^*} denotes probability under $\mu(x) = \mu^*(x)$. Such a CI is guaranteed to have good coverage properties uniformly over all regression functions in $\mathcal{M}$. Following Li (1989), we refer these CIs as *honest with respect to* $\mathcal{M}$. It is desirable for a CI to be honest with respect to some meaningful and interpretable function class. In the following subsections, we discuss two examples of such function classes, outline how honest CIs can be constructed in each case, and illustrate their use in the context of the empirical application from Section 4.2. Both function classes can be interpreted as classes of functions which can be well-approximated by pth-order polynomials, but they differ in how this notion is formalized.

5.1. **Bounds on misspecification errors at the threshold**

One way to restrict $\mu(x)$ is to impose that the magnitude of the left limit of the misspecification bias at the threshold is no larger than the magnitude of the misspecification bias at any point in the support of the running variable below the threshold, and similarly for the right limit. That is, one could assume that $\mu(x)$ takes values in

$$\mathcal{M}_h = \{\mu^* : |\lim_{x\downarrow 0} \delta(x)| \leq \max_{x\in\mathcal{X}_h^+} |\delta(x)| \text{ and } |\lim_{x\uparrow 0} \delta(x)| \leq \max_{x\in\mathcal{X}_h^-} |\delta(x)|\},$$

where $\delta(x) \equiv \delta_{\mu^*}(x) = \mu^*(x) - m(x)'\theta_{h,\mu^*}$, with

$$\theta_{h,\mu^*} = \operatorname*{argmin}_{\theta} \mathbb{E}_{\mu^*}((\mu^*(X_i) - m(X_i)'\theta)^2 \mathbb{I}\{|X_i| \leq h\}),$$

is the difference between $\mu^*(x)$ and its best polynomial approximation $m(x)'\theta_{h,\mu^*}$ over $(-h,h)$, and h is taken to be fixed. Elements of $\mathcal{M}_h$ are thus approximately pth order polynomials in the sense that that the misspecification error is bounded.

The assumption that $\mu \in \mathcal{M}_h$ implies that τ is partially identified, in the sense that we can deduce from the population distribution of the observables (Y,X) that

$$\tau \in \mathcal{T} = (\tau_h - \max_{x\in\mathcal{X}_h^+} |\delta(x)| - \max_{x\in\mathcal{X}_h^-} |\delta(x)|, \tau_h + \max_{x\in\mathcal{X}_h^+} |\delta(x)| + \max_{x\in\mathcal{X}_h^-} |\delta(x)|),$$

where $\mathcal{T}$ is called the identified set. Note that a disadvantage of working with $\mathcal{M}_h$ is that the class depends on the distribution of the running variable through the definition of the best polynomial approximation θ_{h,μ^*}.

To develop a honest CI in this setup, let $\widehat{\delta}(x_g) = \widehat{\mu}(x_g) - m(x_g)'\widehat{\theta}$, with $\widehat{\mu}(x_g) = \sum_{i=1}^N Y_i \mathbb{I}\{X_i = x_g\}/n_g$ and $n_g = \sum_{i=1}^N \mathbb{I}\{X_i = x_g\}$, be the an estimate of $\delta(x_g)$, and let $\widehat{\sigma}_g^2 = \sum_{i=1}^N (Y_i - \widehat{\mu}(x_g))^2 \mathbb{I}\{X_i = x_g\}/n_g$ and $\widehat{\pi}_g = n_g/N$ be estimates of $\sigma_g^2 = \mathbb{V}(Y_i|X_i = x_g)$ and $\pi_g = P(X_i = x_g)$, respectively. It is also useful to write the identified set as

$$\mathcal{T} = \bigcup_{\psi\in\Psi} \mathcal{T}_\psi, \quad \text{where} \quad \mathcal{T}_\psi = (\tau_h - s^+\delta(x^+) - s^-\delta(x^-), \tau_h + s^+\delta(x^+) + s^-\delta(x^-)),$$

$\Psi = G_h^- \times G_h^+ \times \{-1,1\}^2$ and $\psi = (x^-, x^+, s^-, s^+)$ is a generic element of Ψ, in order to avoid the occurrences of the absolute value operators. Note that some $\mathcal{T}_\psi$ are generally empty.

Now suppose for a moment we knew that $\tau \in \mathcal{T}_\psi$ for one particular value of $\psi \in \Psi$, and consider the simpler problem of constructing a $(1-\alpha)$ left-sided CI $[c_{l,\psi}^\alpha, \infty)$ for τ in this case. A natural choice for $c_{l,\psi}^\alpha$ in this context would be $\widehat{\tau} - s^+\widehat{\delta}(x^+) - s^-\widehat{\delta}(x^-) + p_{l,\psi}^\alpha/\sqrt{N}$, where $p_{l,\psi}^\alpha$ is an appropriate critical value. Since we assumed that $\tau \in \mathcal{T}_\psi$ for some known value of $\psi \in \Psi$, it holds that

$$\begin{aligned} P\left(\tau \le \widehat{\tau} - s^+\widehat{\delta}(x_{g^+}) - s^-\widehat{\delta}(x_{g^-}) + p_{l,\psi}^\alpha/\sqrt{N}\right) \\ \le P\left(\sqrt{N}\left(s^+(\widehat{\delta}(x_{g^+}) - \delta(x_{g^+})) + s^-(\widehat{\delta}(x_{g^-}) - \delta(x_{g^-})) - (\widehat{\tau} - \tau_h)\right) \le p_{l,\psi}^\alpha\right), \end{aligned}$$

and since $(\widehat{\mu}(x_1), \ldots, \widehat{\mu}(x_G))'$ and $\widehat{\theta}$ are jointly asymptotically normal, a somewhat tedious application of the delta method outlined in Appendix B yields that

$$\sqrt{N}\left(s^+(\widehat{\delta}(x_{g^+}) - \delta(x_{g^+})) + s^-(\widehat{\delta}(x_{g^-}) - \delta(x_{g^-})) - (\widehat{\tau} - \tau_h)\right) \xrightarrow{d} \mathcal{N}(0, V_\psi),$$

where the asymptotic variance is given by

$$V_\psi = \frac{\sigma^2_{g^-}}{\pi_{g^-}} + \frac{\sigma^2_{g^+}}{\pi_{g^+}} + a'_\psi Q^{-1}\Omega Q^{-1} a_\psi - 2(s^+\sigma^2_{g^+} m_{g^+} + s^-\sigma^2_{g^-} m_{g^-})' Q^{-1} a_\psi,$$

with $a_\psi = (s^+ m_{g^+} + s^- m_{g^-} + e_1)'$. A feasible choice for the critical value is thus

$$p^\alpha_{l,\psi} = \Phi^{-1}(\alpha)\widehat{V}^{1/2}_\psi / \sqrt{N},$$

where $\Phi^{-1}(\alpha)$ denotes the α quantile of the standard normal distribution and

$$\widehat{V}_\psi = \frac{\widehat{\sigma}^2_{g^-}}{\widehat{\pi}_{g^-}} + \frac{\widehat{\sigma}^2_{g^+}}{\widehat{\pi}_{g^+}} + a'_\psi \widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1} a_\psi - 2(s^+\widehat{\sigma}^2_{g^+} m_{g^+} + s^-\widehat{\sigma}^2_{g^-} m_{g^-})' \widehat{Q}^{-1} a_\psi,$$

is the natural estimate of V_ψ.

Because we do not know which of the sets $\mathcal{T}_\psi$ contains τ, we consider the union of the sets $[c^\alpha_{l,\psi}, \infty)$ over $\psi \in \Psi$, which we denote by

$$C^{1-\alpha}_l = \bigcup_{\psi \in \Psi} [\widehat{\tau} - s^+\widehat{\delta}(x^+) - s^-\widehat{\delta}(x^-) + p^\alpha_{l,\psi}/\sqrt{N}, \infty).$$

Following arguments in Berger (1982), $C^{1-\alpha}_l$ is indeed a valid left-sided $(1-\alpha)$ CI for τ. Berger (1982) also shows that this type of CI can be interpreted as being based on inverting the decision of a hypothesis with strong optimality properties. Note that we can write $C^{1-\alpha}_l$ in a somewhat more intuitive form:

$$C^{1-\alpha}_l = [c^\alpha_l, \infty) \quad \text{with} \quad c^\alpha_l = \widehat{\tau} - \max_{\psi \in \Psi} \left(s^+\widehat{\delta}(x^+) + s^-\widehat{\delta}(x^-) + \Phi^{-1}(1-\alpha)\widehat{V}^{1/2}_\psi / \sqrt{N} \right).$$

Through similar reasoning, we can also obtain a right-sided $(1-\alpha)$ CI for τ:

$$C^{1-\alpha}_r = (\infty, c^\alpha_r] \quad \text{with} \quad c^\alpha_r = \widehat{\tau} + \max_{\psi \in \Psi} \left(s^+\widehat{\delta}(x^+) + s^-\widehat{\delta}(x^-) + \Phi^{-1}(1-\alpha)\widehat{V}^{1/2}_\psi / \sqrt{N} \right).$$

Finally, an intersection of the right- and left-sided CIs with an appropriately adjusted nominal level yields a two-sided $(1-\alpha)$ CI for τ:

$$C^{1-\alpha}_t = (c^{\alpha/2}_l, c^{\alpha/2}_r).$$

Proposition 4. *The CIs $C^{1-\alpha}_s$, $s \in \{l, r, t\}$, are honest with respect to $\mathcal{M}_h$:*

$$\lim_{N\to\infty} \inf_{\mu \in \mathcal{M}_h} P_\mu(\tau \in C^{1-\alpha}_s) \geq 1-\alpha \text{ for } s \in \{l, r, t\}.$$

We remark that these confidence intervals are explicitly tailored to settings where X_i is discrete, and at least a moderate number of observations is available for every support point within the estimation window. This is because the approach is built on the asymptotic normality of the estimators $(\widehat{\mu}(x_1), \ldots, \widehat{\mu}(x_G))'$. This construction has no analogue in settings where X_i exhibits near-continuous variation.

5.2. **Bound on the second derivative**

Another way of restricting $\mu(x)$ is through smoothness assumptions. Perhaps the most natural way of imposing smoothness is to assume that μ is twice differentiable on either side of the cutoff, with a bounded second derivative. For technical reasons, we consider a closure of this family (i.e. only require μ to be twice differentiable almost everywhere). Let $\mu_+(x) = \mu(x)\mathbb{I}\{x \geq 0\}$ and $\mu_-(x) = \mu(x)\mathbb{I}\{x < 0\}$ denote the parts of μ above and below the cutoff. We put

$$\mathcal{M} = \mathcal{M}(K) = \{\mu_+ - \mu_- \colon \mu_+ \in \mathcal{M}(K, \mathbb{R}_+),\ \mu_- \in \mathcal{M}(K, \mathbb{R}_-)\},$$

where $\mathcal{M}(K, \mathcal{X})$ is a second-order Hölder class,

$$\mathcal{M}(K, \mathcal{X}) = \{\mu \colon |\mu'(x) - \mu'(y)| \leq K|x - y| \ \ x, y \in \mathcal{X}\}.$$

In other words, $\mathcal{M}(K)$ consists of functions that are continuous on either side of the cutoff, twice-differentiable almost everywhere, and the second derivative is bounded in absolute value by a known constant K.

We want to construct CIs that are honest with respect to $\mathcal{M}(K)$, and that are based on a local polynomial estimator $\widehat{\tau}$ defined in Section 2. This problem was considered in Armstrong and Kolesár (2016b).[16] We outline the solution specialized to the case $p = 1$, so that $\widehat{\tau}$ is a local linear estimator. It turns out that it is easier to construct CIs that satisfy a slightly stronger condition than (5.1) in that they are honest conditional on the running variables X_i. Let $\widetilde{\tau}_h = \mathbb{E}(\widehat{\tau} \mid X_1, \ldots, X_N)$ denote conditional expectation of the estimator, and let $\widehat{\sigma}^2/N_h$ denote an estimator of $\mathbb{V}(\widehat{\tau} \mid X_1, \ldots, X_N)$, the conditional variance of $\widehat{\tau}$. When the conditional expectation function is non-linear, so that $\widetilde{\tau}_h \neq \tau$, Abadie et al. (2014) show that the EHW estimator $\widehat{\sigma}^2_{\text{EHW}}/N_h$ is conservative (i.e. it overestimates the conditional variance), and they propose a nearest-neighbor variance estimator that is consistent under mild regularity conditions. For the honest CI that we describe below, one can use either

[16] The related problem of honest testing for a jump in the density of the running variable at the threshold has been considered in Frandsen (2016), who uses a bound on the second derivative of the density that yields a class of densities similar to $\mathcal{M}(K)$.

variance estimator as we only require that as $N \to \infty$,

$$\frac{\widehat{\sigma}^2/N_h}{\mathbb{V}(\widehat{\tau} \mid X_1, \ldots, X_N)} \geq 1 + o_P(1), \tag{5.2}$$

where the $o_P(1)$ term is uniform over $\mathcal{M}(K)$. To derive the form of the honest CI, decompose the t-statistic based on $\widehat{\tau}$ as

$$\frac{\widehat{\tau} - \tau}{\widehat{\sigma}/\sqrt{N_h}} = \frac{\widehat{\tau} - \widetilde{\tau}_h}{\widehat{\sigma}/\sqrt{N_h}} + \frac{\widetilde{\tau}_h - \tau}{\widehat{\sigma}/\sqrt{N_h}}.$$

Under mild regularity conditions, the first term is normally distributed in large samples with mean zero, and variance at most one (the variance is equal to one if equation (5.2) holds with equality). The second term is bounded in absolute value by

$$b(\widehat{\tau}) = \frac{\sup_{\mu \in \mathcal{M}(K)} |\tau - \widetilde{\tau}_h|}{\widehat{\sigma}/\sqrt{N_h}}.$$

Therefore, the appropriate critical value is $\mathrm{cv}_{1-\alpha}(b(\widehat{\tau}))$, where $\mathrm{cv}_{1-\alpha}(b)$ is the $1-\alpha$ quantile of the $|\mathcal{N}(b,1)|$ distribution (instead of the usual z_α critical value which is the $1-\alpha$ quantile of the $|\mathcal{N}(0,1)|$ distribution). Thus, a honest CI is given by

$$C^{1-\alpha}_{\mathcal{M}(K)} = \left(\widehat{\tau} \pm \mathrm{cv}_{1-\alpha}(b(\widehat{\tau}))\widehat{\sigma}/\sqrt{N_h} \right).$$

The next proposition summarizes these results and gives an explicit expression for $b(\widehat{\tau})$. To state the result, note that the estimator can be written as a linear estimator, $\widehat{\tau} = \sum_{i=1}^N w(X_i) Y_i$, with weights

$$\begin{aligned} w(x) = \mathbb{I}\{h \geq x \geq 0\} \frac{\sum_{i:\, h \geq X_i \geq 0} X_i^2 - x \sum_{i:\, h \geq X_i \geq 0} X_i}{N_{h,+} \sum_{i:\, h \geq X_i \geq 0} X_i^2 - (\sum_{i:\, h \geq X_i \geq 0} X_i)^2} \\ - \mathbb{I}\{h \geq -x > 0\} \frac{\sum_{i:\, h \geq -X_i > 0} X_i^2 - x \sum_{i:\, h \geq -X_i > 0} X_i}{N_{h,-} \sum_{i:\, h \geq -X_i > 0} X_i^2 - (\sum_{i:\, h \geq -X_i > 0} X_i)^2}, \end{aligned}$$

where $N_{h,+} = \sum_{i=1}^N \mathbb{I}\{h \geq X_i \geq 0\}$, and $N_{h,-} = \sum_{i=1}^N \mathbb{I}\{h \geq -X_i > 0\}$.

Proposition 5. *Consider a local linear estimator $\widehat{\tau}$, and let $\widehat{\sigma}^2/N_h$ be an estimator of its asymptotic variance. Then*

$$b(\widehat{\tau}) = -\frac{K}{2} \frac{\sum_{i:\, X_i \geq 0} w(X_i)(\mathbb{I}\{X_i \geq 0\} - \mathbb{I}\{X_i < 0\})}{\widehat{\sigma}/N_h}.$$

Furthermore, if, as $N \to \infty$, $(\widehat{\tau} - \tau_h)/\mathbb{V}(\widehat{\tau} \mid X_1, \ldots, X_N)$ converges, uniformly over $\mathcal{M}(K)$,

to a standard normal random variable and (5.2) *holds, then* $C^{1-\alpha}_{\mathcal{M}(K)}$ *is a honest CI with respect to* $\mathcal{M}(K)$.

As can be seen from the proof given in in Appendix B, the proposition obtains essentially as a special case of a more general setup considered in Armstrong and Kolesár (2016b). This CI has several attractive features. First, as can be seen from the proof, the same construction can be used whether the running variable is discrete or continuous: one need not make a distinction between these two cases. Second, $C^{1-\alpha}_{\mathcal{M}(K)}$ takes into account the exact finite-sample bias of the estimator, and does not rely on asymptotic promises about what the bandwidth would have been had the sample size been larger. In particular, the CI remains valid even if the bandwidth is fixed. In contrast, the usual CIs in RDDs rely on the bandwidth to shrink sufficiently fast relative to the sample size so that equation (2.2) holds, and consequently shrink at a slower rate than $C^{1-\alpha}_{\mathcal{M}(K)}$. Third, to achieve the tightest possible CI, one can simply choose the bandwidth that minimizes the value of $\mathrm{cv}_{1-\alpha}(b(\widehat{\tau}))\widehat{\sigma}$.

While the CIs considered in Proposition 4 use the fit of the polynomial approximation $m(x)'\widehat{\theta}$ away from the threshold point to bound the bias at the threshold, $C^{1-\alpha}_{\mathcal{M}(K)}$ uses the bound on the second derivative K. In practice, the researcher must therefore choose an appropriate bound K; it it not possible to use a data-driven method without additional assumptions. In particular, it follows from the results in Low (1997), Cai and Low (2004) and Armstrong and Kolesár (2016a) that it is not possible to form honest CIs that are tighter using data-dependent tuning parameters, and maintain coverage over the whole function class $\mathcal{M}(K)$, for some conservative upper bound K, relative to the CI in Proposition 5.

5.3. **Empirical illustration**

For illustration, we compute the CIs proposed in the previous two subsections for the empirical application to returns to schooling in Section 4.2. Table 5 shows that the procedure proposed in Section 5.1 leads to CIs that are wider than the EHW CIs. The difference is most pronounced for Oreopoulos' original specification in column (1) and the specifications that use the full data in column (2)–(3). For the cases $h = 6$ and $h = 3$, the honest CI is only slightly wider than the EHW CIs. All honest CIs cover the value zero, which confirms the finding from Section 4.2 that no statistically significant effect can be detected.

Table 5 also shows honest CIs based on the construction in Section 5.2. We consider the values $K = 0.003$ and $K = 0.03$. The former corresponds to a very optimistic bound on the smoothness of the conditional expectation function, while the latter value is more conservative. For $K = 0.003$, the bandwidth that leads to the tightest possible CI is given by $h = 6$, which corresponds to the specification in column (4). The resulting CI is given

Table 5: Alternative confidence intervals for effect of being subject to increases minimum school-leaving age on natural logarithm of annual earnings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimate	.055	-.011	.042	.021	.085	.065	.110
BME CI	(-.245, .354)	(-.334, .313)	(-.220, .303)	(-.132, .175)	(-.105, .275)	(-.072, .202)	(-.150, .371)
BSD CI ($K = .003$)				(-.055, .097)			
BSD CI ($K = .03$)						(-.082, .211)	
Polyn. order	4	1	2	1	2	1	2
Separate fit	No	Yes	Yes	Yes	Yes	Yes	Yes
Localization	No	No	No	$h = 6$	$h = 6$	$h = 3$	$h = 3$
Eff. sample size	73,954	73,954	73,954	20,883	20,883	10,533	10,533

Note: Table reports CI assuming bounded misspecification errors at the threshold (BME; see Proposition 4) and CIs assuming bounds on the second derivative of μ (BSD; see Proposition 5), for the same specifications considered in Table 4 above.

by $(-0.055, 0.097)$, which takes into account that the estimator may be biased, with the bias equal to at most 0.022 in absolute value. For $K = 0.03$, the optimal bandwidth is given by $h = 3$, which corresponds to the specification in column (6). The bias of the estimator is bounded by 0.066, and the resulting CI is given by $(-0.082, 0.211)$. In line with the conclusions in Section 4.2, these results again indicate that the effect of an increase in minimum schooling age on earnings is not significant.

6. CONCLUSIONS

RDDs with a discrete running variable are ubiquitous in empirical practice. In this paper, we show that the commonly used CIs based on standard errors that are clustered by the running variable have poor coverage properties, and therefore recommend that they should not be used in practice. We suggest a more attractive approach to inference that is based on formalizing the notion that the conditional expectation function can be well-approximated by a polynomial. We discuss two such restrictions on the conditional expectation function that are easily interpretable, and construct CIs with guaranteed coverage properties under these restrictions. To bound the bias of the estimator, the first method uses the fit of the polynomial approximation to the conditional expectation function away from the threshold to, and the second method uses a bound on the second derivative of the conditional expectation function. The second method has the advantage that the resulting CI is valid whether the running variable is discrete or continuous.

A. PROOFS OF RESULTS IN SECTION 3

The proof of the propositions in Section 3 follows directly from general results on the properties of $\widehat{\sigma}^2_{\text{CRV}}$ that are given in the following subsection. The proofs of these results are given in turn in Sections A.2–A.4. To state these results, we use the notation $\text{diag}\{a_g\}$ to denote a diagonal matrix with diagonal elements given by $a_1, \ldots, a_G$, and $\text{vec}\{a_g\} = (a_1', \ldots, a_G')'$.

A.1. Properties of $\widehat{\sigma}^2_{\text{CRV}}$ under General Conditions

In this subsection, we consider a setup that is slightly more general than that in Section 3, in that it also allows the bandwidth h to change with the sample size. For convenience, the following assumption summarizes this more general setup.

Assumption 1 (Model)**.** *For each N, the data $\{Y_i, X_i\}_{i=1}^N$ are i.i.d., distributed according to a law P_N. Under P_N, the marginal distribution of X_i is discrete with $G = G_- + G_+$ support points denoted $x_1 < \cdots < x_{G^-} < 0 \leq x_{G^-+1} < \cdots < x_G$. Let $\mu(x) = \mathbb{E}_N(Y_i \mid X_i = x)$ denote*

the conditional expectation under P_N. Let $\varepsilon_i = Y_i - \mu(X_i)$, and let $\sigma_g^2 = \mathbb{V}_N(\varepsilon_i \mid X_i = x_g)$ denote its conditional variance. Let $h = h_N$ denote a non-random bandwidth sequence, and let $\mathcal{G}_h \subseteq \{1, \ldots, G\}$ denote the indices for which $|x_g| \leq h$. Let $\pi_g = P_N(X_i = x_g)$, $\pi = P_N(|X_i| \leq h)$, and $N_h = \sum_{i=1}^N \mathbb{I}\{|X_i| \leq h\}$. For a fixed p, define

$$m(x) = (\mathbb{I}\{x \geq 0\}, 1, x, \ldots, x^p, \mathbb{I}\{x \geq 0\} x, \ldots, \mathbb{I}\{x \geq 0\} x^p)',$$

$M_i = \mathbb{I}\{|X_i| \leq h\} m(X_i)$, and $m_g = \mathbb{I}\{|x_g| \leq h\} m(x_g)$. Let $\widehat{Q} = N_h^{-1} \sum_{i=1}^n M_i M_i'$ and $Q_N = \mathbb{E}_N(M_i M_i')/\pi$. Let $\theta_h = Q_N^{-1} \mathbb{E}_N(m(X_i) Y_i \mid |X_i| \leq h)$, and let $\widehat{\theta} = \widehat{Q}^{-1} \frac{1}{N_h} \sum_{i=1}^N M_i Y_i$. Define $\delta(x) = \mu(x) - m(x)'\theta_h$, and $u_i = Y_i - X_i'\theta_h = \delta(X_i) + \epsilon_i$. Define $\Omega = \mathbb{E}_N[u_i^2 M_i M_i' \mid |X_i| \leq h] = \sum_{g=1}^G (\sigma_g^2 + \delta^2(x_g)) Q_g$, where $Q_g = \frac{\pi_g}{\pi} m_g m_g'$, and suppose that $N\pi \to \infty$.

Note that the setup allows the various quantities that depend on P_N and h to change with N, such as the number of support points G, their locations x_g, the conditional expectation function $\mu(x)$, or the specification errors $\delta(X_i)$.

Assumption 2 (Regularity conditions)**.** *(i)* $\sup_N \max_{g \in \mathcal{G}_h} \mathbb{E}_N(\epsilon_i^4 \mid X_i = x_g) < \infty$, *and* $\sup_N \max_{g \in \mathcal{G}_h} \delta(x_g) < \infty$*; and (ii)* $\det(H^{-1} Q_N H^{-1}) = \det(\sum_{g \in \mathcal{G}_N} \frac{\pi_g}{\pi} m(x_g/h) m(x_g/h)') > C$ *for some $C > 0$ that doesn't depend on N, where $H = \operatorname{diag}\{m(h)\}$, and the limits $\lim_{N\to\infty} H^{-1} \Omega H^{-1}$ and $\lim_{N\to\infty} H^{-1} Q_N H^{-1}$ exist.*

The first part of the assumption corresponds to the regularity assumptions in Section 3. The second part of the assumption ensures that the parameter θ_h and the asymptotic variance of $\widehat{\theta}$ remain well-defined as the bandwidth shrinks to zero.[17]

Our first result is an asymptotic approximation in which G_h^+ and G_h^- are fixed as the sample size increases. Let $B_1, \ldots, B_G$ be a collection of random vectors such that $\operatorname{vec}\{B_g\} \sim \mathcal{N}(0, V)$, with

$$V = \frac{1}{\pi} \operatorname{diag}\{\pi_g(\sigma_g^2 + \delta(x_g))\} - \frac{1}{\pi} \operatorname{vec}\{\pi_g \delta(x_g)\} \operatorname{vec}\{\pi_g \delta(x_g)\}'.$$

Note that if $|x_g| > h$, then $B_g = 0$ and $Q_g = 0$, and that the limiting distribution of the statistic $\sqrt{N_h}(\widehat{\tau} - \tau_h)$ coincides with the distribution of $e_1' Q_N^{-1} \sum_{g=1}^G m_g B_g$. Finally, define

$$W_g = e_1' Q_N^{-1} m_g \left(B_g - \frac{\pi_g}{\pi} m_g' Q_N^{-1} \sum_{j=1}^G m_j B_j + \sqrt{\frac{N}{\pi}} \pi_g \delta(x_g) \right).$$

With this notation, we obtain the following generic result.

[17] It is necessary to normalize the quantities by the inverse of H since if $h \to 0$, elements of $\widehat{Q}$ and $\widehat{\theta}$ converge at different rates.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Suppose also that, as $N \to \infty$, (i) G_h^+ and G_h^- are fixed; and (iii) the limit of V exists. Then*

$$\widehat{\sigma}^2_{\text{CRV}} \overset{d}{=} (1 + o_{P_N}(1)) \sum_{g=1}^{G} W_g^2.$$

Our second result is an asymptotic approximation in which the number of support points of the running variable (or, equivalently, the number of "clusters") that are less than h away from the threshold increases with the sample size.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Suppose also that, as $N \to \infty$, $G_h \to \infty$ and $\max_{g \in \mathcal{G}_h} \pi_g / \pi \to 0$. Then*

$$\widehat{\sigma}^2_{\text{CRV}} = (1 + o_{P_N}(1)) e_1' Q_N^{-1} \left(\Omega + (N-1) \sum_{g=1}^{G} Q_g \cdot \pi_g \delta(x_g)^2. \right) Q_N^{-1} e_1.$$

The assumption that $\max_{g \in \mathcal{G}_h} \pi_g / \pi \to 0$ ensures that each "cluster" comprises a vanishing fraction of the effective sample size.

A.2. **Auxiliary lemma**

Here we state an intermediate result that is used in the proofs of Theorem 1 and 2 below.

Lemma 1. *Suppose that Assumption 1 holds. Then*

$$\frac{N_h/N}{\pi} = 1 + o_{P_N}(1), \tag{A.1}$$

$$H^{-1} \widehat{Q} H^{-1} - H^{-1} Q_N H^{-1} = o_{P_N}(1). \tag{A.2}$$

Suppose, in addition, that Assumption 2 holds. Then

$$\sqrt{N_h} H (\widehat{\theta} - \theta_h) \overset{d}{=} H Q_N^{-1} S + o_{P_N}(1),$$

where $S \sim \mathcal{N}(0, \Omega)$. Let $n_g = \sum_{i=1}^{N} \mathbb{I}\{X_i = x_g\}$, $\widehat{q}_g = H \widehat{Q}^{-1} H m(x_g/h) \mathbb{I}\{|x_g| \leq h\}$, and $A_g = \frac{\mathbb{I}\{|x_g| \leq h\}}{\sqrt{N_h}} \sum_{i=1}^{N} (\mathbb{I}\{X_i = x_g\} \varepsilon_i + (\mathbb{I}\{X_i = x_g\} - \pi_g) \delta(x_g))$. Then $\sum_{g=1}^{G} m_g A_g \overset{d}{=} H^{-1} S + o_{P_N}(1)$, and

$$\widehat{\sigma}^2_{\text{CRV}} = \sum_{g=1}^{G} (e_1' \widehat{q}_g)^2 \left(A_g + \frac{N}{\sqrt{N_h}} \pi_g \delta(x_g) - \frac{n_g}{N_h} \widehat{q}_g' \sum_{j=1}^{G} m(x_j/h) A_j \right)^2. \tag{A.3}$$

Proof. We have $\mathbb{V}_N(N_h/N) = \pi(1-\pi)/N \leq \pi/N$. Therefore, by Markov inequality, $N\pi \to \infty$ implies $\frac{N_h/N}{\pi} = \mathbb{E}_N(N_h/(N\pi)) + o_{P_N}(1) = 1 + o_{P_N}(1)$, which proves (A.1). Secondly, since elements of $H^{-1} M_i$ are bounded by $\mathbb{I}\{|X_i| \leq h\}$, the variance of any element of $\frac{N_h}{N\pi} \widehat{Q}$

is bounded by $(1-\pi)/(N\pi)$, which converges to 0 as $N\pi \to \infty$. Combining this result with (A.1) and Markov inequality yields (A.2).

Next note that $\sum_{i=1}^{G} m_g A_g = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\sqrt{\pi}} H^{-1} M_i u_i$, and that by the central limit theorem, $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\sqrt{\pi}} H^{-1} M_i u_i \xrightarrow{d} H^{-1} S + o_{P_N}(1)$. Therefore,

$$\sqrt{N_h} H(\widehat{\theta} - \theta_h) = \sqrt{\frac{\pi N}{N_h}} (H^{-1} \widehat{Q} H^{-1})^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\sqrt{\pi}} H^{-1} M_i u_i \xrightarrow{d} H Q_N^{-1} S + o_{P_N}(1),$$

as claimed. Next, the cluster-robust variance estimator can be written as

$$\widehat{\sigma}^2_{\text{CRV}} = e_1' \widehat{Q}^{-1} \frac{1}{N_h} \sum_{g=1}^{G} \mathbf{M}_g' \widehat{\mathbf{u}}_g \widehat{\mathbf{u}}_g' \mathbf{M}_g \widehat{Q}^{-1} e_1 = \sum_{g=1}^{G} \left(\frac{1}{\sqrt{N_h}} e_1' \widehat{Q}^{-1} \mathbf{M}_g' \widehat{\mathbf{u}}_g \right)^2.$$

We now decompose the expression in parentheses. Since

$$H\sqrt{N_h}(\widehat{\theta} - \theta_h) = H \widehat{Q}^{-1} H \sum_{g=1}^{G} m(x_g/h) A_g,$$

we have

$$\begin{aligned}
\frac{1}{\sqrt{N_h}} e_1' \widehat{Q}^{-1} \mathbf{M}_g' \widehat{\mathbf{u}}_g &= e_1' \widehat{q}_g \frac{1}{\sqrt{N_h}} \sum_{i=1}^{N} \mathbb{I}\{X_i = x_g\} \widehat{u}_i \\
&= e_1' \widehat{q}_g \left(\frac{1}{\sqrt{N_h}} \sum_{i=1}^{N} \mathbb{I}\{X_i = x_g\} u_i + \frac{n_g}{\sqrt{N_h}} m(x_g/h) H(\theta_h - \widehat{\theta}) \right) \\
&= e_1' \widehat{q}_g \left(A_g + \frac{N}{\sqrt{N_h}} \pi_g \delta(x_g) - \frac{n_g}{N_h} \widehat{q}_g' \sum_{j=1}^{G} m(x_j/h) A_j \right),
\end{aligned}$$

which yields the result. □

A.3. **Proof of Theorem 1**

Let $q_g = H Q_N^{-1} H m(x_g/h) \mathbb{I}\{|x_g| \leq h\}$, and define $\widehat{q}_g$, A_g, and n_g as in the statement of Lemma 1. By Lemma 1, $\widehat{q}_g = q_g(1 + o_{P_N}(1))$, and by Markov inequality, $\pi_g/(N\pi) = \pi_g/\pi + o_{P_N}(1)$ for $g \in \mathcal{G}_h$. Since G_h is fixed, combining these results with Equations (A.1) and (A.3), it follows that cluster-robust variance estimator satisfies

$$\widehat{\sigma}^2_{\text{CRV}} = (1 + o_{P_N}(1)) \sum_{g=1}^{G} (e_1' q_g)^2 \left(A_g + \sqrt{N\pi} \frac{\pi_g}{\pi} \delta(x_g) - \frac{\pi_g}{\pi} q_g' \sum_{j=1}^{G} m(x_j/h) A_j \right)^2,$$

To prove the theorem, it therefore suffices to show that

$$e_1' q_g \left(A_g - \frac{\pi_g}{\pi} q_g' \sum_{j=1}^{G} m(x_j/h) A_j + \sqrt{N\pi} \frac{\pi_g}{\pi} \delta(x_g) \right) = W_g(1 + o_{P_N}(1)) \tag{A.4}$$

In turn, this expression follows from Slutsky's lemma if we can show that

$$\text{vec}\{A_g\} \stackrel{d}{=} \text{vec}\{B_g\}(1 + o_{P_N}(1)), \tag{A.5}$$

This in turn follows by $N\pi \to \infty$ and the central limit theorem.

A.4. **Proof of Theorem 2**

Throughout the proof, write $a \preceq b$ to denote $a < Cb$ for some constant C that does not depend on N. By Lemma 1, we can write the cluster-robust estimator as

$$\begin{aligned} \widehat{\sigma}^2_{\text{CRV}} = \sum_{g=1}^{G} (e_1' \widehat{q}_g)^2 \left(A_g + \frac{N}{\sqrt{N_h}} \pi_g \delta(x_g) \right)^2 + \widehat{S}' H^{-1} \sum_{g=1}^{G} (\widehat{q}_g' e_1)^2 \frac{n_g^2}{N_h^2} \widehat{q}_g \widehat{q}_g' \cdot H^{-1} \widehat{S} \\ - 2 \sum_{g=1}^{G} (e_1' \widehat{q}_g)^2 \left(A_g + \frac{N}{\sqrt{N_h}} \pi_g \delta(x_g) \right) \frac{n_g}{N_h} \widehat{q}_g' H^{-1} \widehat{S}, \end{aligned}$$

where $\widehat{S} = H \sum_{j=1}^{G} m(x_j/h)' A_j$, and n_g, A_g and $\widehat{q}_g$ are defined in the statement of the Lemma. Denote the three summands by $\mathcal{C}_1, \mathcal{C}_2$, and $\mathcal{C}_3$, respectively.

We first show that $\mathcal{C}_2 = o_{P_N}(1)$. Since $H^{-1}\widehat{S} = O_{P_N}(1)$ by Lemma 1, it suffices to show that

$$\sum_{g=1}^{G} (\widehat{q}_g' e_1)^2 \frac{n_g^2}{N_h^2} \widehat{q}_g \widehat{q}_g' = o_{P_N}(1).$$

To this end, note that since elements of $m(x_g/h)$ are bounded by 1, for any j, by Cauchy-Schwarz inequality, $|\widehat{q}_g' e_j| \leq \|e_j' H \widehat{Q}^{-1} H\|_2 2(p+1)$, where $\|v\|_2$ denotes the Euclidean norm of a vector v. Since $\|e_j' H \widehat{Q}^{-1} H\|_2 = O_{P_N}(1)$ and $N_h/\pi N = 1 + o_{P_N}(1)$ by Lemma 1,

$$\left| \sum_{g=1}^{G} (\widehat{q}_g' e_1)^2 \frac{n_g^2}{N_h^2} e_j \widehat{q}_g \widehat{q}_g' e_k \right| \leq O_{P_N}(1) \sum_{g \in \mathcal{G}_h} \frac{n_g^2}{N_h^2} = O_{P_N}(1) \sum_{g \in \mathcal{G}_h} \frac{n_g^2}{\pi^2 N^2}$$

Now, since $\mathbb{E}_N(n_g^2) = N\pi_g(1-\pi_g) + N^2\pi_g^2$, and $\sum_{g \in \mathcal{G}_h} \pi_g = \pi$,

$$\mathbb{E}_N \sum_{g \in \mathcal{G}_h} \frac{n_g^2}{N^2\pi^2} = \sum_{g \in \mathcal{G}_h} \frac{\pi_g(1-\pi_g)}{N\pi^2} + \sum_{g \in \mathcal{G}_h} \frac{\pi_g^2}{\pi^2} \leq \left(\frac{1}{N\pi} + \frac{\max_{g \in \mathcal{G}_h} \pi_g}{\pi} \right) \sum_{g \in \mathcal{G}_h} \frac{\pi_g}{\pi} \to 0.$$

Therefore, by Markov inequality, $\sum_{g \in \mathcal{G}_h} \frac{n_g^2}{\pi^2 N^2} = o_{P_N}(1)$, so that $\mathcal{C}_2 = o_{P_N}(1)$ as claimed.

Now consider $\mathcal{C}_1$. Let $q_g = HQ_N^{-1}Hm(x_g/h)\mathbb{I}\{|x_g| \leq h\}$. We have

$$\begin{aligned}
\mathcal{C}_1 &= \frac{1}{N_h}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{g=1}^{G}(e_1'\widehat{q}_g)^2\mathbb{I}\{X_i = x_g\}\,\mathbb{I}\{X_j = x_g\}\,(\varepsilon_i + \delta(x_g))(\varepsilon_j + \delta(x_g)) \\
&= (1 + o_{P_N}(1))\frac{1}{N\pi}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,\mathbb{I}\{X_j = x_g\}\,(\varepsilon_i + \delta(x_g))(\varepsilon_j + \delta(x_g)) \\
&= (1 + o_{P_N}(1))\left(\mathcal{C}_{11} + 2(\mathcal{C}_{12} + \mathcal{C}_{13} + \mathcal{C}_{14} + \mathcal{C}_{15} + \mathcal{C}_{16})\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{C}_{11} &= \frac{1}{N\pi}\sum_{i=1}^{N}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,(\varepsilon_i + \delta(x_g))^2, \\
\mathcal{C}_{12} &= \frac{1}{N\pi}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,\mathbb{I}\{X_j = x_g\}\,\varepsilon_i\varepsilon_j, \\
\mathcal{C}_{13} &= \frac{1}{N\pi}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,\mathbb{I}\{X_j = x_g\}\,\varepsilon_j\delta(x_g), \\
\mathcal{C}_{14} &= \frac{1}{N\pi}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,\mathbb{I}\{X_j = x_g\}\,\varepsilon_i\delta(x_g), \\
\mathcal{C}_{15} &= \frac{1}{N\pi}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{g=1}^{G}(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,(\mathbb{I}\{X_j = x_g\} - \pi_g)\delta(x_g)^2 \\
\mathcal{C}_{16} &= \frac{1}{N\pi}\sum_{g=1}^{G}\sum_{i=1}^{N}(i-1)(e_1'q_g)^2\mathbb{I}\{X_i = x_g\}\,\pi_g\delta(x_g)^2.
\end{aligned}$$

We have

$$\mathbb{E}_N(\mathcal{C}_{11}) = \frac{1}{\pi}\sum_{g=1}^{G}(e_1'q_g)^2\pi_g(\sigma_g^2 + \delta(x_g)^2) = e_1'Q_N^{-1}\Omega Q_N^{-1}e_1,$$

and

$$\mathbb{V}(\mathcal{C}_{11}^2) \leq \frac{1}{N\pi^2}\sum_{g=1}^{G}(e_1'q_g)^4\pi_g\mathbb{E}_N[(\varepsilon_i + \delta(x_g))^4 \mid X_i = x_g] \preceq \frac{\sum_{g\in\mathcal{G}_h}\pi_g}{N\pi^2} = \frac{1}{N\pi} \to 0.$$

Next, $\mathbb{E}_N(\mathcal{C}_{12}) = 0$, and

$$\mathbb{V}(\mathcal{C}_{12}) = \frac{N-1}{2N\pi^2}\sum_{g=1}^{G}(e_1'q_g)^2\pi_g^2\sigma_g^2\sigma_g^2 \preceq \frac{\max_g\pi_g\sum_{g=1}^{G}\pi_g}{\pi^2} = \frac{\max_g\pi_g}{\pi} \to 0.$$

The expectations for the remaining terms satisfy $\mathbb{E}_N(\mathcal{C}_{13}) = \mathbb{E}_N(\mathcal{C}_{14}) = \mathbb{E}_N(\mathcal{C}_{15}) = 0$, and

$$\mathbb{E}_N(\mathcal{C}_{16}) = \frac{N-1}{2\pi}\sum_{g=1}^{G}(e_1'q_g)^2\pi_g^2\delta(x_g)^2.$$

The variances of $\mathcal{C}_{13},\ldots,\mathcal{C}_{16}$ are all of smaller order than this expectation:

$$\begin{aligned}
\mathbb{V}(\mathcal{C}_{13}) &= \frac{1}{N^2\pi^2}\sum_{g}^{G}\sum_{i,k=1}^{N}\sum_{j=1}^{\min\{i,k\}-1}(e_1'q_g)^4\pi_g^3\sigma_g^2\delta(x_g)^2 \preceq \frac{N\max_g\pi_g}{\pi^2}\sum_{g}^{G}(e_1'q_g)^2\pi_g^2\delta(x_g)^2 \\
&= o(\mathbb{E}_N(\mathcal{C}_{16})) \\
\mathbb{V}(\mathcal{C}_{14}) &= \frac{1}{N^2\pi^2}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{k=1}^{i-1}\sum_{g=1}^{G}(e_1'q_g)^4\pi_g^3\sigma_g^2\delta(x_g)^2 = o(\mathbb{E}_N(\mathcal{C}_{16})) \\
\mathbb{V}(\mathcal{C}_{15}) &= \frac{1}{N^2\pi^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{\min\{i,k\}-1}\sum_{g,f=1}^{G}\left(\mathbb{I}\{g=f\}\,\pi_g-\pi_g\pi_f\right)\pi_g\pi_f(e_1'q_g)^2(e_1'q_f)^2\delta(x_f)^2\delta(x_g)^2 \\
&\leq \frac{1}{N^2\pi^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\sum_{j=1}^{\min\{i,k\}-1}\sum_{g=1}^{G}\pi_g^3(e_1'q_g)^4\delta(x_g)^4 = o(\mathbb{E}_N(\mathcal{C}_{16})),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}(\mathcal{C}_{16}^2) &= \frac{1}{N^2\pi^2}\sum_{g=1}^{G}\sum_{f=1}^{G}\sum_{i=1}^{N}(i-1)^2(\mathbb{I}\{g=f\}\,\pi_g-\pi_g\pi_f)\pi_g\pi_f\delta(x_g)^2\delta(x_f)^2(e_1'q_g)^2(e_1'q_f)^2 \\
&\leq \frac{N}{\pi^2}\sum_{g=1}^{G}\pi_g^3\delta(x_g)^4(e_1'q_g)^4 = o(\mathbb{E}_N(\mathcal{C}_{16})).
\end{aligned}$$

It therefore follows that

$$\mathcal{C}_1 = (1+o_{P_N}(1))\mathbb{E}_N(\mathcal{C}_1) = (1+o_{P_N}(1))\left(e_1'Q_N^{-1}\Omega Q_N^{-1}e_1 + \frac{N-1}{\pi}\sum_{g=1}^{G}(e_1'q_g)^2\pi_g^2\delta(x_g)^2\right).$$

Finally, the cross-term $\mathcal{C}_3$ is $o(\mathbb{E}_N(\mathcal{C}_1)^{1/2})$ by Cauchy-Schwarz inequality, so that $\widehat{\sigma}^2_{\mathrm{CRV}} = (1+o_{P_N}(1))\mathbb{E}_N(\mathcal{C}_1)$, which yields the result.

B. PROOFS OF RESULTS IN SECTION 5

B.1. **Proof of Proposition 4**

Under the conditions of the proposition, it holds that

$$\sqrt{N}\begin{pmatrix}\mathrm{vec}(\widehat{\delta}(x_g)-\delta(x_g)) \\ \widehat{\tau}_h-\tau_h\end{pmatrix} \xrightarrow{d} \mathcal{N}(0,\Sigma),$$

where the variance matrix Σ is given by

$$\Sigma = \begin{pmatrix} \text{diag}(\sigma_g^2/\pi_g) + \mathcal{W} & \text{vec}(\sigma_g^2 m_g' Q^{-1} e_1 - m_g' Q^{-1} \Omega Q^{-1} e_1) \\ \text{vec}(\sigma_g^2 m_g' Q^{-1} e_1 - m_g' Q^{-1} \Omega Q^{-1} e_1)' & e_1' Q^{-1} \Omega Q^{-1} e_1 \end{pmatrix},$$

and $\mathcal{W}$ is a $G \times G$ matrix with (g, g^*) element equal to $m_g' Q^{-1} \Omega Q^{-1} m_{g^*} - (\sigma_g^2 + \sigma_{g^*}^2) m_g' Q^{-1} m_{g^*}$. To see that this is true, note that by the central limit theorem,

$$\begin{aligned} &\sqrt{N} \begin{pmatrix} \text{vec}(n_g/N) - \text{vec}(\pi_g) \\ \text{vec}(n_g \widehat{\mu}_g/N) - \text{vec}(\pi_g \mu(x_g)) \end{pmatrix} \\ &\xrightarrow{d} N\left(0, \begin{pmatrix} \text{diag}(\pi_g) & \text{diag}(\pi_g \mu(x_g)) \\ \text{diag}(\pi_g \mu(x_g)) & \text{diag}(\pi_g(\mu(x_g)^2 + \sigma_g^2)) \end{pmatrix} - \begin{pmatrix} \text{vec}(\pi_g) \\ \text{vec}(\pi_g \mu(x_g)) \end{pmatrix} \begin{pmatrix} \text{vec}(\pi_g) \\ \text{vec}(\pi_g \mu(x_g)) \end{pmatrix}' \right). \end{aligned}$$

By the delta method

$$\sqrt{N} \begin{pmatrix} \text{vec}(\widehat{\mu}_g - \mu_g) \\ N^{-1} \sum_{i=1}^N M_i u_i \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \text{diag}(\sigma_g^2/\pi_g) & \text{vec}(\sigma_g^2 m_g') \\ \text{vec}(\sigma_g^2 m_g')' & \Omega \end{pmatrix}\right),$$

where $\Omega = \sum_g \pi_g(\delta_g^2 + \sigma_g^2) m_g m_g'$. Therefore, we get

$$\begin{aligned} \sqrt{N} \begin{pmatrix} \text{vec}(\widehat{\delta}(x_g) - \delta(x_g)) \\ \widehat{\tau}_h - \tau_h \end{pmatrix} &= \sqrt{N} \begin{pmatrix} I_G & -\text{vec}(m_g' Q^{-1}) \\ 0 & e_1' Q^{-1} \end{pmatrix} \begin{pmatrix} \text{vec}(\widehat{\mu}_g - \mu_g) \\ \frac{1}{N} \sum_i M_i u_i \end{pmatrix} (1 + o_P(1)) \\ &\xrightarrow{d} \mathcal{N}(0, \Sigma), \end{aligned}$$

as claimed, where $\widehat{Q} = E_n[M_i M_i'] \xrightarrow{P} Q = \mathbb{E}[M_i M_i']$. It also follows from standard arguments that the $o_P(1)$ term in the previous equations is uniformly $o_P(1)$ over $\mathcal{M}_h$. Honesty of the confidence intervals then follows from Berger (1982), along the lines of the argument given in the main part of the paper.

B.2. **Proof of Proposition 5**

We first derive the expression for $b(\widehat{\tau})$, following the arguments in Theorem B.1 in Armstrong and Kolesár (2016a). Put $w_+(x) = w(x)\mathbb{I}\{x \geq 0\}$, and $w_-(x) = -w(x)\mathbb{I}\{x < 0\}$. Note that these weights satisfy $\sum_{i=1}^N X_i w_+(X_i) = 0$, and $\sum_{i=1}^N X_i w_-(X_i) = 0$, so that the conditional bias $\widehat{\tau}$ at μ is that same as the bias at $\mu_+(x) - x\mu_+'(0) - (\mu_-(x) - x\mu_-'(0))$, using the convention that $\mu_-(0) = \lim_{x \uparrow 0} \mu_-(x)$. We can therefore without loss of generality assume $\mu_+'(0) = \mu_-'(x) = 0$. By assumption, the first derivatives of the functions μ_+ and μ_- are Lipschitz, and hence absolutely continuous, so that, by the Fundamental Theorem of Calculus and Fubini's theorem, we can write, for $x \geq 0$, $\mu_+(x) = \mu_+(0) + \int_0^x \mu''(s)(x - s)\, ds$, and for

$x \leq 0$, $\mu_-(x) = \mu_-(0) + \int_x^0 \mu''(s)(x-s)\,ds$. The conditional bias can therefore be written as:

$$\begin{aligned}
\tilde{\tau}_h - \tau &= \sum_i w_+(X_i)(\mu_+(X_i) - \mu_+(0)) - \sum_i w_-(X_i)(\mu_-(X_i) - \mu_-(0)) \\
&= \sum_{i\colon X_i \geq 0} w(X_i) \int_0^{X_i} \mu''(s)(X_i - s)\,ds + \sum_{i\colon X_i < 0} w(X_i) \int_{X_i}^0 \mu''(s)(X_i - s)\,ds \\
&= \int_0^\infty \mu''(s) \sum_{i\colon X_i \geq s} w(X_i)(X_i - s)\,ds + \int_{-\infty}^0 \mu''(s) \sum_{i\colon X_i \leq -s} w(X_i)(X_i - s)\,ds,
\end{aligned}$$

where the first line uses $\sum_{i=1}^N w_+(X_i) = \sum_{i=1}^N w_-(X_i) = 1$ and $\tau = \mu_+(0) - \mu_-(0)$, and the last line uses Fubini's theorem to change the order of summation and integration. Next, note that $\bar{w}_+(s) = \sum_{i\colon X_i \geq s} w(X_i)(X_i - s)$ is negative for all $s \geq 0$, because $\bar{w}_+(0) = 0$, $\bar{w}_+(s) = 0$ for $s \geq h$, and $\bar{w}'_+(s) = -\sum_{X_i \geq s} w(X_i)$ is monotone on $[0, h]$ with $\bar{w}'_+(0) = -1$. Similarly, $\bar{w}_-(s) = \sum_{i\colon X_i \geq s} w(X_i)(X_i - s)$ is positive for all $s \geq 0$. Therefore, the expression in the preceding display is maximized by setting $\mu''(x) = -K \operatorname{sign}(x)$, and minimized by setting $\mu''(x) = K \operatorname{sign}(x)$, which leads to

$$\sup_{\mu \in \mathcal{M}(K)} |\tilde{\tau}_h - \tau| = -\frac{K}{2} \sum_{i\colon X_i \geq 0} w(X_i)(\mathbb{I}\{X_i \geq 0\} - \mathbb{I}\{X_i < 0\}),$$

proving the first part of the proposition. The second part of the proposition follows from Theorem E.1 in Armstrong and Kolesár (2016a).

REFERENCES

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for Misspecified Models with Fixed Regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Armstrong, T. B. and M. Kolesár (2016a): "Optimal inference in a class of regression models," ArXiv:1511.06028.

——— (2016b): "Simple and honest confidence intervals in nonparametric regression," ArXiv:1606.01200.

Berger, R. (1982): "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, 24, 295–300.

Cai, T. T. and M. G. Low (2004): "An adaptation theory for nonparametric confidence intervals," *Annals of Statistics*, 32, 1805–1840.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

Cameron, C. A., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414–427.

Cameron, C. A. and D. L. Miller (2014): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50, 317–372.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2015): "Randomization Tests under an Approximate Symmetry Assumption," Technical Report No. 2014-13, Stanford University.

Card, D., C. Dobkin, and N. Maestas (2008): "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," *American Economic Review*, 98, 2242–58.

Chetty, R., J. N. Friedman, and E. Saez (2013): "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review*, 103, 2683–2721.

Clark, D. and H. Royer (2013): "The Effect of Education on Adult Mortality and Health: Evidence from Britain," *American Economic Review*, 103, 2087–2120.

Devereux, P. J. and R. A. Hart (2010): "Forced to be Rich? Returns to Compulsory Schooling in Britain," *Economic Journal*, 120, 1345–1364.

Dong, Y. (2015): "Regression discontinuity applications with rounding errors in the running variable," *Journal of Applied Econometrics*, 30, 422–446.

Frandsen, B. R. (2016): "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete," Working Paper.

Gelman, A. and G. Imbens (2014): "Why high-order polynomials should not be used in regression discontinuity designs," NBER Working Paper No. 20405.

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

Imbens, G. W. and M. Kolesár (2016): "Robust Standard Errors in Small Samples: Some Practical Advice," *Review of Economics and Statistics*, forthcoming.

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

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.

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

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

Martorell, P. and I. McFarlin (2011): "Help or hindrance? The effects of college remediation on academic and labor market outcomes," *Review of Economics and Statistics*, 93, 436–454.

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

Sacks, J. and D. Ylvisaker (1978): "Linear estimation for approximately linear models," *Annals of Statistics*, 1122–1137.

Schochet, P., T. Cook, J. Deke, G. Imbens, J. Lockwood, J. Porter, and J. Smith (2010): "Standards for Regression Discontinuity Designs." *What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education.*

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