Bias-aware inference in the regression discontinuity design

Master Thesis Presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Dominik Liebl

Submitted in March 2023 by:
Sven Jacobs
Matriculation Number: 2879411

Contents

1 Introduction **1**

2 Regression discontinuity designs **2**

2.1 Sharp regression discontinuity design 2

2.2 Fuzzy and other designs 4

3 Identification **4**

4 Estimation **5**

4.1 Parametric estimation: Global polynomials 6

4.2 Nonparametric estimation: Local polynomials 6

4.3 Bandwidth selection 8

5 Inference **10**

5.1 Undersmoothing 11

5.2 Robust bias-correction 11

5.3 Inflated critical value 12

5.4 Including covariates 14

6 Validation **14**

6.1 Manipulation of the assignment variable 15

6.2 Placebo cutoffs 16

6.3 Bandwidth sensitivity and donut holes 16

7 Application **16**

8 Simulation **21**

8.1 Setup 22

8.2 Results and discussion 24

9 Conclusion **27**

Appendix **28**

References **34**

1 Introduction

The regression discontinuity (RD) design is regarded as one of the most credible non-experimental research designs, widely applied in economics for treatment effect analysis. The key feature of any RD design is that treatment assignment changes abruptly at a known threshold value of an assignment variable, which is called the cutoff. Such cutoff rules can often be found in practice. An example is the award of a scholarship when a pupil scores above a certain point count in a standardized test. Here, the assignment variable is the test score, the cutoff is the required point count, and the treatment is being awarded the scholarship. A research question might be what effect the scholarship has on later academic achievements. In the basic setup which we focus on ("sharp" design), the actual treatment status relates one-to-one to the treatment assignment. In contrast, when the treatment probability changes discontinuously at the cutoff, but not sharply by one, the design is labeled "fuzzy".

Under certain assumptions, mainly one that ensures the comparability of just-treated and just-untreated units, the effect of the treatment at the cutoff can be estimated as the (potential) jump in the outcome of interest at the cutoff. For reasons we discuss later, RD estimation is generally considered a nonparametric estimation problem. Whereas estimation is straightforward once a bandwidth is selected, inference for the RD treatment effect poses a challenge. The reason is the (smoothing) bias associated with the nonparametric estimation, rendering conventional inference, in general, invalid.

The goal of this thesis is to provide an overview and illustration of modern (sharp) RD analysis, focusing on recent approaches to conduct valid bias-aware inference. We discuss identification, estimation, validation and present the three main inference procedures taking bias into account. An application to real data illustrates all the steps in a sharp RD analysis, and a Monte Carlo simulation investigates the finite-sample performance of the asymptotically valid inference procedures.

The RD design was initially proposed and applied in 1960 by Thistlethwaite and Campbell (impact on merit awards on future academic outcomes), but received little attention until the late 1990s, when some studies exploiting the RD design were published in leading journals (e.g. Angrist and Lavy (1999), Black (1999)). From then on the RD design has become an active and still rapidly expanding research area. Arguably the most important theoretical contribution is due to Hahn et al. (2001), who presented formal identification results and recommended local linear estimation. The two main contributions to the bias-aware RD literature came from Calonico et al. (2014) with their proposed robust bias-correction, and Armstrong and Kolesár (2020) proposing to use inflated critical values based on the maximal possible bias. For early general reviews see Imbens and Lemieux (2008), and Lee and Lemieux (2010); for an up-to-date review Cattaneo and Titiunik (2022).

The structure of the thesis is as follows. In the next section, we formally introduce the sharp RD design and briefly consider extensions to the basic setup. Section 3 discusses

identification, and Section 4 estimation. In Section 5 the challenge of valid inference is emphasized and the bias-aware approaches are presented. All the steps of a sound RD analysis are illustrated with a real-data application in Section 7, which is a reanalysis of a prominent study by Ludwig and Miller (2007). Afterwards, the finite-sample behavior of the bias-aware procedures is evaluated by means of a Monte Carlo study. Section 9 concludes and the appendix contains mostly tables and figures from the application.

2 Regression discontinuity designs

We start with an overview of RD designs. First, we formally introduce the so-called sharp design, where compliance with treatment assignment is perfect. We then touch on departures from the standard sharp design. Particularly the case of imperfect compliance, i.e. when treatment assignment and receipt do not coincide. The remainder of the thesis, however, focuses on the sharp design.

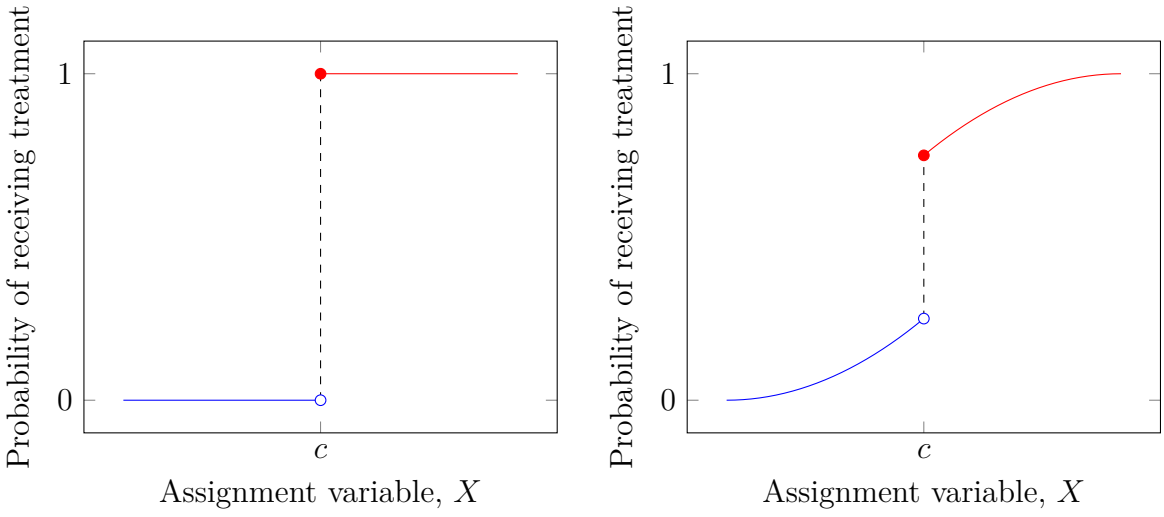
2.1 Sharp regression discontinuity design

We assume to have a random sample of n units, indexed by $i = 1, \ldots, n$. For each unit we observe X_i, the value of a continuous assignment variable X. There exists a known cutoff c, such that units with $X_i \geq c$ are assigned to treatment. Units with $X_i < c$ are assigned to the control group. However, it is important to distinguish between being assigned to treatment and receiving treatment. When some units do not comply with their assigned condition, we call this imperfect compliance. In the sharp RD design such imperfect compliance is ruled out. That is, all units assigned to the treatment group receive the treatment ($T = 1$), while all units assigned to the control group do not receive the treatment ($T = 0$). As is shown in Figure 1a, the conditional probability of receiving treatment $\mathrm{P}(T = 1|X = x)$ therefore jumps "sharply" from zero to one at the cutoff.

To formalize treatment analysis, we employ the potential outcomes framework (Rubin, 2005). The framework assumes that each unit has two potential outcomes, $Y_i(0)$ and $Y_i(1)$. $Y_i(0)$ is the outcome that would be observed under control, and $Y_i(1)$ is the outcome that would be observed under treatment. The fundamental problem of causal inference is that for each unit we only observe one outcome,

$$Y_i = (1 - T_i) \cdot Y_i(0) + T_i \cdot Y_i(1) = \begin{cases} Y_i(0) & , X_i < c \\ Y_i(1) & , X_i \geq c \end{cases}.$$

Figure 2 illustrates this fundamental problem for the sharp design. The figure shows two expected potential outcome functions, $\mathrm{E}[Y(0)|X = x]$ and $\mathrm{E}[Y(1)|X = x]$. Both functions are plotted partly solid and partly dashed. The reason is that in practice we can never observe realizations of the dashed function segments. All units below the cutoff ($X_i < c$) are in the control group. Thus, outcomes under treatment for this group are not observable



(a) Sharp (b) Fuzzy

Figure 1: Conditional probability of receiving treatment in RD designs. Units to the left of the cutoff c are assigned to control (in blue), units to the right to treatment (in red).

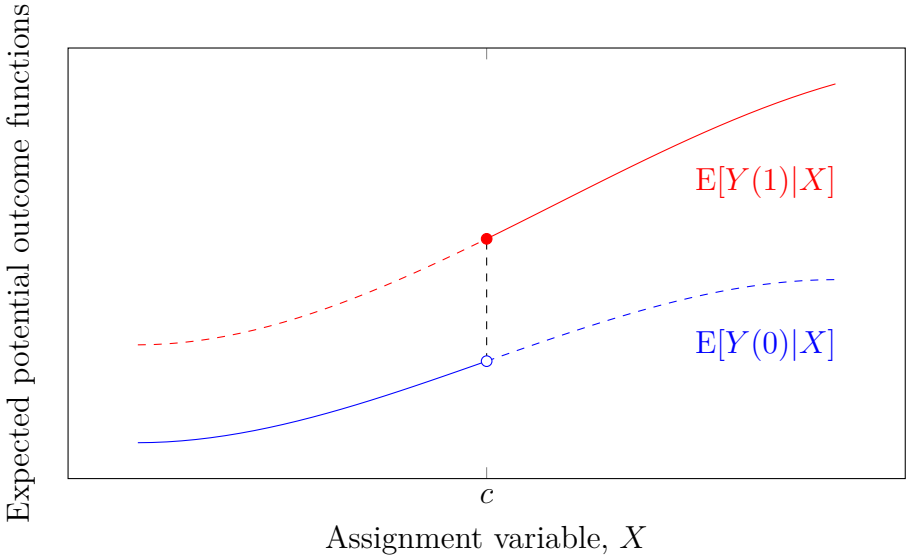


Figure 2: Expected potential outcome function under control (in blue) and treatment (in red), respectively. For units to the left of the cutoff c, we only observe control outcomes (according to the solid blue line) and no treatment outcomes (according to the dashed red line). The opposite applies to units to the right of the cutoff.

(dashed red line). All units above the cutoff ($X_i \geq c$) are treated. Consequently, outcomes under control cannot be observed (dashed blue line). Due to this lack of common support, the estimation of average (population) treatment effects, $\mathrm{E}[Y(1) - Y(0)|X = x]$, seems unfeasible. In the next section, however, we state that under relatively mild conditions the average treatment effect (ATE) at the cutoff is identified in the sharp RD design.

2.2 Fuzzy and other designs

In this thesis, we assume the basic RD setup, where all units comply with the treatment assignment (sharp design), there is one continuous assignment variable, and one cutoff. In recent years, several extensions of the basic setup have been considered. Especially, what is known as the fuzzy RD design. In the fuzzy design there is still a discontinuity in the probability of receiving treatment at the cutoff, but not from zero to one. The reason is imperfect compliance, i.e. some units assigned to control are treated, and/or some units supposed to be treated are not. To give an example, in an early RD study, Angrist and Lavy (1999) investigate the effect of class size on scholastic achievement. The authors exploit that primary schools in Israel are required to have no more than 40 pupils in a class. Hence, at an enrollment-cutoff of 41, class size should drop sharply. Some schools, however, still had classes with more than 40 pupils (e.g. due to scarcity of teachers), resulting in a fuzzy design. An illustration of the conditional probability of being treated in a fuzzy design (with two-sided non-compliance) is depicted in Figure 1b. Such compliance issues hamper the treatment effect study, but many aspects from the sharp RD analysis carry over to the fuzzy case. See, e.g., Lee and Lemieux (2010).

At this point, we just name a few other extensions. The assignment variable can be discrete and have mass points (Kolesár and Rothe, 2018). A common example is age that is only available at an annual level. RD designs can have multiple assignment variables (Papay et al., 2011), like two test scores (language and math). A special case is the Geographic RD design, where treatment assignment changes at the border separating two areas. Typically, latitude and longitude are then used for the analysis. RD designs can also have multiple cutoffs but one assignment variable (Cattaneo et al., 2016). For example, when regions choose a different cutoff to implement a federal program.

3 Identification

As described above, a fundamental problem is that we never observe treated and untreated units with the same value of the assignment variable. Near the cutoff, though, we do observe treatment and control units, respectively, that exhibit a similar value of the assignment variable. If in the absence of the treatment, the average potential outcomes would also be similar, we could compare these control and treatment units to learn about the average effect of the treatment near the cutoff. Hahn et al. (2001) were the first to present a formal identification result motivated by this kind of continuity consideration.

The key assumption to identify the sharp RD treatment effect

$$\tau_{\text{SRD}} \equiv \text{E}[Y(1) - Y(0)|X = c]$$

is that the expected potential outcome functions are continuous at the cutoff. That is, $\text{E}[Y(0)|X = x]$ and $\text{E}[Y(1)|X = x]$ are continuous at $x = c$.[1] Then,

$$\begin{aligned}\tau_{\text{SRD}} &= \lim_{x \downarrow c} \text{E}[Y(1)|X = x] - \lim_{x \uparrow c} \text{E}[Y(0)|X = x] \\ &= \lim_{x \downarrow c} \text{E}[Y|X = x] - \lim_{x \uparrow c} \text{E}[Y|X = x]\,.\end{aligned}$$

Therefore, the RD treatment effect can be estimated as the vertical distance between the two regression functions at the cutoff. Under the continuity assumption, the average outcome of the units just below the cutoff constitutes a valid counterfactual for the just treated units. We also notice that for identification no additional covariates are required. As we discuss later, however, covariates can be included to increase precision.

In Figure 2 both potential outcome functions are continuous in the assignment variable. The treatment effect τ_{SRD} is then identified as the vertical distance between the functions at the cutoff (dashed black line).

A peculiarity of the parameter τ_{SRD} is its very local nature. The sharp RD design can, in general, only recover the ATE at the cutoff. That is, for units with an assignment value of $X_i = c$. In general, the effect of the treatment is heterogeneous and varies with the assignment variable. Without further assumptions, we cannot make any statement about the ATE for units away from the cutoff. In the scenario of Figure 2, the treatment effect τ_{SRD} is indeed representative for the whole population, due to the shape of the average potential outcome functions (approximately a homogeneous ATE). In practice one would have to assume such shapes. In the past years, different approaches have been suggested to increase the limited external validity of the RD treatment effect (e.g. Angrist and Rokkanen (2015), Bertanha and Imbens (2020)).

At the end of the section, we note that the continuity-based identification of RD effects following Hahn et al. (2001) is the standard. Another framework for RD analysis assumes that the treatment is as-if randomly assigned near the cutoff. In this randomization-based framework, the RD analysis is closely connected to the analysis of experiments (e.g. randomization inference). The interested reader is referred to Lee (2008), and Sekhon and Titiunik (2017).

4 Estimation

Before turning to the main theoretical part, inference for the RD treatment effect, we discuss how to estimate the parameter τ_{SRD}, which in the previous section was identified

[1] Some researchers impose the stronger assumption that continuity holds over the full support (e.g. Imbens and Lemieux, 2008).

as the vertical distance between the regression functions at the cutoff. The assignment variable X is continuous. Therefore, the estimation of $\mathrm{E}[Y(0)|X = c]$ and $\mathrm{E}[Y(1)|X = c]$ will rely on observations further away from the cutoff. In the parametric approach, all observations are used to fit global polynomials (typically of higher order). In the nonparametric approach, only observations near the cutoff are used to fit local polynomials (typically of lower order). A challenge for either approach is that estimation occurs at a single point of interest, the cutoff, which is a boundary point. After discussing the problems of the traditional parametric approach, we present the favorable local polynomial estimation.

4.1 Parametric estimation: Global polynomials

In the parametric estimation, for all $i = 1, \ldots, n$ the model

$$Y_i = \beta_0 + \alpha T_i + \beta_1(X_i - c) + \beta_2(X_i - c)T_i + \cdots + \beta_{2p-1}(X_i - c)^p + \beta_{2p}(X_i - c)^p T_i + \varepsilon_i$$

is postulated, where p is the polynomial order. Because the assignment variable is centered at the cutoff, $\hat{\alpha}$ constitutes the estimate of τ_{SRD}. The interactions with the treatment indicator allow for a distinct influence of X on the two groups. In early applications the polynomials used to be of higher order (e.g. $p = 4$ or $p = 5$). As usual, if the functional form is incorrectly specified, τ_{SRD} is estimated with systematic bias. But it also seems unappealing to include units far away from the cutoff for point estimation at the cutoff. In a recent paper, Gelman and Imbens (2019) pointed out three major problems of higher-order polynomials in RD analyses and advise against their usage. First, implicitly a high weight is given to units far from the cutoff. Second, the estimator is sensitive to the degree of the polynomial. And third, the confidence intervals have poor coverage. For these reasons, RD estimation is now generally considered a nonparametric estimation problem.

4.2 Nonparametric estimation: Local polynomials

The most often applied estimation technique in RD analyses is local polynomial estimation. The method approximates a function by locally fitting a polynomial, where the neighborhood is chosen by a bandwidth h. Moreover, weights (based on a kernel function K) are assigned within the neighborhood, such that typically observations closer to the point of evaluation receive more weight. Adapted for the sharp RD design, the estimation of τ_{SRD} proceeds as follows. We elaborate on each step below.

1. Choose a kernel K as the weighting scheme.

2. Choose the order of the polynomial p.

3. Choose the bandwidths h_{l} and h_{r}.

4. For control observations ($X_i < c$), fit a weighted least squares regression of order p with weight $K\left(\frac{X_i - c}{h_{\mathrm{l}}}\right)$ for each observation:

$$\hat{\mu}_{\mathrm{l}} = \underset{\mu_{\mathrm{l}}}{\arg\min} \sum_{i:X_i<c} \{Y_i - \mu_{\mathrm{l}} - \mu_{\mathrm{l},1}(X_i - c) - \cdots - \mu_{\mathrm{l},p}(X_i - c)^p\}^2 K\left(\frac{X_i - c}{h_{\mathrm{l}}}\right)$$

   The estimated intercept $\hat{\mu}_{\mathrm{l}}$ constitutes the estimate of $\mathrm{E}[Y(0)|X = c]$.

5. For treatment observations ($X_i \geq c$), fit a weighted least squares regression of order p with weight $K\left(\frac{X_i - c}{h_{\mathrm{r}}}\right)$ for each observation:

$$\hat{\mu}_{\mathrm{r}} = \underset{\mu_{\mathrm{r}}}{\arg\min} \sum_{i:X_i\geq c} \{Y_i - \mu_{\mathrm{r}} - \mu_{\mathrm{r},1}(X_i - c) - \cdots - \mu_{\mathrm{r},p}(X_i - c)^p\}^2 K\left(\frac{X_i - c}{h_{\mathrm{r}}}\right)$$

   The estimated intercept $\hat{\mu}_{\mathrm{r}}$ constitutes the estimate of $\mathrm{E}[Y(1)|X = c]$.

6. The estimated RD treatment effect is $\hat{\tau}_{\mathrm{SRD}} = \hat{\mu}_{\mathrm{r}} - \hat{\mu}_{\mathrm{l}}$.

The purpose of the kernel is to ensure that observations closer to the cutoff contribute more to the estimation at the cutoff. The kernel most often applied in RD estimation is the triangular kernel, depicted in Figure A1. The weight is maximized at the cutoff and declines linearly until an observation lies outside the interval $[c - h, c + h]$. These observations receive zero weight. It can be shown (Cheng et al., 1997) that the triangular kernel is the (asymptotically) optimal kernel. Nevertheless, in practice sometimes the uniform kernel assigning equal weights is used. Another option is the Epanechnikov kernel with optimal properties for non-boundary estimation.[2] The triangular kernel is recommended, albeit the particular choice is of minor importance.

As for the order of the local polynomial, an order of zero is discouraged for the RD design. Estimation takes place at a boundary, i.e. only one-sided observations are available. It is well known that the local constant estimator (so-called Nadaraya-Watson estimator) suffers from poor boundary behavior (e.g. Hansen, 2022, Section 19.10). By contrast, the local linear estimator ($p = 1$) was shown to be boundary-adaptive (Fan, 1992), leading to a boundary bias of lower order; the same order as for estimation at interior points. In general, a small $p \neq 0$ is preferable, as the trade-off between accuracy and variability should be governed by the bandwidth. The default is local linear regression. A data-driven polynomial order selection procedure was proposed by Pei et al. (2022).

[2] See Appendix 1 for a comparison of the three kernels.

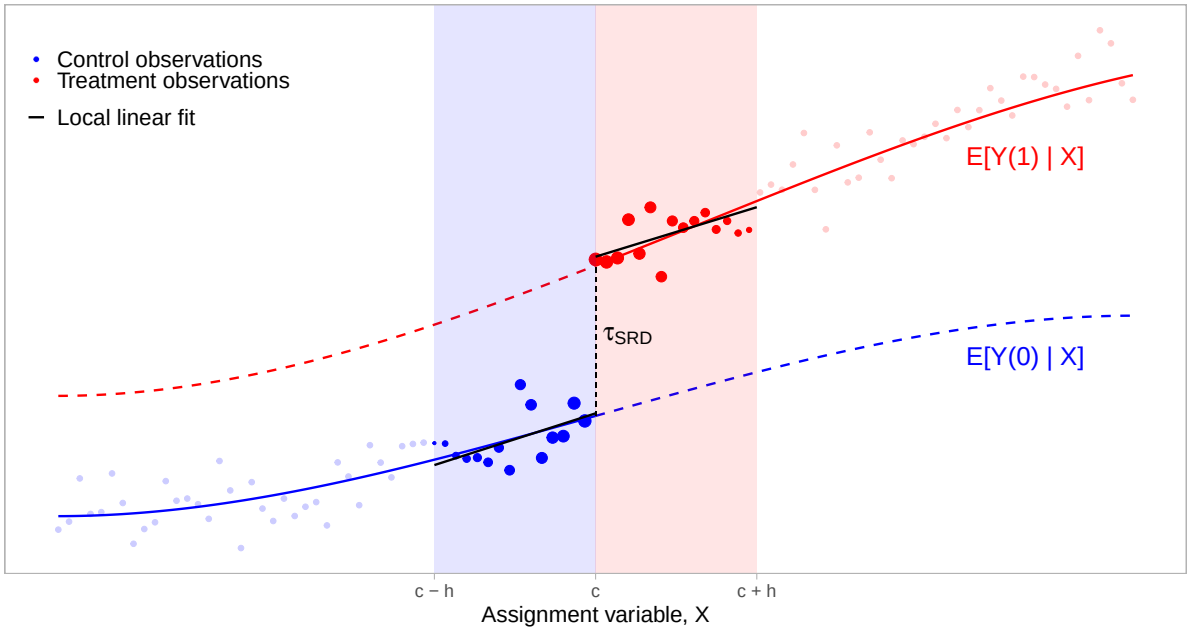


Figure 3: Local linear estimation of the RD treatment effect τ_{SRD}. Observed are random outcomes for the setting from Figure 2. The treatment effect is estimated by fitting locally a weighted linear regression, separately for control and treatment units. The estimation windows, $[c-h, c]$ (control) and $[c, c+h]$ (treatment), are determined by the common bandwidth h. The weights, as given by the triangular kernel, are represented by the size of the dots in the shaded windows.

The bandwidth h controls the size of the neighborhood around the cutoff, that is, how many observations are included for estimation. The bandwidths h_{l} and h_{r}, for estimation of $\text{E}[Y(0)|X=c]$ and $\text{E}[Y(1)|X=c]$, respectively, can differ in principle. For example, a different bandwidth is reasonable when the data suggest a different degree of curvature. In applications, often a common bandwidth $h = h_{\text{l}} = h_{\text{r}}$ is chosen. In this thesis we restrict ourselves to the choice of a common bandwidth. Bandwidth selection is a crucial and challenging step. We address it in more detail below.

All estimation steps together are illustrated in Figure 3. In the example, the triangular kernel, $p = 1$, and a common bandwidth are applied. According to steps (4) and (5), two weighted local linear regressions are fit separately (black lines). The estimate $\hat{\tau}_{\text{SRD}}$ from step (6) is the difference between these two fitted regression lines at the cutoff c, which is close to the actual τ_{SRD}.

4.3 Bandwidth selection

The bandwidth is referred to as the smoothing parameter. Heuristically, a larger bandwidth leads to more (smoothing) bias, as the lower-order polynomial approximation of the unknown regression function near the cutoff gets worse. At the same time, the variance of the estimator decreases, as more observations are used. Analogously, a smaller bandwidth generally leads to less bias, but increased variance. Therefore, the bandwidth selection involves a bias-variance trade-off. In Figure 3 both functions are well approximated by a

first-order polynomial in the respective neighborhood; bias will be small.

An ad hoc way of choosing the bandwidth is by "eyeballing". However, RD estimation and inference results are sensitive to the bandwidth choice, requiring automatic, data-dependent procedures. Two popular procedures are cross-validation (CV) and plug-in. Imbens and Lemieux (2008) proposed a version of CV specifically for the RD design, aimed at estimation at the boundary. To obtain a bandwidth with higher accuracy near the cutoff, it may be prudent to discard units far away from the cutoff before computing the CV criterion.

The plug-in method for a common bandwidth consists of deriving an explicit formula for the bandwidth that minimizes an asymptotic approximation to the mean squared error (MSE) of $\hat{\tau}_{\text{SRD}}$, and then to plug in estimators for the unknown quantities in this formula. The asymptotic or leading (conditional) MSE of $\hat{\tau}_{\text{SRD}}$ is

$$\text{AMSE}(\hat{\tau}_{\text{SRD}}) = \text{ABias}^2(\hat{\tau}_{\text{SRD}}) + \text{AVar}(\hat{\tau}_{\text{SRD}})\,.$$

For the bias term it can be shown (Cattaneo et al., 2020a, Section 4.2.2) that

$$\text{ABias}(\hat{\tau}_{\text{SRD}}) = h^{p+1}\mathcal{B}\,, \tag{1}$$

with

$$\mathcal{B} = \mathcal{B}_{\text{r}} - \mathcal{B}_{\text{l}}\,, \quad \mathcal{B}_{\text{l}} = \mu_{\text{l}}^{(p+1)} B_{\text{l}}\,, \quad \mathcal{B}_{\text{r}} = \mu_{\text{r}}^{(p+1)} B_{\text{r}}\,,$$
$$\mu_{\text{l}}^{(p+1)} \equiv \lim_{x\uparrow c} \frac{\mathrm{d}^{p+1}\,\text{E}[Y(0)|X=x]}{\mathrm{d}x^{p+1}}\,, \quad \mu_{\text{r}}^{(p+1)} \equiv \lim_{x\downarrow c} \frac{\mathrm{d}^{p+1}\,\text{E}[Y(1)|X=x]}{\mathrm{d}x^{p+1}}\,.$$

The constants B_{l} and B_{r} depend on the chosen kernel K and polynomial order p. The derivatives are related to the approximation errors of the pth-order polynomial approximations. For instance, in case of local linear regression the error is driven by the second derivative. The error increases in the curvature of the regression function. Notice that, if $\mu_{\text{l}}^{(p+1)}$ equals $\mu_{\text{r}}^{(p+1)}$, the bias term can be offset.[3] For the asymptotic (conditional) variance it can be shown that

$$\text{AVar}(\hat{\tau}_{\text{SRD}}) = \frac{1}{nh}\mathcal{V}\,, \tag{2}$$

with

$$\mathcal{V} = \mathcal{V}_{\text{l}} + \mathcal{V}_{\text{r}}\,, \quad \mathcal{V}_{\text{l}} = \frac{\sigma_{\text{l}}^2}{f_X(c-)}V\,, \quad \mathcal{V}_{\text{r}} = \frac{\sigma_{\text{r}}^2}{f_X(c+)}V\,,$$
$$\sigma_{\text{l}}^2 \equiv \lim_{x\uparrow c} \text{Var}(Y(0)|X=x)\,, \quad \sigma_{\text{r}}^2 \equiv \lim_{x\downarrow c} \text{Var}(Y(1)|X=x)\,,$$
$$f_X(c-) \equiv \lim_{x\uparrow c} f_X(x)\,, \quad f_X(c+) \equiv \lim_{x\downarrow c} f_X(x)\,.$$

The constant V depends on the chosen kernel K and polynomial order p. The presence of the density of the assignment variable f_X at the cutoff reflects that variance will be larger

[3] If p is uneven (e.g. as in local linear estimation), then $B_{\text{l}} = B_{\text{r}}$. See Calonico et al. (2014, Lemma A.1).

if fewer observations close to the cutoff are available. Notice that due to local estimation the effective sample size is nh and not n.

The bias-variance trade-off is formally expressed by the optimization problem

$$\min_{h>0} \ \mathrm{AMSE}(\hat{\tau}_{\mathrm{SRD}}) = h^{2(p+1)}\mathcal{B}^2 + \frac{1}{nh}\mathcal{V}\,.$$

A smaller bandwidth leads to reduced bias, but larger variance, and vice versa. The resulting (A)MSE-optimal bandwidth is

$$h_{\mathrm{MSE}} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2}\right)^{1/(2p+3)} \cdot n^{-1/(2p+3)}\,. \tag{3}$$

Thus, for the local linear estimator $h_{\mathrm{MSE}} \sim n^{-1/5}$. This bandwidth is not directly applicable, as several ingredients are unknown in practice and estimates need to be plugged in first. The bias estimation is more involved. Estimates of $\mu_{\mathrm{l}}^{(p+1)}$ and $\mu_{\mathrm{r}}^{(p+1)}$ are obtained by separate local polynomial regressions, with polynomial order $q \geq p+1$ and preliminary bandwidth b. This bandwidth often relies on a rule of thumb. Imbens and Kalyanaraman (2012) were the first to implement a plug-in bandwidth selector according to (3) by substituting consistent estimators, albeit for $p = 1$ only. The bandwidth selector of Calonico et al. (2014) instead applies for any p. Besides, their selector improves on the estimation of the variance component $\mathcal{V}$ and how preliminary bandwidths are chosen. Under standard assumptions, both estimators are consistent and optimal in the sense of Li (1987). Also, in distinction to h_{MSE}, both incorporate a (positive) regularization term in the denominator to adjust for potential low precision in estimating $\mu_{\mathrm{l}}^{(p+1)}$ and $\mu_{\mathrm{r}}^{(p+1)}$ (Imbens and Kalyanaraman, 2012, Section 4). The motivation is that even when substantial curvature is present, due to low precision the estimated biases may be small, resulting in a very large and poor-performing bandwidth.

5 Inference

The estimator $\hat{\tau}_{\mathrm{SRD}}$ is MSE-optimal when constructed with the MSE-optimal bandwidth (or two distinct MSE-optimal bandwidths). In the next step, we want to conduct inference for τ_{SRD}, i.e. building confidence intervals and testing hypotheses. Valid inference for the RD treatment effect, however, is non-trivial. The challenge is the present bias associated with the nonparametric estimation. Under suitable regularity conditions (Calonico et al., 2014, Lemma A.1) it can be shown that as $n \to \infty$, $h \to 0$, $nh \to \infty$,

$$\frac{\hat{\tau}_{\mathrm{SRD}} - \tau_{\mathrm{SRD}}}{\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})} - \frac{\mathrm{Bias}(\hat{\tau}_{\mathrm{SRD}})}{\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})} \xrightarrow{d} \mathcal{N}(0,1)\,, \tag{4}$$

where the standard error $\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})$ estimates the standard deviation of $\hat{\tau}_{\mathrm{SRD}}$. The asymptotic distribution of the t-statistic is not centered at zero. Hence, if bias is not negligible,

standard least squares inference (e.g. standard t-test) will be invalid. A correct null hypothesis of no treatment effect would be over-rejected. Valid inference procedures have to be aware of the (unknown) bias and incorporate it in some way. In the following, we present the three main bias-aware approaches. Their performance in practice will be assessed and compared in a simulation study as well as an application to real data.

5.1 Undersmoothing

The first approach is called undersmoothing. The idea is to employ conventional least squares inference by choosing a smaller ("undersmoothed") bandwidth relative to the MSE-optimal one. The theoretical argument is that shrinking the bandwidth at a faster rate, e.g. $h \sim n^{-\delta}$ with $\delta > 1/5$ for $p = 1$, eliminates the bias in the approximate large-sample distribution. Formally, the ratio of bias and standard error in (4) converges in probability to zero, $\text{Bias}(\hat{\tau}_{\text{SRD}})/\,\text{SE}(\hat{\tau}_{\text{SRD}}) \xrightarrow{p} 0$. An asymptotic $(1-\alpha)100\%$ confidence interval can then be constructed by

$$\text{CI}_{\text{US}} = \left[\, \hat{\tau}_{\text{SRD}} \pm |z_{\alpha/2}| \cdot \text{SE}(\hat{\tau}_{\text{SRD}}) \,\right] ,$$

where z_α denotes the α-quantile of the standard normal distribution and the subscript "US" refers to "Undersmoothing". For the standard error different suitable options are available. The usual choice is a weighted nearest neighbor or Eicker-Huber-White estimator. Alternatively an estimator based on $\text{AVar}(\hat{\tau}_{\text{SRD}})$ in (2).

There are some drawbacks with undersmoothing. There exists no clear guidance on how much to undersmooth, i.e. how much the bandwidth for constructing CI_{US} should deviate from h_{MSE}. In addition, a smaller bandwidth means that fewer observations are included and variance is increased. Lastly, bias is still existent in finite samples.

5.2 Robust bias-correction

An alternative approach is to estimate the smoothing bias and to base inference on the bias-corrected estimate. The t-statistic will be modified and we obtain bias-corrected confidence intervals, centered at the bias-corrected estimate. Compared to undersmoothing, the MSE-optimal bandwidth (and hence the same data) can be used for estimation and inference. Calonico et al. (2014) proposed what they call a robust bias-correction, which accounts for the additional variability of the bias estimation.

In a first step, the leading bias $\text{ABias}(\hat{\tau}_{\text{SRD}})$ is estimated and removed from the point estimator $\hat{\tau}_{\text{SRD}}$. From discussing bandwidth selection in the previous section, we already know how an estimator (of $\mathcal{B}$ in (1)) is constructed: The unknown derivatives are estimated by preliminary local polynomial regressions, with a higher polynomial order $q \geq p+1$ and a separate bandwidth b. The estimated bias is already available when the MSE-optimal bandwidth has been implemented for point estimation.

Standard bias-correction then would use the conventional $\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})$ for the confidence interval, thereby ignoring the introduced variability from the additional bias-correction step. This is known to translate into poor coverage in applications (e.g. Hall, 1992). In their robust bias-correction, Calonico et al. (2014) account for the additional variability by allowing the estimated bias to converge in distribution to a random variable and contribute to the distributional approximation of $\hat{\tau}_{\mathrm{SRD}}$. The new asymptotic variance is larger, and the resulting standard error $\mathrm{SE}_{\mathrm{RBC}}(\hat{\tau}_{\mathrm{SRD}})$ will be larger than $\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})$. All together, an asymptotically valid $(1-\alpha)100\%$ confidence interval for τ_{SRD} is

$$\mathrm{CI}_{\mathrm{RBC}} = \left[\left(\hat{\tau}_{\mathrm{SRD}} - \widehat{\mathrm{ABias}}(\hat{\tau}_{\mathrm{SRD}}) \right) \pm |z_{\alpha/2}| \cdot \mathrm{SE}_{\mathrm{RBC}}(\hat{\tau}_{\mathrm{SRD}}) \right],$$

where the subscript "RBC" refers to "Robust Bias-Correction". The interval is not centered at the MSE-optimal estimate $\hat{\tau}_{\mathrm{SRD}}$, instead it is centered at the bias-corrected estimate.

Building on this robust bias-correction procedure, Calonico et al. (2020) examine what would be an optimal bandwidth choice for inference (as compared to the MSE-optimal bandwidth for point estimation). The authors' objective is to minimize the coverage error (CE) of the robust bias-corrected confidence interval $\mathrm{CI}_{\mathrm{RBC}}$, i.e. the discrepancy between the empirical coverage and the nominal level. Their bandwidth h_{CE} minimizes an asymptotic approximation to the coverage error of $\mathrm{CI}_{\mathrm{RBC}}$. Consequently, when constructed with the CE-optimal bandwidth, $\mathrm{CI}_{\mathrm{RBC}}$ will be both valid and CE-optimal in large samples. Even though no closed-form solution for h_{CE} exists, it declines at a faster rate than h_{MSE} and is therefore undersmoothed: $h_{\mathrm{CE}} \sim n^{-1/(p+3)}$ compared to $h_{\mathrm{MSE}} \sim n^{-1/(2p+3)}$.

5.3 Inflated critical value

The third approach by Armstrong and Kolesár (2020) takes the potential bias of $\hat{\tau}_{\mathrm{SRD}}$ into account by using a larger critical value compared to the conventional or undersmoothing confidence interval. In the following we present the general concept, for (technical) details we refer to Armstrong and Kolesár (2020, 2018).

Let $\mathrm{E}_g[\hat{\tau}_{\mathrm{SRD}}|X_1,\ldots,X_n]$ denote the conditional expectation of $\hat{\tau}_{\mathrm{SRD}}$ when the conditional expectation function $\mathrm{E}[Y|X=x]$ is g. The authors assume that g belongs to a certain class of functions $\mathcal{G}$ (described below), and construct confidence intervals achieving asymptotically correct coverage uniformly over $\mathcal{G}$. The worst-case bias of $\hat{\tau}_{\mathrm{SRD}}$ over the function class $\mathcal{G}$ is

$$\overline{\mathrm{Bias}}(\hat{\tau}_{\mathrm{SRD}}) \equiv \sup_{g\in\mathcal{G}} \left| \mathrm{E}_g[\hat{\tau}_{\mathrm{SRD}}|X_1,\ldots,X_n] - \tau_{\mathrm{SRD}} \right|.$$

The second term in (4) is bounded in absolute value by $\overline{\mathrm{Bias}}(\hat{\tau}_{\mathrm{SRD}})/\,\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})$. Then, asymptotically, the $(1-\alpha)$-quantile of the absolute value of the t-statistic is bounded by $\mathrm{cv}_{1-\alpha}(\overline{\mathrm{Bias}}(\hat{\tau}_{\mathrm{SRD}})/\,\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}}))$, where $\mathrm{cv}_{1-\alpha}(r)$ corresponds to the $(1-\alpha)$-quantile of the

$|\mathcal{N}(r,1)|$ distribution. Thus, an asymptotically valid $(1-\alpha)100\%$ confidence interval is obtained as

$$\mathrm{CI}_{\mathrm{AK}} = \left[\hat{\tau}_{\mathrm{SRD}} \pm \mathrm{cv}_{1-\alpha} \left(\frac{\overline{\mathrm{Bias}}(\hat{\tau}_{\mathrm{SRD}})}{\mathrm{SE}(\hat{\tau}_{\mathrm{SRD}})} \right) \cdot \mathrm{SE}(\hat{\tau}_{\mathrm{SRD}}) \right] ,$$

where the subscript "AK" refers to "Armstrong and Kolesár".

Armstrong and Kolesár consider two function classes. For the description we focus on local linear estimation. Let $g_{\mathrm{l}}(x) = g(x)\mathbf{1}(x < c)$ and $g_{\mathrm{r}}(x) = g(x)\mathbf{1}(x \geq c)$ denote the part of the conditional expectation function below and above the cutoff, respectively. For the first class

$$\mathcal{G}_{\mathrm{SRD,Tay}}(M) = \{ g_{\mathrm{l}} + g_{\mathrm{r}} \,|\, g_{\mathrm{l}}, g_{\mathrm{r}} \in \mathcal{G}_{\mathrm{Tay},2}(M) \} ,$$

the function g is assumed to lie in the second-order Taylor class $\mathcal{G}_{\mathrm{Tay},2}(M)$ on either side of the cutoff. This Taylor class requires g_{l} and g_{r} to be twice differentiable in a neighborhood left and right to the cutoff, respectively, with the second derivative bounded in absolute value by M in that neighborhood. The class does not impose smoothness away from the cutoff, which might be undesirable in empirical applications. Therefore, for the alternative class

$$\mathcal{G}_{\mathrm{SRD,Höl}}(M) = \{ g_{\mathrm{l}} + g_{\mathrm{r}} \,|\, g_{\mathrm{l}}, g_{\mathrm{r}} \in \mathcal{G}_{\mathrm{Höl},2}(M) \} ,$$

the function g is assumed to lie in the second-order Hölder class $\mathcal{G}_{\mathrm{Höl},2}(M)$ on either side of the cutoff. That is, g has to be twice differentiable on either side of the cutoff, and M bounds the magnitude of the second derivative globally.[4]

The smoothness constant M needs to be specified. However, to maintain validity uniformly over the whole function class, without further restrictions M has to be set manually and cannot be data-driven. Therefore, the advice is to use problem-specific knowledge to decide what choice is reasonable a priori. Otherwise, two methods can provide guidance. First, it is possible to place a lower bound on M from the data (Kolesár and Rothe, 2018, Supplement). Second, under the restriction that the second derivative in a neighborhood of the cutoff is bounded by the maximum second derivative of a $\tilde{p}$th-order global polynomial approximation, the bound M can be chosen by a global polynomial rule of thumb: A global polynomial of order $\tilde{p}$ is fit on either side of the cutoff, and M estimated to be the largest second derivative (Armstrong and Kolesár, 2020, Supplement).

The confidence interval $\mathrm{CI}_{\mathrm{AK}}$ can be either constructed with a bandwidth that is MSE-optimal or inference-optimal. For the former case, Armstrong and Kolesár (2020) derive the bandwidth minimizing the maximum MSE over the function class $\mathcal{G}$. For the latter case, the bandwidth is optimized for length and coverage, and will be slightly over-smoothed.

[4] For the definition of $\mathcal{G}_{\mathrm{Tay},2}(M)$ and $\mathcal{G}_{\mathrm{Höl},2}(M)$ see Armstrong and Kolesár (2020, p. 16).

5.4 Including covariates

We conclude the section with covariate-adjusted RD analysis. Researchers may want to include additional covariates for two reasons. First, to increase precision and obtaining shorter confidence intervals. Second, to restore identification when control and treated units near the cutoff differ systematically, rendering the assumption of continuity of the average potential outcomes implausible. For the latter purpose, usually additional parametric assumptions are required if the estimation target remains τ_{SRD}. In the following, we only consider covariate inclusion for efficiency gains.

Let $\boldsymbol{Z}_i(0)$ and $\boldsymbol{Z}_i(1)$ be two vectors of potential covariates, where $\boldsymbol{Z}_i(0)$ contains the covariate values that would be observed under control, and $\boldsymbol{Z}_i(1)$ what would be observed under treatment. The observed covariate vector for unit i is then

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{Z}_i(0) & , X_i < c \\ \boldsymbol{Z}_i(1) & , X_i \geq c \end{cases}.$$

A natural approach is to directly include the covariates in the local polynomial regression. Calonico et al. (2019) recommend to do this linearly, additive-separably and without interacting the covariates with the treatment. That is, adding $\boldsymbol{Z}_i$ to a fully interacted weighted local regression (the latter is algebraically equivalent to the two-step estimation from above). Let $\tilde{\tau}_{\text{SRD}}$ denote the covariate-adjusted estimator. The authors show that, when the covariates are included this way, under weak regularity conditions a zero RD treatment effect on the covariates (i.e. $\text{E}[\boldsymbol{Z}_i(1) - \boldsymbol{Z}_i(0)|X_i = c] = 0$) is sufficient for $\tilde{\tau}_{\text{SRD}}$ to consistently estimate τ_{SRD}. The condition should hold, for example, for covariates determined prior to the treatment assignment.

Notice that including covariates results in general in a different bandwidth, as the optimal bandwidth formulas depend on the covariates.

6 Validation

An important part of any RD analysis is to provide evidence for its validity. Evidence that the RD design can be used to learn about the local effect of the treatment on the outcome of interest. Even though the continuity assumption on the expected potential outcome functions cannot be tested itself, we present different checks to offer indirect evidence. It should also be emphasized that the graphical illustration of the RD design is a powerful first step. A plot of the outcome against the assignment variable can reveal the presence or absence of a jump at the cutoff. In addition, one gets an impression of the shape of the underlying functions. For the sake of clarity, often local means after binning instead of all observations are displayed. To indicate the overall structure of the data, global polynomial fits (separately for control and treatment) are added. Graphical RD presentation and optimal bin selection is discussed by Calonico et al. (2015).

6.1 Manipulation of the assignment variable

When the treatment is beneficial, units being placed into control have an incentive to manipulate their value of the assignment variable. Successful manipulation can result in units barely receiving treatment being systematically different from units barely missing treatment, apart from the treatment status. If these differences affect the outcome, the units close to the cutoff cannot be compared to each other. Continuity of the average potential outcomes would be violated, and the RD design would be invalid. A prominent example is the award of a scholarship when a pupil scores above a grade threshold in a test. A certain type of parent (e.g. highly ambitious) of just-failing children might try to somehow get the score improved. Assume the question is what effect the scholarship has on later academic achievement. If these parents are successful, an RD design would likely be invalid. The reason is that after manipulation the average potential academic achievement of pupils just receiving the scholarship is likely to be higher, independently from the scholarship. Besides formal testing, administrative knowledge about the treatment-assigning process is helpful to learn about potential manipulation. In the scholarship example, for instance, information about the test design, grading process, grade threshold, how parents might appeal. In some cases, based on this knowledge manipulation can credibly be ruled out.

When the assignment variable cannot be precisely manipulated, we expect a continuous distribution. A systematic difference in the number of units near the cutoff would question the RD validity. For example, when there are unexpectedly many pupils just above the grade threshold. A carefully created histogram can reveal such a jump in the density. A more formal way is a density-continuity-test, testing whether the density of the assignment variable is continuous at the cutoff (the null hypothesis). Failing to reject provides evidence against manipulation. Any such test proceeds by estimating the density near the cutoff, separately below and above the cutoff. McCrary (2008) was the first to propose a test, Cattaneo et al. (2020b) developed a superior implementation. To illustrate, Figure A2 in the appendix shows two densities, one with and one without manipulation.

Another approach to detect manipulation is to check whether the units near the cutoff are similar with respect to covariates that could not have been affected by the treatment (e.g. predetermined covariates). For example, whether parents' education of both pupil types (score slightly below and slightly above, respectively) is similar. A systematic difference would point to manipulative behavior. A formal analysis consists in an RD analysis, replacing the outcome Y with a covariate. For each covariate, a separate analysis (i.e. bandwidth selection, treatment effect estimation, bias-aware inference) is conducted. We should not find jumps that are distinguishable from zero.

6.2 Placebo cutoffs

The key identification assumption in the sharp RD design, continuity of the average potential outcome functions at the cutoff, is inherently untestable. What can be tested instead, is whether the respective function, $\mathrm{E}[Y(0)|X = x]$ for $x < c$ and $\mathrm{E}[Y(1)|X = x]$ for $x \geq c$, is continuous away from the cutoff. To do so, we perform an RD analysis for placebo cutoffs, where the treatment status does not change. For placebo cutoffs below the true cutoff, only control observations are used. For placebo cutoffs above, only treated observations. Imbens and Lemieux (2008) suggest to choose the two median values. If we find jumps away from the cutoff, that cannot be explained plausibly, this would doubt the interpretation of a jump at the real cutoff as the treatment effect.

6.3 Bandwidth sensitivity and donut holes

Finally, it is always advisable to check how sensitive the results are to the bandwidth choice. It is reassuring if the findings are stable for different neighborhoods. A plot of the estimated treatment effect together with its robust confidence interval over a range of bandwidths is typically insightful. We expect that, in line with the bias-variance trade-off, for bandwidths larger than the MSE-optimal one the confidence intervals will get shorter but displaced.

Another sensitivity check is to exclude some of the closest observations to the cutoff from the RD analysis, known as the donut-hole-approach (Barreca et al., 2011). At first sight, it seems illogical to exclude the observations that are usually the most informative. The motivation is that if manipulation has taken place, likely the units closest to the cutoff are involved. In practice, estimation and inference results for different sizes of the donut hole should be compared.

7 Application

To illustrate all the discussed steps in a bias-aware sharp RD analysis (identification, estimation, inference, validation), we present an extensive application before the simulation study, which then investigates inferential performance exclusively. For the implementation of the application and simulation, we particularly make use of the two R packages `rdrobust` by Calonico et al. and `RDHonest` by Armstrong and Kolesár. All materials (data, code, etc.) to replicate the results are available on GitHub: `https://github.com/svjaco/master_thesis`

Our application is based on a prominent paper by Ludwig and Miller (2007, "Does Head Start improve children's life chances? Evidence from a regression discontinuity design"). In 1965 the U.S. launched the federal poverty alleviation program "Head Start", targeted at poor children. To receive funding, counties had to apply to a competitive selection process. However, the federal government provided assistance to the 300 poorest counties

Table 1: Overview Head Start data set

Assignment variable (X)	County poverty rate in 1960	Min: 15.21 | Mean: 36.73 | Max: 81.57
Outcome (Y)	Average mortality rate per 100000, Head Start related causes, Ages 5–9, 1973–83	Min: 0 | Mean: 2.17 | Max: 73.53
Cutoff (c)	$X = 59.1984$	Poverty rate of 300th poorest county
Treatment (T)	Grant-writing assistance	
Additional covariates	14 (post: 4, prior: 10)	See Table A2
Number of observations	2781	Control: 2487, Treatment: 294

to develop funding proposals. Ludwig and Miller (2007, e.g. Figure I and II) document that this grant-writing assistance substantially increased Head Start participation and funding. For example, 80 percent of the 300 treated counties received funding. In their main analysis, Ludwig and Miller used a sharp RD design to determine the local effect of grant-writing assistance on child mortality due to causes the program's goal was to reduce.[5] For estimation local linear regression is applied, but the bandwidth is chosen ad hoc, and their inference procedure ignores any potential bias. Therefore, we review the original analysis and conduct validation checks, optimal bandwidth selection and bias-aware inference, which have been developed after the paper's publication.

Table 1 provides a summary. The assignment variable is the county poverty rate, obtained from the 1960 U.S. Census. The outcome is the average mortality rate per 100000 for children of age 5–9, during 1973–83, due to Head Start related causes (mainly anemia, meningitis and respiratory problems). A county receives the treatment of grant-writing assistance when its poverty rate is above the cutoff 59.1984. The raw data are plotted in Figure A3. We have data on 2781 counties, with 294 of them above the poverty cutoff, receiving treatment. We obtain more insights from the standard RD plot in Figure 4. The mortality rate increases nearly linearly in the poverty rate until the cutoff. At the cutoff, the plot suggests a clearly visible downward discontinuity. Afterwards, the mortality seems to rise nonlinearly. Such a visualization provides valuable guidance, but can never replace a formal statistical treatment effect analysis (remember the above mentioned problems of higher-order polynomial fitting).

The RD treatment effect τ_{SRD} in the Head Start application is the ATE of receiving grant-writing assistance on the Head Start (HS) related child mortality for a county with poverty rate in 1960 equal to 59.1984, which reflects a very high poverty. This effect can be identified with a sharp RD design if the continuity assumption holds. That is, continuity (at the cutoff) of the expected potential mortality functions (with and without grant-writing assistance). We have to be aware of two threats. First, the cutoff might have been used in other federal social spending programs. And second, the poverty rate might have been manipulated. Ludwig and Miller addressed the first concern and did not find any evidence for a discontinuity in other federal social spending at the Head Start cutoff. Also, manipulation of the assignment variable can credibly been ruled out with the

[5] Alternatively, we can think about the effect as an intent-to-treat effect of Head Start funding.

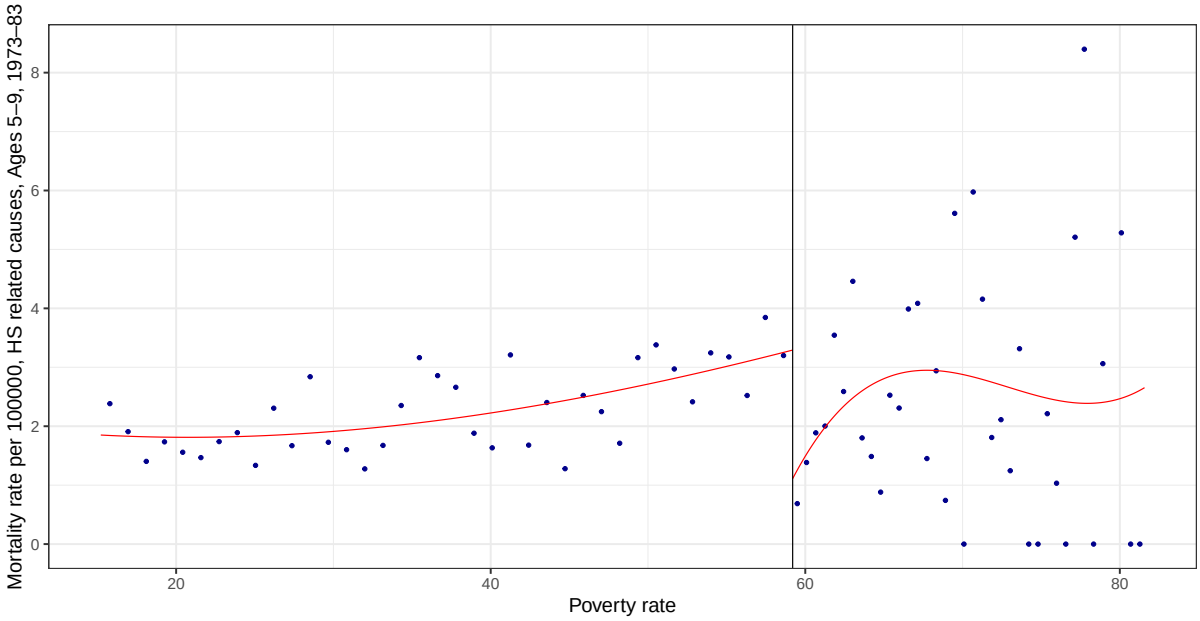


Figure 4: Standard RD plot for the outcome of interest. The dots depict the mean within each bin, where the bins are evenly spaced and their number is chosen to mimic the variance of the raw data (see Calonico et al., 2015). The solid red lines are third-order polynomial fits for control (left) and treated (right) counties separately.

knowledge that treatment was assigned in 1965 on the basis of official census information from 1960. Nevertheless, we conduct the falsification tests from Section 6.

The histogram of the poverty rate in Figure A4 does not reveal any indication of sorting around the cutoff. This is supported by a formal density-continuity-test, which clearly does not reject the null of continuity at the cutoff, as illustrated in Figure 5. Next, we check whether statistically significant discontinuities can be found in variables where no treatment effect is expected. For example, for other causes of death (mainly injuries) and the outcome of interest during the years 1959–64, before Head Start started. For both variables we show the RD plot in Figure A5. The plot for the pre-treatment HS related mortality does show a noticeable jump that requires further investigation. Therefore, we perform a robust bias-corrected RD analysis and can debilitate the concern (Figure A6b combined with a robust p-value of 0.509 in Table A2). The size of the jump in the RD plot is driven by the erratic behavior of the third-order polynomial at the cutoff. Table A2 reports the results of a separate RD analysis (local linear estimation, triangular kernel, MSE-optimal bandwidth, RBC inference) for each of the additional available covariates in the Head Start data set. All 95% robust confidence intervals contain zero. For convenience the results are presented for the MSE-optimal instead of the CE-optimal bandwidth, which might be a more suitable choice, as we are mostly interested in inference here. The conclusions, however, remain the same. For completeness, we also report results from an RD analysis for placebo cutoffs (Table A3) and the donut-hole-approach (Table A4), as described at the end of Section 6. As a benchmark the results for $\hat{\tau}_{\text{SRD}}$ (i.e. effect at the true cutoff without excluded counties at the cutoff) are already included. As expected,

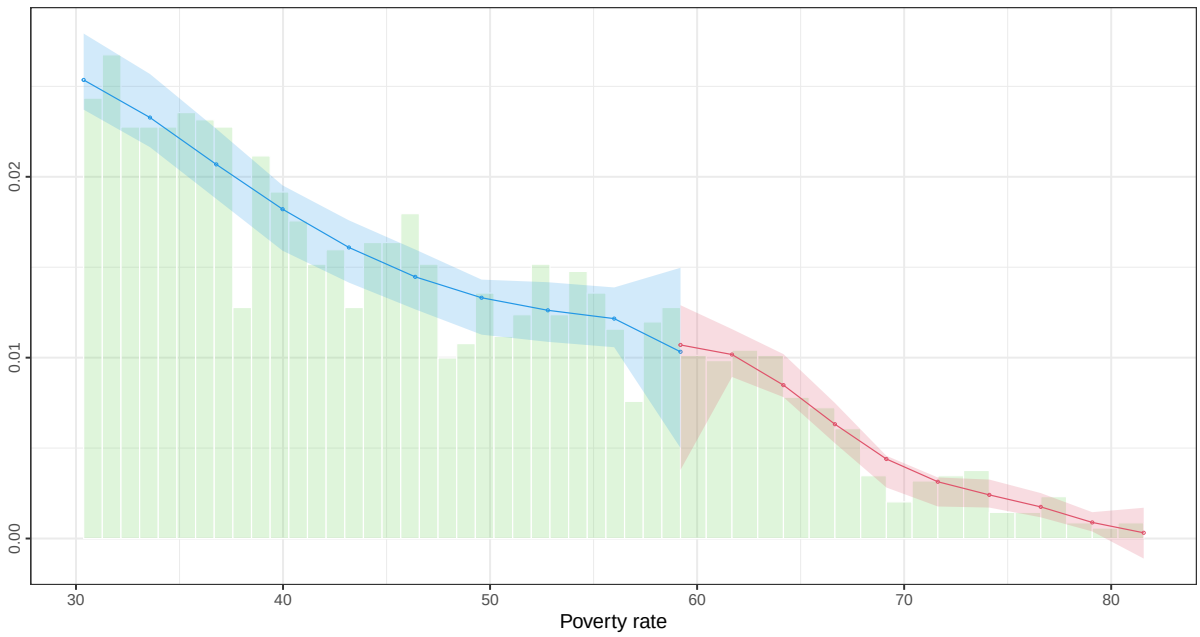


Figure 5: Density-continuity-test for the poverty rate (assignment variable). The density is estimated separately below (blue) and above (red) the cutoff, using the local polynomial density estimator proposed by Cattaneo et al. (2020b). The point estimates are non-centered, as the 95% confidence intervals are robust bias-corrected. A histogram is added to the background.

no significant effect at the artificial cutoffs occurs. The donut-hole-analysis reveals that the results are somewhat sensitive to the degree of extrapolation at the cutoff. Overall, the presented checks support the validity of the RD design.

Let us now turn to the estimation of and inference for the effect of grant-writing assistance on the child mortality (per 100000, ages 5–9, period 1973–83) due to HS related causes. Our results are obtained for local linear estimation and the triangular kernel. We consider the three bias-aware inference approaches from Section 5: undersmoothing, robust bias-correction, inflated critical value. Standard errors are obtained via the nearest neighbor method. The results are collected in Table 2 and visualized in Figure 6. For the sake of comparison, the original findings from Ludwig and Miller (2007, Table III) are reported as well. Their bandwidths of $h_1^{\text{LM}} = 9$ and $h_2^{\text{LM}} = 18$ are ad hoc choices and their inference procedure (bootstrapping the t-statistic) ignores potential bias. Moreover, due to an error in the code, the applied kernel is the uniform kernel, and not (as stated) the triangular. For the bandwidth of 9, Ludwig and Miller report an estimate of -1.895, significant at the 5% level.

We first compare with the RBC approach. The estimated MSE-optimal bandwidth (using the plug-in selector of Calonico et al., 2014) is $\hat{h}_{\text{MSE}} = 6.913$, leading to an estimated effect of -2.389 with robust p-value of 0.043. Since the estimated untreated mortality at the cutoff for $\hat{h}_{\text{MSE}}$ is 3.585, this is a large reduction of more than 60%. The MSE-optimal bandwidth suggests that the bandwidths of Ludwig and Miller are too large ("oversmoothed"). Including the nine predetermined variables from the U.S. Census 1960

Table 2: Bias-aware analysis for the RD treatment effect in the Head Start application

Approach	Bandwidth		RD estimate	Inference	
				95% CI	p-value
Undersmoothing					
– By factor 2	$h^{\text{US}}_{1/2}$	3.457	−3.428	[−5.856, −1.001]	0.006
– By factor 3	$h^{\text{US}}_{1/3}$	2.304	−2.590	[−5.488, 0.307]	0.080
Robust bias-correction					
– MSE-optimal	$\hat{h}_{\text{MSE}}$	6.913	−2.389	[−5.426, −0.083]	0.043
– CE-optimal	$\hat{h}_{\text{CE}}$	4.650	−3.248	[−6.092, −0.749]	0.012
– MSE-optimal, incl. covariates	$\tilde{h}_{\text{MSE}}$	7.115	−2.445	[−5.155, −0.339]	0.025
– Ad hoc, Ludwig and Miller	h^{LM}_1	9.000	−2.182	[−5.722, −0.350]	0.027
– Ad hoc, Ludwig and Miller	h^{LM}_2	18.000	−1.681	[−4.564, −0.101]	0.040
Armstrong and Kolesár					
– minimax MSE	$h^{\text{AK}}_{\text{MSE}}$	4.551	−3.283	[−6.039, −0.526]	0.019
– Length-optimal	$h^{\text{AK}}_{\text{CIL}}$	4.670	−3.239	[−6.022, −0.456]	0.022
– Ad hoc, Ludwig and Miller	h^{LM}_1	9.000	−2.182	[−6.229, 1.866]	0.520
– Ad hoc, Ludwig and Miller	h^{LM}_2	18.000	−1.681	[−11.700, 8.337]	>0.9
Ludwig and Miller (bias ignored, uniform kernel)					
– Ad hoc, Ludwig and Miller	h^{LM}_1	9.000	−1.895		0.036
– Ad hoc, Ludwig and Miller	h^{LM}_2	18.000	−1.198		0.081

Note: Results are based on local linear estimation and the triangular kernel. The included additional covariates are all the variables from the Census 1960, as described in Table A2. For the approach by Armstrong and Kolesár, the Hölder class and the global polynomial ROT ($M_{\text{ROT}} = 0.299$) are used. Due to an error in their code, the results from Ludwig and Miller (2007, Table III) are obtained for the uniform kernel, and not (as reported) for the triangular. Their p-values are t-statistic-bootstrapped, ignoring potential bias.

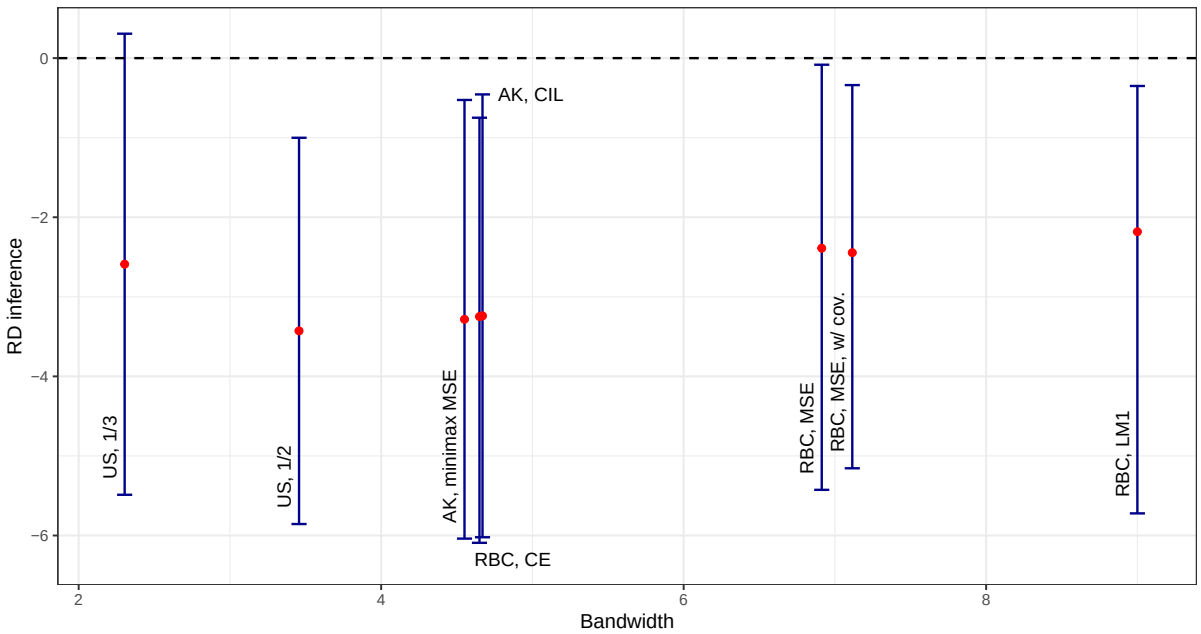


Figure 6: Main results for the RD treatment effect in the Head Start application: Point estimates and 95% confidence intervals for the bias-aware inference approaches, as presented in Table 2.

(Table A2) showcases the potential gains from incorporating additional covariates. The point estimate remains essentially the same, but the confidence interval is about 10% tighter. The CE-optimal bandwidth $\hat{h}_{\text{CE}} = 4.650$ is smaller and we obtain a larger (in magnitude) effect of -3.248. To perform undersmoothing, we choose a bandwidth that is one half and one third of the MSE-optimal bandwidth, respectively. Undersmoothing by 50% yields the largest discontinuity out of all specifications (-3.428) and a relatively narrow confidence interval. Estimating the bias for this undersmoothed bandwidth, however, indicates that bias is not negligible, and the inference is too optimistic. In contrast, for $h^{\text{US}}_{1/3}$ we get an estimate similar to the MSE-optimal one with a relatively wide confidence interval, containing as the only one zero. The third approach by Armstrong and Kolesár requires the specification of the function class $\mathcal{G}$ and smoothness constant M (Section 5.3). We choose the Hölder class (imposing smoothness away from the cutoff) and the global polynomial rule of thumb to bound the second derivative of the conditional expectation function globally. For these choices, the minimax bandwidth $h^{\text{AK}}_{\text{MSE}}$, minimizing the maximum MSE of the local linear estimator over the specified class, is 4.551 and thus 34% smaller than $\hat{h}_{\text{MSE}}$. The resulting RD estimate and confidence interval are close to the CE-optimal RBC results. When the criterion is to optimize interval length while maintaining coverage, the bandwidth $h^{\text{AK}}_{\text{CIL}}$ is only slightly increased.

We conclude the following from the above presented. Our results indicate that the bandwidths in the original paper are too large. To theoretically justify these bandwidths within the approach of Armstrong and Kolesár, the conditional expectation function has to be nearly linear (i.e. M has to be substantially smaller). This is not consistent with the data, when looking at the evolution of the average mortality rate in the treatment group in Figure 4. Both, using the uniform kernel (estimate of -1.895 vs. -2.182 for the triangular kernel) and the oversmoothing contribute to an understatement of the effect size. Our bias-aware analysis yields larger confidence intervals (due to smaller bandwidths plus bias-awareness), but at the same time larger treatment effects, leaving the original conclusion unchanged. In summary, our analysis reaffirms the HS targeted mortality reducing effect of grant-writing assistance at the cutoff.

The application already indicates that how much to undersmooth is a delicate question, and we might easily run into undercoverage if the bandwidth is still "too large". This is one of the points addressed in the upcoming simulation.

8 Simulation

To study the finite-sample performance of the asymptotically valid confidence intervals from Section 5, we conduct a Monte Carlo experiment. We compare empirical coverage and average interval length for three realistic conditional expectation functions.

8.1 Setup

The setup for the simulation study is the same as in Calonico et al. (2014), which itself is based on the data generating process in Imbens and Kalyanaraman (2012). For each replication the data are generated as $n = 500$ i.i.d. draws, $i = 1, \ldots, n$, from the model

$$\begin{aligned} Y_i &= \mu_j(X_i) + \varepsilon_i \,, \quad j = 1, 2, 3 \,, \\ X_i &\sim 2\,\text{Beta}(2, 4) - 1 \,, \\ \varepsilon_i &\sim \mathcal{N}(0, 0.1295^2) \,, \end{aligned}$$

where $\text{Beta}(\alpha, \beta)$ denotes the beta distribution with shape parameters α and β, and the index j specifies the functional form of the conditional expectation function. Three models are considered (Figure 7), with the following functional forms:

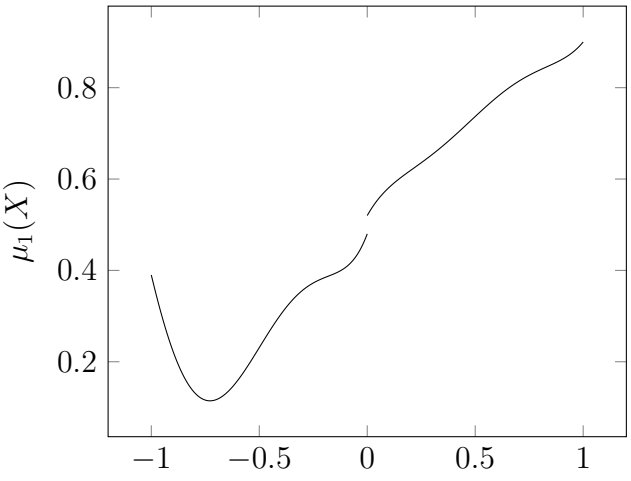
$$\mu_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & , x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & , x \geq 0 \end{cases};$$

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

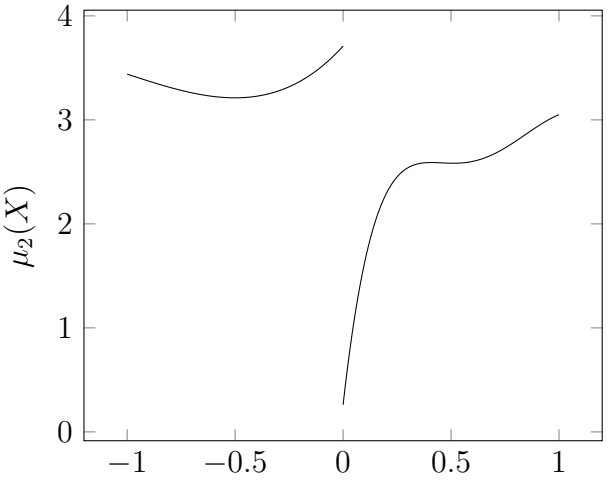
$$\mu_3(x) = \begin{cases} 0.48 + 1.27x - 3.59x^2 + 14.147x^3 + 23.694x^4 + 10.995x^5 & , x < 0 \\ 0.52 + 0.84x - 0.30x^2 - 2.397x^3 - 0.901x^4 + 3.56x^5 & , x \geq 0 \end{cases}.$$

The first function $\mu_1(x)$ is obtained by fitting a fifth-order global polynomial with different coefficients for $X_i < 0$ and $X_i \geq 0$ to the data of Lee (2008, "Randomized experiments from non-random selection in U.S. House elections"). Lee used an RD design to study the incumbency advantage in U.S. House elections. The second function $\mu_2(x)$ is obtained in the same way for the Head Start data (Ludwig and Miller, 2007), known from our application. The assignment variable is rescaled to shift the cutoff to zero. Notice that the population RD treatment effect according to μ_2 is -3.45, larger in magnitude than estimated in the application above. Lastly, the specification $\mu_3(x)$ is obtained by altering some of the coefficients in Model 1, in order to increase the overall curvature. For convenience, the probability density function of $\text{Beta}(2, 4)$ is plotted in Figure 8. On average, 18.75% of the units will be in the treatment group (as $\text{P}(Z \geq 0.5) = 0.1875$ for $Z \sim \text{Beta}(2, 4)$). In the Head Start application the share of treated counties was 10.57%.

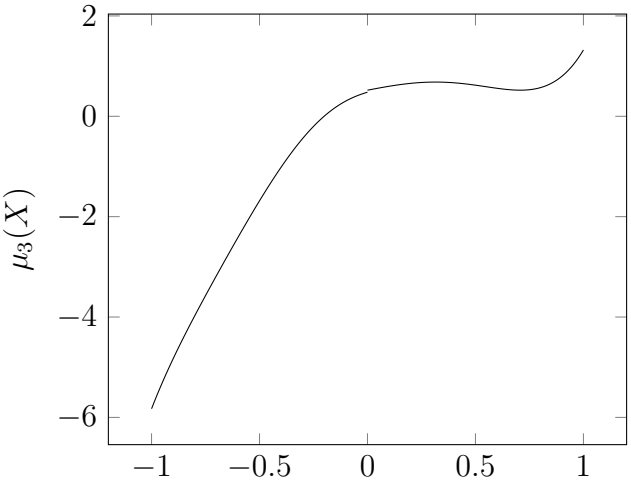
For each model and bias-aware inference procedure we report the average bandwidth, approximated bias and root MSE, empirical coverage and average interval length across 5000 replications. We consider the following eight specifications: Undersmoothing (ad hoc) by dividing the estimated MSE-optimal bandwidth by two and three, robust bias-correction with bandwidths $\hat{h}_{\text{MSE}}$ and $\hat{h}_{\text{CE}}$, and the approach by Armstrong and Kolesár (2020) with the minimax MSE-optimal and length-optimal bandwidth (respectively for the rule of thumb choice of the smoothness constant M_{ROT} and the truth $\overline{M}$). In all



(a) Model 1 – Lee (2008)



(b) Model 2 – Ludwig and Miller (2007)



(c) Model 3

Figure 7: Conditional expectation functions for the simulation study

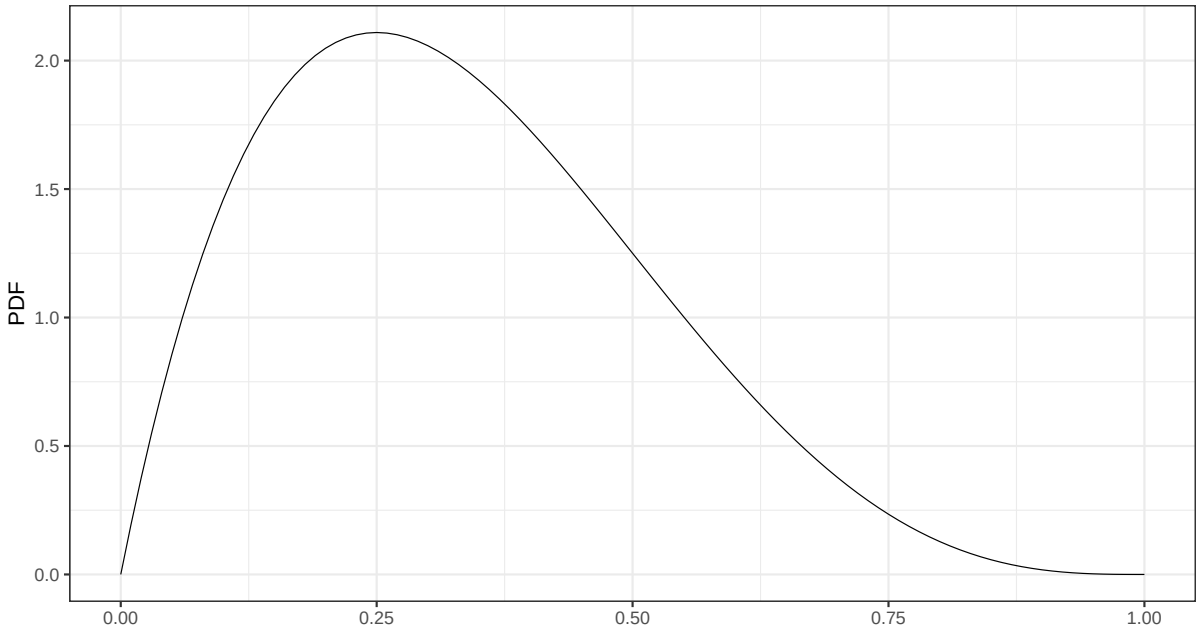


Figure 8: Probability density function of the beta distribution Beta(α, β) with shape parameters $\alpha = 2$ and $\beta = 4$

cases, we conduct local linear estimation with the triangular kernel. Standard errors are obtained by heteroskedasticity-robust nearest neighbor (with three as the minimum number of neighbors). For reference, we also report the infeasible MSE-optimal bandwidth h_{MSE}, which can be derived because of the known data generating process. All confidence intervals are computed at the nominal 95% coverage level.

8.2 Results and discussion

The simulation results are collected in Table 3, Table 4 and Table 5. We begin with Model 1 (Lee, 2008). Due to the overall linear shape of μ_1, the true MSE-optimal bandwidth $h_{\text{MSE}} = 0.166$ is relatively large. The estimated version $\hat{h}_{\text{MSE}} = 0.196$ is larger, and results for RBC in some undercoverage (91.70%). The CE-optimal bandwidth is substantially smaller and coverage is slightly improved. For undersmoothing, bias can be reduced and coverage is similar to RBC with $\hat{h}_{\text{CE}}$, but with wider confidence intervals (the cost of increased variance). The results for the minimax MSE bandwidth for the correctly specified $\overline{M}$ are very close to the CE-optimal RBC results in all aspects, except for improved coverage of 95.38%, almost exactly the nominal level. Since the rule of thumb (ROT) calibrated choice is larger than $\overline{M}$, the corresponding bandwidth is smaller (as the maximal bias is larger). The coverage is even better, at the expense of a slightly increased average interval length. Results for the length-optimal bandwidth selector are essentially the same. On balance, undersmoothing and RBC exhibit some degree of undercoverage, with undersmoothing having on average the widest intervals. In contrast, all specifications of the procedure by Armstrong and Kolesár achieve accurate coverage.

Table 3: Simulation results for Model 1 (Lee, 2008) and nominal coverage of 95%

Approach	Bandwidth	Bias	RMSE	CI coverage [%]	CI length
Undersmoothing (ad hoc)					
– By factor 2	0.098	0.005	0.082	92.60	0.297
– By factor 3	0.065	0.001	0.101	92.24	0.370
Robust bias-correction					
– MSE-optimal	0.196	0.019	0.064	91.70	0.246
– CE-optimal	0.144	0.013	0.071	92.44	0.266
Armstrong and Kolesár					
– minimax MSE (M_{ROT})	0.126	0.011	0.069	95.12	0.289
– minimax MSE ($\overline{M}$)	0.146	0.014	0.063	95.38	0.266
– Length-optimal (M_{ROT})	0.130	0.011	0.068	95.50	0.289
– Length-optimal ($\overline{M}$)	0.150	0.015	0.062	95.66	0.266

Note: The columns are average bandwidth, approximated bias and root MSE, empirical coverage and average interval length across the 5000 replications. The infeasible MSE-optimal bandwidth is $h_{\text{MSE}} = 0.166$. Results are based on local linear estimation and the triangular kernel. For the approach by Armstrong and Kolesár, the Hölder class is used with M as either the rule of thumb M_{ROT} (= 22.399 on average) or the truth $\overline{M} = 14.36$.

For the second model (Ludwig and Miller, 2007) the infeasible MSE-optimal bandwidth is $h_{\text{MSE}} = 0.082$, which is half the size compared to Model 1, as expected from looking at the plots in Figure 7. Again, the estimate $\hat{h}_{\text{MSE}} = 0.1$ is too high ($\approx 20\%$) and the RBC method undercovers by two percentage points. Unfortunately, undersmoothing led to some numerical instability during the simulations for Model 2. The number of successful replications was too low to obtain reliable results. The results for the four specifications of Armstrong and Kolesár are similar to each other, as the ROT estimate of M is close to the true value. The empirical coverage is by one to two percentage points above the nominal level. Compared to RBC, we obtain longer confidence intervals, because the relatively high curvature of μ_2 is taken into account by accordingly increased critical values.

The results for Model 3 in Table 5 are similar to those for Model 1 in Table 3. The increased overall curvature translates into somewhat smaller bandwidths. Now, empirical coverage for the ad hoc undersmoothing is slightly worse, while it is slightly improved for RBC. The higher degree of curvature means that the bound on the second derivative M is enlarged, widening on average the Armstrong and Kolesár confidence intervals. Interestingly, for Model 3 M_{ROT} is considerably below $\overline{M}$ and similar to M_{ROT} for Model 1. This bias of M_{ROT}, however, does not have a relevant impact on coverage (but tightens the intervals).

We summarize the Monte Carlo study as follows. Undersmoothing achieves a decent coverage, but overall offers the worst combination of coverage and average interval length. Robust bias-correction undercovers the true treatment effect slightly (similar to undersmoothing), but yields the shortest intervals. The CE-optimal bandwidth choice offers some small refinements in empirical coverage relative to the MSE-optimal bandwidth se-

Table 4: Simulation results for Model 2 (Ludwig and Miller, 2007) and nominal coverage of 95%

Approach	Bandwidth	Bias	RMSE	CI coverage [%]	CI length
Undersmoothing (ad hoc)					
– By factor 2	0.050	NA	NA	NA	NA
– By factor 3	0.033	NA	NA	NA	NA
Robust bias-correction					
– MSE-optimal	0.100	0.048	0.098	92.98	0.345
– CE-optimal	0.073	0.027	0.100	93.02	0.396
Armstrong and Kolesár					
– minimax MSE (M_{ROT})	0.077	0.032	0.095	96.08	0.440
– minimax MSE ($\overline{M}$)	0.075	0.031	0.095	96.62	0.447
– Length-optimal (M_{ROT})	0.080	0.035	0.095	96.16	0.442
– Length-optimal ($\overline{M}$)	0.078	0.033	0.094	96.80	0.448

Note: The columns are average bandwidth, approximated bias and root MSE, empirical coverage and average interval length across the 5000 replications. The infeasible MSE-optimal bandwidth is $h_{\text{MSE}} = 0.082$. Results are based on local linear estimation and the triangular kernel. For the approach by Armstrong and Kolesár, the Hölder class is used with M as either the rule of thumb M_{ROT} (= 103.418 on average) or the truth $\overline{M} = 109.62$. Undersmoothing led to numerical instability, wherefore results are not available.

Table 5: Simulation results for Model 3 and nominal coverage of 95%

Approach	Bandwidth	Bias	RMSE	CI coverage [%]	CI length
Undersmoothing (ad hoc)					
– By factor 2	0.087	−0.004	0.086	91.90	0.314
– By factor 3	0.058	−0.003	0.107	91.76	0.393
Robust bias-correction					
– MSE-optimal	0.173	−0.013	0.065	92.94	0.252
– CE-optimal	0.127	−0.006	0.072	92.96	0.275
Armstrong and Kolesár					
– minimax MSE (M_{ROT})	0.119	−0.007	0.071	95.94	0.301
– minimax MSE ($\overline{M}$)	0.095	−0.005	0.079	95.64	0.339
– Length-optimal (M_{ROT})	0.122	−0.007	0.070	96.14	0.301
– Length-optimal ($\overline{M}$)	0.098	−0.005	0.077	96.08	0.339

Note: The columns are average bandwidth, approximated bias and root MSE, empirical coverage and average interval length across the 5000 replications. The infeasible MSE-optimal bandwidth is $h_{\text{MSE}} = 0.260$. Results are based on local linear estimation and the triangular kernel. For the approach by Armstrong and Kolesár, the Hölder class is used with M as either the rule of thumb M_{ROT} (= 26.629 on average) or the truth $\overline{M} = 45.406$.

lector. The best coverage is obtained for the procedure of Armstrong and Kolesár. The global polynomial rule of thumb to choose the required smoothness constant M works reasonably well. Still, the necessity to specify this constant to bound the second derivative on both sides of the cutoff constitutes a drawback. Each method, however, imposes (explicitly or implicitly) restrictions on the smoothness of the conditional expectation function. Estimating the bias in RBC via local quadratic regression assumes that (in a neighborhood around the cutoff) the regression functions are three times continuously differentiable. The Monte Carlo experiment suggests to report in practice the confidence intervals of Armstrong and Kolesár, $\mathrm{CI}_{\mathrm{AK}}(h^{\mathrm{AK}}_{\mathrm{MSE}}, M_{\mathrm{ROT}})$ and $\mathrm{CI}_{\mathrm{AK}}(h^{\mathrm{AK}}_{\mathrm{CIL}}, M_{\mathrm{ROT}})$, together with the robust bias-corrected confidence intervals, $\mathrm{CI}_{\mathrm{RBC}}(\hat{h}_{\mathrm{MSE}})$ and $\mathrm{CI}_{\mathrm{RBC}}(\hat{h}_{\mathrm{CE}})$. In doing so, we can likely complement a more accurate but wider interval with a tighter but less accurate interval. The sensitivity of the confidence intervals to the choice of the bandwidth and smoothing constant can then be further examined.

9 Conclusion

This thesis provided an overview and illustration of modern (sharp) RD analysis that takes potential bias into account to obtain valid inference for the nonparametrically estimated RD treatment effect. We discussed identification, estimation, validation and presented the three main bias-aware inference approaches: undersmoothing, robust bias-correction (Calonico et al., 2014), and inflated critical values (Armstrong and Kolesár, 2020). We reanalyzed an older prominent RD study by Ludwig and Miller (2007), particularly by conducting bias-aware inference, which provided support for the original findings. In a Monte Carlo study we assessed and compared the finite-sample performance of the asymptotically valid inference procedures. Based on the insights from all the parts, we discourage the application of ad hoc undersmoothing, especially as the main approach to achieve valid bias-aware inference. Instead, we recommend to report results for the two other approaches with the corresponding optimal bandwidth choices, and the rule of thumb M_{ROT} if no problem-specific knowledge for the choice is available. If necessary, further sensitivity checks for the confidence intervals can be conducted.

The Monte Carlo experiment can be extended in several directions. To ensure that the results are not sensitive to the chosen data generating process, we can alter the distribution of the assignment variable (e.g. to a uniform distribution), and the distribution of the error (e.g. to a log-normal distribution). Furthermore, we can consider heteroskedastic errors, a different error variance, and a different sample size. A natural extension of this thesis is bias-aware inference in fuzzy regression discontinuity designs.

Appendix

Appendix 1: Kernels

Table A1: Prominent kernel functions

Kernel	Function	Support
Triangular	$K_{\mathrm{T}}(u) = 1 - \lvert u \rvert$	$[-1, 1]$
Uniform	$K_{\mathrm{U}}(u) = 0.5$	$[-1, 1]$
Epanechnikov	$K_{\mathrm{E}}(u) = 0.75(1 - u^2)$	$[-1, 1]$

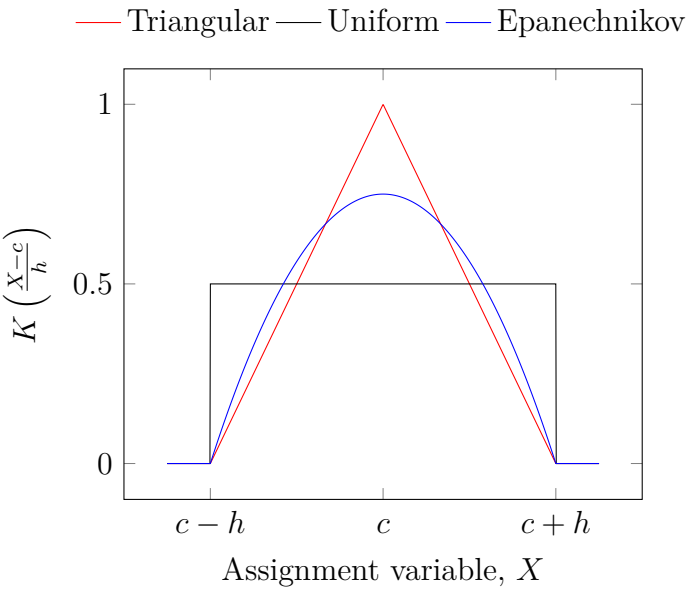


Figure A1: Commonly used kernels as given in Table A1, applied to the RD design

Appendix 2: Manipulation of the assignment variable

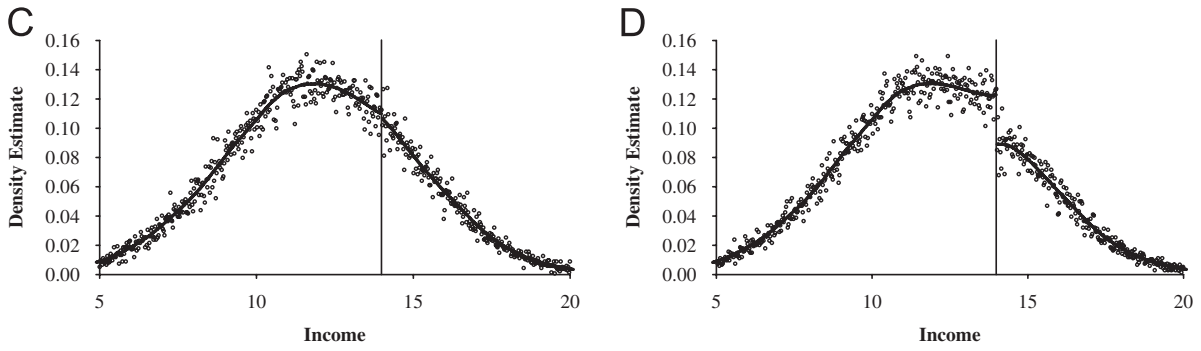


Figure A2: Hypothetical example of an income-tested job training program. (C) density of income without manipulation; (D) density of income with manipulation. From McCrary (2008, Fig. 2).

Appendix 3: Application

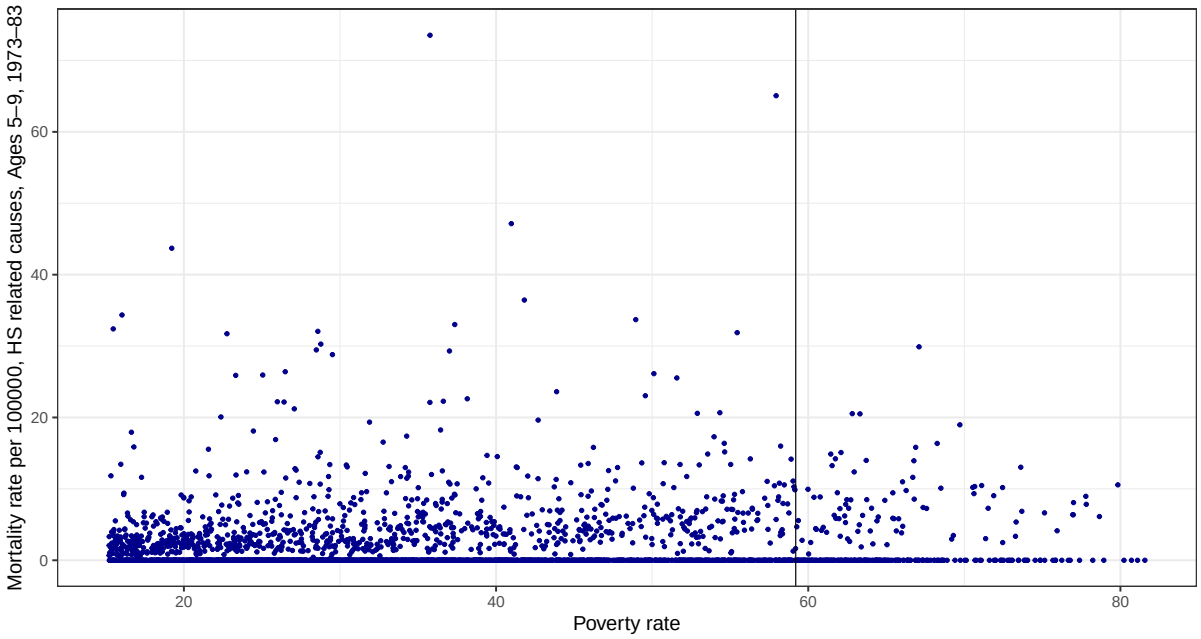


Figure A3: Scatter plot of the raw data underlying the Head Start RD analysis. The cutoff of 59.1984 is indicated by the vertical line.

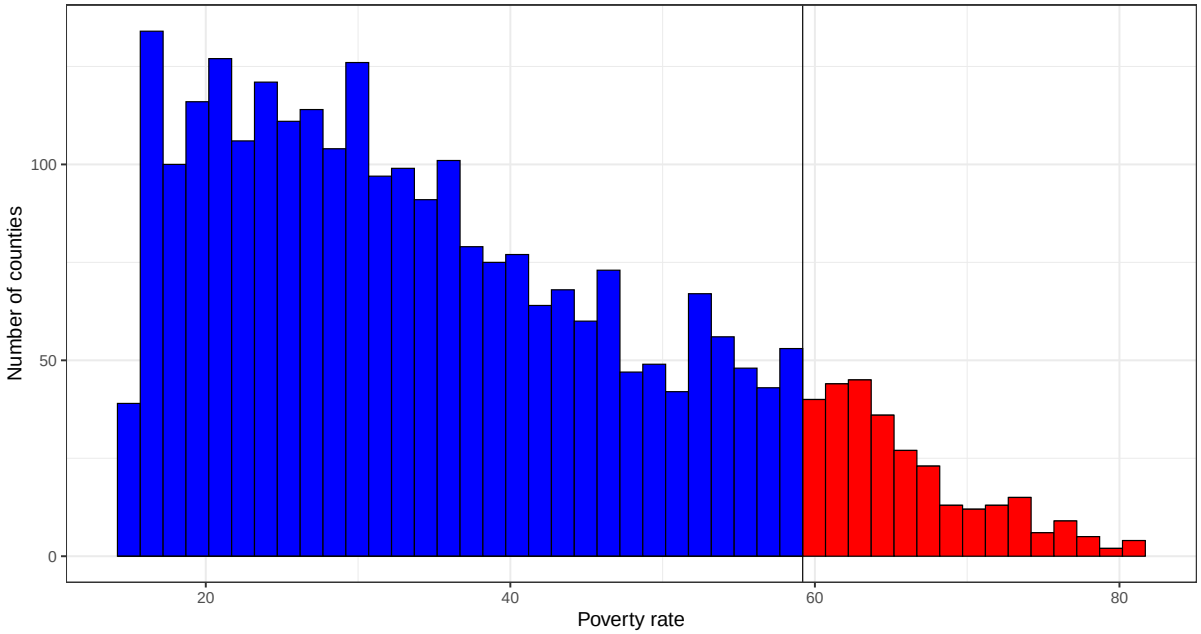
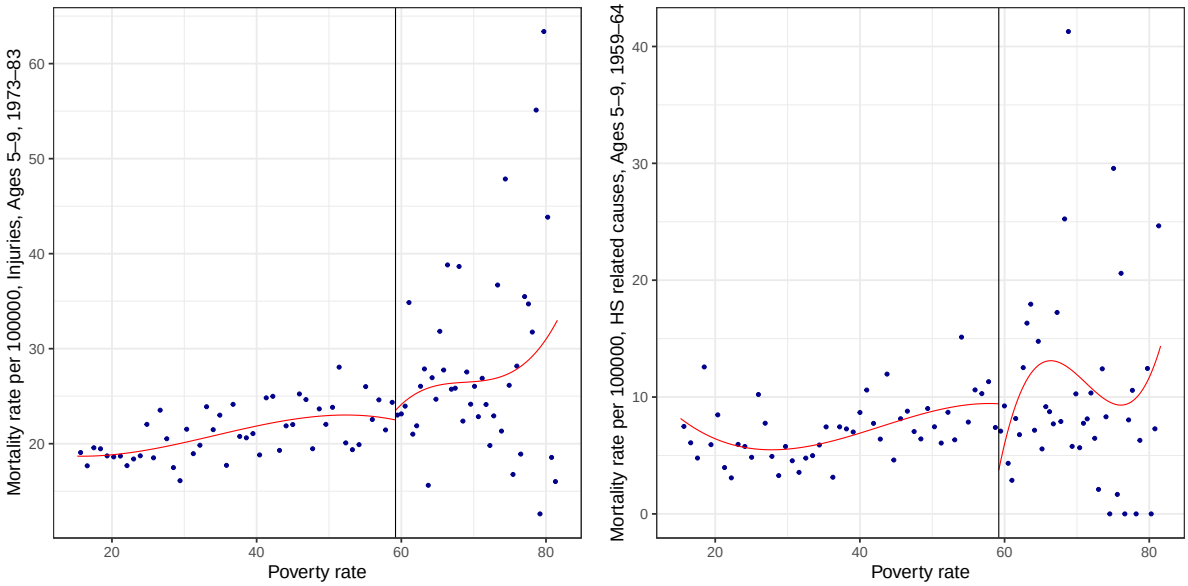


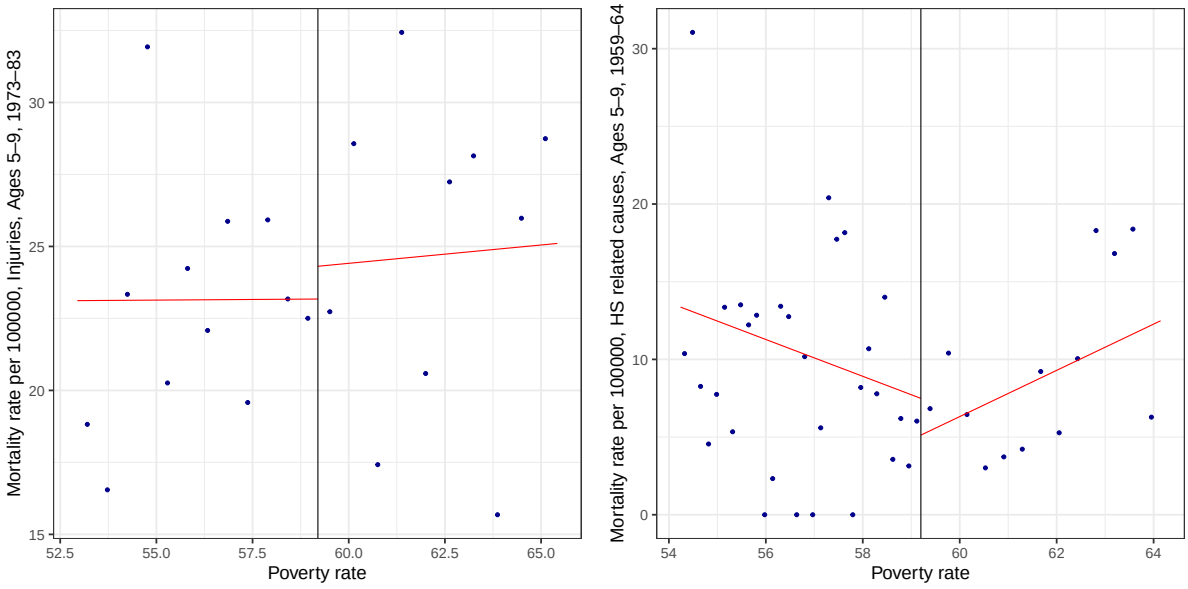
Figure A4: Histogram of the poverty rate (assignment variable). The histogram is constructed such that the cutoff (vertical line) constitutes a boundary, separating control (blue) and treated (red) counties.



(a) Post-treatment placebo outcome injury-related mortality

(b) Outcome of interest (Y) before treatment

Figure A5: RD plot for two covariates where no treatment effect is expected, to provide evidence for the validity of the RD design. The dots depict the mean within each bin, where the bins are evenly spaced and their number is chosen to mimic variance. The solid red lines are third-order polynomial fits for control (left) and treated (right) counties separately.



(a) Post-treatment placebo outcome injury-related mortality

(b) Outcome of interest (Y) before treatment

Figure A6: As Figure A5, but instead of global polynomial fits the local linear fits (triangular kernel, estimated MSE-optimal bandwidth) are displayed.

Table A2: RD effect analysis for covariates

Covariate	$\hat{h}_{\text{MSE}}$	RD estimate	RBC inference	
			95% CI	p-value
Mortality, Injuries, Ages 5–9, 1973–1983	6.262	1.133	[−7.074, 10.120]	0.728
Mortality, All causes, Ages 5–9, 1973–1983	6.345	−3.501	[−14.679, 7.371]	0.516
Mortality, Head Start related causes, Ages 25+, 1973–1983	8.054	2.032	[−12.557, 14.913]	0.866
Mortality, Injuries, Ages 25+, 1973–1983	6.205	0.052	[−13.815, 9.693]	0.731
Mortality, Head Start related causes, Ages 5–9, 1959–1964	4.963	−2.376	[−7.009, 3.476]	0.509
Census 1960: County population	9.352	3084.605	[−5920.470, 11918.011]	0.510
Census 1960: % attending school, Ages 14–17	10.083	0.557	[−4.335, 5.832]	0.773
Census 1960: % attending school, Ages 5–34	6.194	0.906	[−1.345, 3.325]	0.406
Census 1960: % high school or more, Ages 25+	7.651	0.621	[−1.490, 2.801]	0.549
Census 1960: Population, Ages 14–17	9.642	308.801	[−352.227, 1017.850]	0.341
Census 1960: Population, Ages 5–34	9.703	1648.581	[−2977.399, 6400.919]	0.474
Census 1960: Population, Ages 25+	8.312	1574.977	[−2894.802, 5741.418]	0.518
Census 1960: % urban population	8.846	2.339	[−5.567, 11.564]	0.493
Census 1960: % black population	7.166	0.793	[−10.237, 11.035]	0.941

Note: Results are based on local linear estimation and the triangular kernel.

Table A3: RD effect analysis for placebo cutoffs

Cutoff	$\hat{h}_{\text{MSE}}$	RD estimate	RBC inference	
			95% CI	p-value
		Below true cutoff		
31.463	4.595	0.032	[−1.142, 1.420]	0.831
40	6.114	0.015	[−1.494, 2.019]	0.769
50	2.340	1.894	[−1.689, 6.397]	0.254
		True cutoff		
59.198	6.913	−2.389	[−5.426, −0.083]	0.043
		Above true cutoff		
64.428	2.024	−0.490	[−2.802, 2.223]	0.821
70	3.054	−1.733	[−15.334, 10.639]	0.723

Note: The median poverty rate for control is 31.463, and for treatment 64.428. Results are based on local linear estimation and the triangular kernel.

Table A4: RD effect analysis for the donut-hole-approach

Radius donut hole	$\hat{h}_{\text{MSE}}$	RD estimate	RBC inference		Excluded counties	
			95% CI	p-value	Control	Treatment
0.0	6.913	−2.389	[−5.426, −0.083]	0.043	0	0
0.1	6.524	−2.515	[−5.821, −0.030]	0.048	3	3
0.2	6.371	−2.208	[−5.744, 0.564]	0.107	6	8
0.3	6.407	−1.966	[−5.604, 0.979]	0.168	9	10
0.4	6.438	−2.143	[−6.136, 1.045]	0.165	12	12
0.5	6.468	−2.478	[−7.196, 1.309]	0.175	18	16

Note: Results are based on local linear estimation and the triangular kernel.

References

Angrist, J. and V. Lavy (1999). "Using Maimonides' rule to estimate the effect of class size on scholastic achievement". *The Quarterly Journal of Economics* 114 (2), pp. 533–575.

Angrist, J. and M. Rokkanen (2015). "Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff". *Journal of the American Statistical Association* 110 (512), pp. 1331–1344.

Armstrong, T. and M. Kolesár (2018). "Optimal inference in a class of regression models". *Econometrica* 86 (2), pp. 655–683.

– (2020). "Simple and honest confidence intervals in nonparametric regression". *Quantitative Economics* 11 (1), pp. 1–39.

Barreca, A., M. Guldi, J. Lindo, and Waddell G. (2011). "Saving babies? Revisiting the effect of very low birth weight classification". *The Quarterly Journal of Economics* 126 (4), pp. 2117–2123.

Bertanha, M. and G. Imbens (2020). "External validity in fuzzy regression discontinuity designs". *Journal of Business & Economic Statistics* 38 (3), pp. 593–612.

Black, S. (1999). "Do better schools matter? Parental valuation of elementary education". *The Quarterly Journal of Economics* 114 (2), pp. 577–599.

Calonico, S., M. Cattaneo, and M. Farrell (2020). "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs". *The Econometrics Journal* 23 (2), pp. 192–210.

Calonico, S., M. Cattaneo, M. Farrell, and R. Titiunik (2019). "Regression discontinuity designs using covariates". *The Review of Economics and Statistics* 101 (3), pp. 442–451.

Calonico, S., M. Cattaneo, and R. Titiunik (2014). "Robust nonparametric confidence intervals for regression-discontinuity designs". *Econometrica* 82 (6), pp. 2295–2326.

– (2015). "Optimal data-driven regression discontinuity plots". *Journal of the American Statistical Association* 110 (512), pp. 1753–1769.

Cattaneo, M., N. Idrobo, and R. Titiunik (2020a). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge: Cambridge University Press.

Cattaneo, M., M. Jansson, and X. Ma (2020b). "Simple local polynomial density estimators". *Journal of the American Statistical Association* 115 (531), pp. 1449–1455.

Cattaneo, M., L. Keele, R. Titiunik, and G. Vazquez-Bare (2016). "Interpreting regression discontinuity designs with multiple cutoffs". *The Journal of Politics* 78 (4), pp. 1229–1248.

Cattaneo, M. and R. Titiunik (2022). "Regression discontinuity designs". *Annual Review of Economics* 14, pp. 821–851.

Cheng, M.-Y., J. Fan, and J. Marron (1997). "On automatic boundary corrections". *The Annals of Statistics* 25 (4), pp. 1691–1708.

Fan, J. (1992). "Design-adaptive nonparametric regression". *Journal of the American Statistical Association* 87 (420), pp. 998–1004.

Gelman, A. and G. Imbens (2019). "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics* 37 (3), pp. 447–456.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). "Identification and estimation of treatment effects with a regression-discontinuity design". *Econometrica* 69 (1), pp. 201–209.

Hall, P. (1992). "Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density". *The Annals of Statistics* 20 (2), pp. 675–694.

Hansen, B. (2022). *Econometrics*. New Jersey: Princeton University Press.

Imbens, G. and K. Kalyanaraman (2012). "Optimal bandwidth choice for the regression discontinuity estimator". *The Review of Economic Studies* 79 (3), pp. 933–959.

Imbens, G. and T. Lemieux (2008). "Regression discontinuity designs: A guide to practice". *Journal of Econometrics* 142 (2), pp. 615–635.

Kolesár, M. and C. Rothe (2018). "Inference in regression discontinuity designs with a discrete running variable". *American Economic Review* 108 (8), pp. 2277–2304.

Lee, D. (2008). "Randomized experiments from non-random selection in U.S. House elections". *Journal of Econometrics* 142 (2), pp. 675–697.

Lee, D. and T. Lemieux (2010). "Regression discontinuity designs in economics". *Journal of Economic Literature* 48 (2), pp. 281–355.

Li, K.-C. (1987). "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set". *The Annals of Statistics* 15 (3), pp. 958–975.

Ludwig, J. and D. Miller (2007). "Does Head Start improve children's life chances? Evidence from a regression discontinuity design". *The Quarterly Journal of Economics* 122 (1), pp. 159–208.

McCrary, J. (2008). "Manipulation of the running variable in the regression discontinuity design: A density test". *Journal of Econometrics* 142 (2), pp. 698–714.

Papay, J., J. Willett, and R. Murnane (2011). "Extending the regression-discontinuity approach to multiple assignment variables". *Journal of Econometrics* 161 (2), pp. 203–207.

Pei, Z., D. Lee, D. Card, and A. Weber (2022). "Local polynomial order in regression discontinuity designs". *Journal of Business & Economic Statistics* 40 (3), pp. 1259–1267.

R Core Team (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rubin, D. (2005). "Causal inference using potential outcomes: Design, modeling, decisions". *Journal of the American Statistical Association* 100 (469), pp. 322–331.

Sekhon, J. and R. Titiunik (2017). "On interpreting the regression discontinuity design as a local experiment". *Regression Discontinuity Designs: Theory and Applications*. Ed. by M. Cattaneo and J. Escanciano. Advances in Econometrics 38. Bingley: Emerald Group Publishing, pp. 1–28.

Thistlethwaite, D. and D. Campbell (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment". *Journal of Educational Psychology* 51 (6), pp. 309–317.

Statement of authorship

I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case.

Place, Date	Signature